DFFMamba: A Novel Remote Sensing Change Detection Method with Difference Feature Fusion Mamba

*PENG Daifeng**, *DONG Fengxu*, *GUAN Haiyan*

School of Remote Sensing and Geomatics Engineering, Nanjing University of Information Science and Technology, Nanjing 210044, P. R. China

(Received 1 August 2025; revised 15 October 2025; accepted 1 December 2025)

Abstract: Change detection (CD) plays a crucial role in numerous fields, where both convolutional neural networks (CNNs) and Transformers have demonstrated exceptional performance in CD tasks. However, CNNs suffer from limited receptive fields, hindering their ability to capture global features, while Transformers are constrained by high computational complexity. Recently, Mamba architecture, which is based on state space models(SSMs), has shown powerful global modeling capabilities while achieving linear computational complexity. Although some researchers have incorporated Mamba into CD tasks, the existing Mamba-based remote sensing CD methods struggle to effectively perceive the inherent locality of changed regions when flattening and scanning remote sensing images, leading to limitations in extracting change features. To address these issues, we propose a novel Mamba-based CD method termed difference feature fusion Mamba model (DFFMamba) by mitigating the loss of feature locality caused by traditional Mamba-style scanning. Specifically, two distinct difference feature extraction modules are designed: Difference Mamba (DMamba) and local difference Mamba (LDMamba), where DMamba extracts difference features by calculating the difference in coefficient matrices between the state-space equations of the bi-temporal features. Building upon DMamba, LDMamba combines a locally adaptive state-space scanning (LASS) strategy to enhance feature locality so as to accurately extract difference features. Additionally, a fusion Mamba (FMamba) module is proposed, which employs a spatial-channel token modeling SSM (SCTMS) unit to integrate multi-dimensional spatio-temporal interactions of change features, thereby capturing their dependencies across both spatial and channel dimensions. To verify the effectiveness of the proposed DFFMamba, extensive experiments are conducted on three datasets of WHU-CD, LEVIR-CD, and CLCD. The results demonstrate that DFFMamba significantly outperforms state-of-the-art CD methods, achieving intersection over union (IoU) scores of 90.67%, 85.04%, and 66.56% on the three datasets, respectively.

Key words: change detection; state space model (SSM) change feature fusion; deep learning; difference Mamba (DMamba); local difference Mamba(LDMamba); spatial-channel token modeling SSM (SCTMS)

CLC number: P237 **Document code:** A **Article ID:** 1005-1120(2025)06-0728-21

0 Introduction

Change detection (CD) is the process of identifying changes of an object or phenomenon by using images acquired at different times but same geographic areas[1]. It plays a vital role in numerous fields, including monitoring of land-use and land cover[2], urban sprawl[3], and geological-hazard monitoring[4]. In the past decade, the advances in satellite observation technology have made it increasingly feasible to acquire multi-temporal, high-resolution optical imagery with enhanced spatial detail and rich semantic features. However, how to efficiently and rapidly extract useful features and information from massive optical remote sensing (RS) data still poses great challenges, especially

*Corresponding author, E-mail address: daifeng@nuist.edu.cn.

How to cite this article: PENG Daifeng, DONG Fengxu, GUAN Haiyan. DFFMamba: A novel remote sensing change detection method with difference feature fusion Mamba[J]. Transactions of Nanjing University of Aeronautics and Astronautics, 2025,42(6):728-748.

http://dx.doi.org/10.16356/j.1005-1120.2025.06.003

for accurate multi-temporal image CD task[5].

In the literature, most traditional CD methods focus on detecting changed pixels and classifying them to generate a change map[6]. While these methods have demonstrated promising results on certain types of imagery, their effectiveness is highly dependent on domain-specific knowledge, often resulting in limited accuracy and poor generalization capabilities. The emergence of deep learning (DL) has introduced new models and paradigms for CD. Owing to its exceptional feature representation and nonlinear modeling capabilities, DL has substantially enhanced the efficiency and accuracy of CD, exerting a profound influence on the field. Consequently, deep learning-based change detection (DLCD) techniques continue to emerge, including convolutional neural network (CNN)-based methods[7], Transformer-based methods[8] and Mamba-based methods[9]. Specifically, CNNs enable powerful automatic feature extraction in CD, thereby capturing key semantic features of changes from bi-temporal images. However, the capacities of modeling long-range dependencies is severely constrained by limited receptive fields. Differently, based on self-attention units, Transformer architectures inherently possess global context modeling abilities. However, their application potential on remote sensing (RS) images is significantly limited due to the quadratic computational complexity, especially when processing high-resolution remote sensing imagery for pixel-level prediction tasks.

Recently, inspired by the capabilities of state-space models (SSMs)[10], especially Mamba[11], it is possible to effectively capture global information while maintaining linear computational complexity across a variety of computer vision tasks. Consequently, a growing body of research has been dedicated to the development and adaptation of Mamba-based architectures for remote sensing change detection (RSCD) task. Particularly, RS Mamba (RS-Mamba)[12] introduces an omnidirectional selective scan module(OSSM) to globally model image context across multiple directions. ChangeMamba[9], building upon the visual state space model (VMamba)[13] architecture, employs a cross-scanning mechanism to achieve effective modeling of global contextual information of images. While these methods broaden the perspective of RSCD by incorporating global awareness, the employed image flattening approach easily leads to a loss of locality in changed regions and compromises spatial consistency. LocalMamba[14] introduces locality by dividing the image into several fixed windows that are scanned individually. However, this fixed-window strategy not only introduces irrelevant background locality but also results in incomplete incorporation of the locality of changes. CD-Lamba[15] designed a locally adaptive state-space scanning (LASS) strategy that employs dynamic and adaptive windows, which enhances the locality of changes while preserving global context. The aforementioned methods primarily enhance global background modeling by refining the sequence scanning mechanism for image data. However, they lack a dedicated feature extraction mechanism for change regions, which hinders their ability to accurately and effectively capture discriminative change features for improved change detection. Although CD-Lamba successfully alleviates the feature locality loss associated with conventional scanning strategies, it similarly suffers from the absence of a targeted mechanism for extracting change-related features. As a result, it cannot reliably identify discriminative change patterns, and thus fails to achieve significant performance gains.

In addition, current bitemporal feature fusion methods commonly employ concatenation or subtraction to integrate change features. The emergence of Mamba has provided a new perspective for change feature integration. In such context, visual state space model for land cover change detection (LCCDMamba)[16] introduces a multi-scale information spatio-temporal fusion (MISF) module that integrates CNNs and Mamba to capture multi-scale spatio-temporal change information. However, this method merely concatenates bi-temporal features for processing, failing to achieve sufficient interaction between them. By contrast, CD-Lamba[15] designs a cross-temporal state-space scanning (CTSS) strategy, which allows any pixel in the bi-temporal data

to integrate information from all other pixels across different spatial directions and temporal states. However, CTSS only performs pixel-wise cross-connections for feature sequences in the spatial dimension, neglecting pixel-level dependencies across channel dimensions. As a result, the representation at each pixel fails to adequately incorporate cross-spatial, cross-channel, and cross-temporal information, which limits its effective multi-dimensional fusion.

To address the aforementioned challenges, we propose a difference feature fusion Mamba (DFFMamba) model , which retains the core advantages of Mamba in state space modeling while effectively enhancing the locality of early-stage features. By integrating the multi-dimensional spatio-temporal interactions of change features to enhance interactions across dimensions, DFFMamba can effectively detect dynamic changes in complex spatio-temporal environments, thereby significantly improving the accuracy and reliability of CD. Specifically, to address the issue that current Mamba-based change detection methods lack a dedicated feature extraction mechanism for change regions, we design a difference Mamba (DMamba) module, which effectively extracts change features by computing the difference between coefficient matrices in the state space equations. To further mitigate the loss of locality in early-stage features, we introduce LASS into DMamba and propose a local difference Mamba (LDMamba) module, which enhances feature locality and strengthens the extraction of discriminative change features by separating the input features into local-related and background-related components and individually scanning them. Finally, to address CTSS's inability to integrate pixel-level information from both spatial and channel dimensions, we construct a fusion Mamba (FMamba) module incorporating a spatial-channel token modeling SSM (SCTMS), which allows the model to interact the multi-dimensional spatio-temporal interactions of change features, effectively learning global contextual information. It promotes the unified fusion of spatio-temporal features and captures dependencies of change features across both spatial and channel dimensions, thereby enabling the network to accurately identify dynamic changes and correlated characteristics in complex spatio-temporal scenarios.

The primary contributions of this work are as follows:

(1) This paper proposes a novel DFFMamba for CD task where change-related features are effectively captured through the DMamba, while the loss of locality in early-stage features is mitigated via the LDMamba. Ultimately, the multi-dimensional spatio-temporal interactions of change features are integrated through the FMamba module. This integrated design significantly enhances the model capability to capture and identify change information, leading to competitive CD performance.

(2) A DMamba module is proposed to obtain reliable difference information by computing the difference between the coefficient matrices within the state-space equations. Building upon DMamba, an LDMamba module is proposed to address the loss of feature locality caused by conventional scanning mechanisms. In such a way, the global contextual information is preserved while the locality of early-stage features is enhanced, enabling precise extraction of change features.

(3) To integrate multi-dimensional spatio-temporal interactions of change features and enhance their cross-dimensional coupling, we introduce the FMamba module. This module employs the SCTMS method to interleave features across spatial and channel dimensions, enabling each pixel to integrate information from all other pixels across spatial, channel, and temporal dimensions.

1 Related Work

1.1 CNN-based change detection

Due to the exceptional capability in extracting local features, CNN architectures have been widely used to address CD tasks. Ref.[7] pioneered a UNet-based architecture named fully convolutional early fusion (FC-EF) by introducing a fully convolutional network. This approach concatenates bi-temporal images along the channel dimension before inputting them into the network. Two variants of fully convolutional Siamese-concatenation (FC-

Siam-Conc) and fully convolutional Siamese-difference (FC-Siam-Diff) are further developed by employing twin Siamese branches with shared weights. However, such methods struggle to efficiently learn highly discriminative change features. To this end, Siamese network and NestedUNet (SNUNet)[17] was introduced, which employed a densely connected Siamese network to mitigate the loss of deep spatial information. In Ref.[18], a deeply supervised image fusion network (DSIFN) was proposed to enhance learning of discriminative change characteristics. Similarly, an attention-based deeply supervised network(ADS-Net)[19] was proposed by devising an adaptive spatial and channel fusion attention (ASCFA) mechanism, which concurrently enhanced change features in both spatial and channel dimensions. In addition, it is challenging to mitigate the interference from pseudo-changes. To address this limitation, a feature-output space dual-alignment (FODA) framework is proposed to suppress spurious variations by modeling relational constraints in invariant regions across multi-temporal images[20]. In Ref.[21], a weighted double-margin contrastive loss was introduced to encourage the focus on change features while penalizing attention to invariant features, thereby effectively mitigating interference from spurious variations.

Despite achieving promising results, CNN-based approaches suffer from limited receptive fields due to fixed kernel sizes, which hinders global dependency capture and compromises consistency between local and global representations. In this paper, we adopt the recently proposed Mamba framework to leverage its exceptional global modeling capabilities, thereby attaining compelling performance.

1.2 Transformer-based change detection

Due to its powerful long-range dependencies modeling capabilities, vision Transformers (ViT)[22] have been extensively introduced into CD task, achieving superior performance against CNN-based counterparts. In a pioneer work, Ref.[23] employed Transformer encoders to capture rich contextual information from images, a Transformer decoder was subsequently used to refine the original feature representations. Ref.[8] introduced a pure Transformer-based Siamese network architecture for CD. This framework unifies hierarchically structured Transformer encoders with multilayer perceptron (MLP) decoders, eliminating the need for CNN-based feature extractors. In a similar work, Ref.[24] used Swin Transformer blocks as foundational units for both the encoders and decoders. While effectively mitigating the receptive field limitation of CNNs, these methods introduce substantial computational costs. To address this issue, existing research has focused on refining attention mechanisms to substantially reduce computational costs. Among these efforts, Ref.[25] proposed a lightweight structure-aware Transformer network (LSAT), which replaced the standard self-attention (SA) module in ViT with a cross-dimension interactive self-attention (CISA) module that operated with linear computational complexity, thereby significantly decreasing computational overhead. Similarly, Ref.[26] used a lightweight multi-head attention mechanism to optimize computational efficiency. Despite improvements in computational efficiency, striking a balance between efficiency and high accuracy remains a major challenge for Transformer-based approaches.

1.3 Mamba-based change detection

Mamba architectures, which are based on structured state space sequence models (S4) have recently garnered significant research attention due to their efficient contextual modeling capabilities with linear complexity. Particularly, VMamba[13] effectively adapts SSMs to visual domains through the introduction of the state space for 2D (SS2D) module. By using a cross-scanning module that traverses image spaces, SS2D converts non-causal visual data into ordered patch sequences for efficient processing. Note that a fundamental challenge in Mamba-based CD lies in optimizing scanning strategies for sequence modeling. While VMamba's cross-scanning approach mitigates directional sensitivity, it flattens spatial tokens and increases distances between adjacent elements, resulting in a loss of locality in regions of change. To mitigate this issue,

LocalMamba[14] is proposed by segmenting images into fixed-size windows and performing independent scans within each partition. However, this fixed-window strategy not only introduces irrelevant background locality, misleading the identification of changed regions, but also leads to incomplete integration of change locality. To address these limitations, CD-Lamba[15] designed a LASS strategy that employs dynamic and adaptive windows, which enhanced the locality of changes while preserving global context. However, due to the limitation of scanning mechanism, Mamba-based CD methods struggle to capture detail information. To overcome this drawback, the scaled residual ConvMamba (SRCM)[27] was proposed by synergistically harnessing Mamba for global context modeling while employing convolutional operations to enhance local details, thereby mitigating the deficiency of detail-specific cues. Ref.[28] first explored the potential of hybrid CNN-SSM by introducing a simple feature interaction module (FIM), enabling the simultaneous capture of global information and local features. Similarly, Ref.[29] designed CWMamba by utilizing Mamba modules for global feature integration and CNN-based feature extraction block (BCGF) for local feature enhancement. In Ref.[15], LASS was proposed to compensate for the missing of local information. However, it fails to capture change-specific features, leading to suboptimal CD performance. Instead, by incorporating the LASS concept into LD-Mamba module, significant performance gain is observed in our proposed DFFMamba.

2 Methodology

2.1 Preliminaries

Mamba is an emerging sequence modeling architecture that is attracting growing attention in the field of deep learning. This architecture exhibits close connections to CNNs, recurrent neural networks (RNNs), and classical SSMs. Particularly, SSMs are typically formulated as linear time-invariant (LTI) systems, mathematically grounded in a set of linear ordinary differential equations (ODEs), shown as

$$\boldsymbol{h}'(t)=\boldsymbol{A}\boldsymbol{h}(t)+\boldsymbol{B}\boldsymbol{x}(t) \tag{1}$$

$$\boldsymbol{y}(t)=\boldsymbol{C}\boldsymbol{h}(t)+\boldsymbol{D}\boldsymbol{x}(t) \tag{2}$$

where $\boldsymbol{x}(t)\in\mathbf{R},\boldsymbol{h}(t)\in\mathbf{R}^N,\boldsymbol{y}(t)\in\mathbf{R}$ represent the input sequence, the hidden state, and the output sequence; and $\boldsymbol{A}\in\mathbf{R}^{N\times N},\boldsymbol{B}\in\mathbf{R}^{N\times L},\boldsymbol{C}\in\mathbf{R}^N,\boldsymbol{D}\in\mathbf{R}$ the learnable parameters. N is the state size and L the input dimension.

In addition, to address the challenge of discretizing continuous systems for integration into deep learning, S4 is proposed by introducing a timescale parameter Δ. In such case, the continuous parameters $\boldsymbol{A}$ and $\boldsymbol{B}$ are converted into discrete parameters $\bar{\boldsymbol{A}}$ and $\bar{\boldsymbol{B}}$, leading to a commonly used discretization method of zero-order hold (ZOH), shown as

$$\bar{\boldsymbol{A}}=\exp(\Delta\boldsymbol{A}) \tag{3}$$

$$\bar{\boldsymbol{B}}=(\Delta\boldsymbol{A})^{-1}(\exp\Delta\boldsymbol{A}-\boldsymbol{I})\times\Delta\boldsymbol{B} \tag{4}$$

where $\boldsymbol{I}$ is the identity matrix. Therefore, the continuous ODE can be converted into a discrete form, i.e.

$$\boldsymbol{h}(t)=\bar{\boldsymbol{A}}\boldsymbol{h}_{t-1}+\bar{\boldsymbol{B}}\boldsymbol{x}(t) \tag{5}$$

$$\boldsymbol{y}(t)=\boldsymbol{C}\boldsymbol{h}(t)+\boldsymbol{D}\boldsymbol{x}(t) \tag{6}$$

where $\boldsymbol{D}$ acts as a residual connection and is often omitted from the equation, i.e.

$$\boldsymbol{y}(t)=\boldsymbol{C}\boldsymbol{h}(t) \tag{7}$$

Finally, the output is obtained via a global convolution, shown as

$$\bar{\boldsymbol{K}}=\left(\boldsymbol{C}\bar{\boldsymbol{B}},\boldsymbol{C}(\bar{\boldsymbol{A}}\bar{\boldsymbol{B}}),\cdots,\boldsymbol{C}\bar{\boldsymbol{A}}^{M-1}\bar{\boldsymbol{B}}\right) \tag{8}$$

$$\boldsymbol{y}=\boldsymbol{x}*\bar{\boldsymbol{K}} \tag{9}$$

where $\boldsymbol{M}$ is the length of the input sequence; $\bar{\boldsymbol{K}}\in\mathbf{R}^M$ the structured convolution kernel, and * the convolution operation.

Note that Mamba achieves a breakthrough enhancement over classical SSMs by introducing a selective scan mechanism. By dynamically adjusting model parameters according to the current input, this mechanism selectively propagates or forgets information, effectively overcoming limitations of traditional models in handling discrete and information-dense data.

2.2 Overall network architecture

DFFMamba consists of two weight-sharing encoders, three intermediate modules, and one decoder, as illustrated in Fig.1. The encoders are imple-

mented using VMamba V2 pre-trained on the ImageNet-1K dataset, which are composed of patch embedding layers, patch merging layers, and visual state space (VSS) block. The intermediate modules consist of DiffMamba, LDMamba, and FMamba. Specifically, DMamba and LDMamba are dedicated to enhance the locality of early-stage features and extract discriminative change features, while FMamba aims to integrate multi-dimensional spatio-temporal interactions of change features, enabling the model to effectively capture change characteristics in complex spatio-temporal environments. It should be noted that LDMamba, due to its unique scanning mechanism, is specifically employed to enhance the locality of shallow features and is therefore used only at the first stage. Meanwhile, DMamba also faces challenges in accurately localizing discriminative information within deep features. Therefore, we apply DMamba to extract difference information from the two intermediate sets of features. In addition, inspired by multi-scale vision Mamba UNet (MSVM-UNet)[30] model, the decoder consists of large kernel patch expanding (LKPE) layers and multi-scale visual state space (MSVSS) blocks. The LKPE layer performs up-sampling on the feature maps, which incorporates large-kernel depth-wise convolution prior to expanding the channel dimension to obtain more discriminative feature representations. MSVSS block captures and aggregates fine-grained multi-scale information while learning multi-dimensional spatio-temporal interactions from spatio-temporal features provided through skip connections, which mitigates directional sensitivity issues in 2D visual data and enhances the model's ability to comprehensively capture the features of the changed regions.

In general, bi-temporal images are processed by the encoder to generate multi-scale features, which are subsequently delivered to intermediate modules. Specifically, the LDMamba module enhances the locality of the first set of features to achieve more accurate extraction of difference information. These features are subsequently fed into the FMamba module to integrate multi-dimensional spatio-temporal interactions of change features, resulting in interactive spatio-temporal features. The second and third sets of features are processed by the DMamba module to extract difference information and are then passed to the FMamba module. The last set of features is directly input into the FMamba module to generate spatio-temporal features. These multi-scale spatio-temporal features

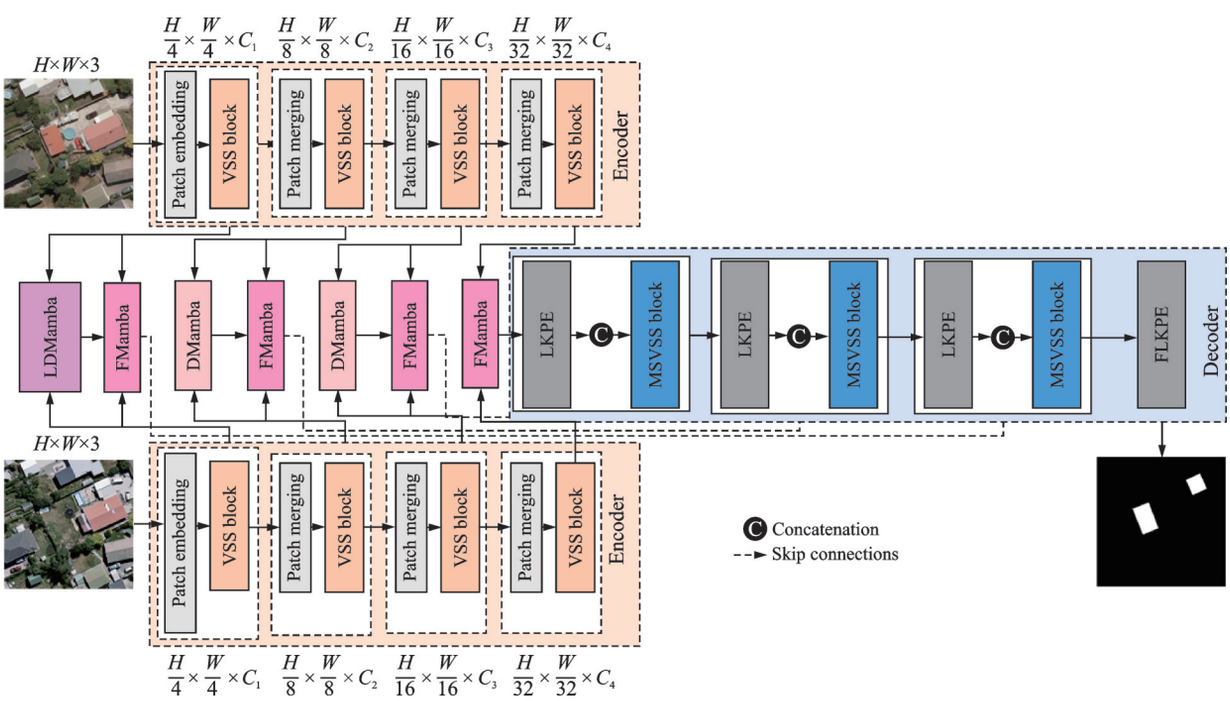

Fig.1 Overall framework of DFFMamba

are delivered to the corresponding layers in the decoder. Subsequently, the MSVSS blocks capture and aggregate fine-grained multi-scale information from the contracting path, while learning multi-dimensional spatio-temporal interactions from the skip connections. Finally, the CD map is obtained through the final LKPE (FLKPE) layer.

2.3 DMamba and LDMamba

Accurate extraction of difference information from bi-temporal images is essential for the decoder to capture robust change features. To this end, we propose the DMamba module. Furthermore, to address the loss of feature locality inherent in traditional Mamba scanning mechanisms, we introduce the LDMamba module. LDMamba separates features into local-related and background-related components. Each part is processed independently by DMamba for differential feature extraction. This strategy effectively preserves global contextual information while strengthening local feature coherence, leading to significant improvements in the discriminative capability of change representations.

2.3.1 DMamba

The DMamba module is integrated into the skip connections of the network, where it processes feature representations from the second and third encoder stages and performs specialized extraction of difference information. Assuming the feature representation from the kth encoder block is denoted as $F^k_{T1,T2} \in \mathbf{R}^{H_k \times W_k \times C_k}$, the entire process can be expressed as

$$\widehat{\mathrm{FD}}^{\,k}_{T1}, \widehat{\mathrm{FD}}^{\,k}_{T2} = \mathrm{DMamba}(F^k_{T1}, F^k_{T2}) \tag{10}$$

As shown in Fig.2, DMamba consumes corresponding bi-temporal features as input and produces two outputs while retaining the original spatial dimensions of the features. The input features are first processed through a linear projection layer (Linear) and a depthwise separable convolution (DWConv) layer, and are then fed into the difference selective scan (DSS) module.

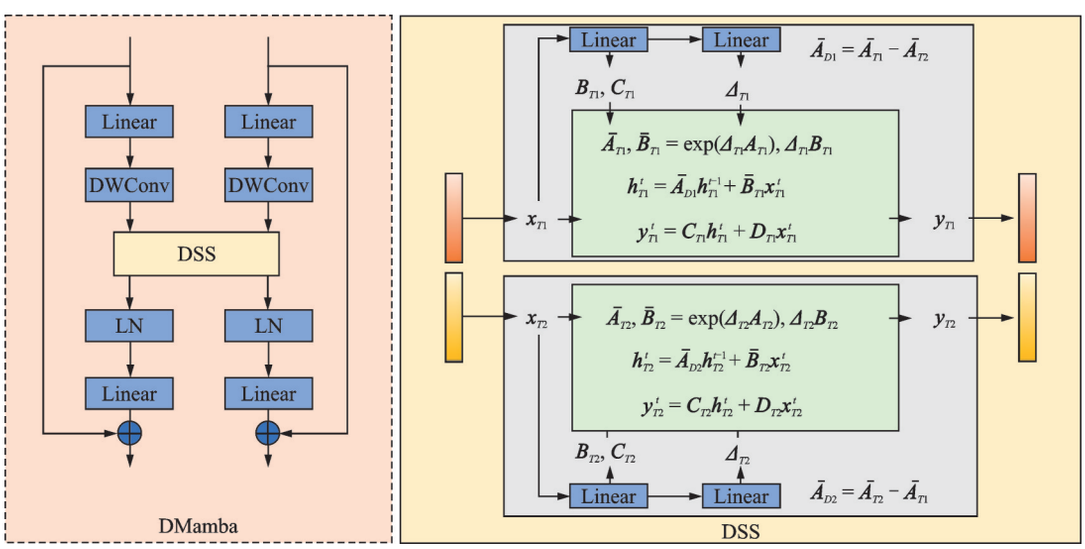

Fig.2 Overall structure of DMamba

Following the selective mechanism of Mamba, the coefficient matrices B, C, and Δ are generated from the input to enable the model's context-aware capabilities. Here, linear projection layers are utilized to generate these coefficient matrices. According to Eq.(5), matrix $\bar{A}$ is used to capture information from previous states in order to construct new states. To extract the difference information from the bi-temporal features, our study computes the difference of the coefficient matrix $\bar{A}$ to serve as the new coefficient matrix. The formula is as follows

$$\bar{A}_{T1}, \bar{B}_{T1} = \exp(\Delta_{T1}A_{T1}), \Delta_{T1}B_{T1} \tag{11}$$

$$\bar{A}_{T2}, \bar{B}_{T2} = \exp(\Delta_{T2}A_{T2}), \Delta_{T2}B_{T2} \tag{12}$$

$$\bar{A}_{D1} = \bar{A}_{T1} - \bar{A}_{T2} \tag{13}$$

$$\bar{A}_{D2} = \bar{A}_{T2} - \bar{A}_{T1} \tag{14}$$

$$h'_{T1} = \bar{A}_{D1}h'^{-1}_{T1} + \bar{B}_{T1}x'_{T1} \tag{15}$$

$$\boldsymbol{h}_{T2}^{t}=\bar{\boldsymbol{A}}_{D2}\boldsymbol{h}_{T2}^{t-1}+\bar{\boldsymbol{B}}_{T2}\boldsymbol{x}_{T2}^{t} \tag{16}$$

$$\boldsymbol{y}_{T1}^{t}=\boldsymbol{C}_{T1}\boldsymbol{h}_{T1}^{t}+\boldsymbol{D}_{T1}\boldsymbol{x}_{T1}^{t} \tag{17}$$

$$\boldsymbol{y}_{T2}^{t}=\boldsymbol{C}_{T2}\boldsymbol{h}_{T2}^{t}+\boldsymbol{D}_{T2}\boldsymbol{x}_{T2}^{t} \tag{18}$$

where $\boldsymbol{x}_{T1}^{t}$ and $\boldsymbol{x}_{T2}^{t}$ represent the inputs at time step t; $\bar{\boldsymbol{A}}_{D1}$ and $\bar{\boldsymbol{A}}_{D2}$ the computed difference coefficient matrices; and $\boldsymbol{y}_{T1}^{t}$ and $\boldsymbol{y}_{T2}^{t}$ the outputs of the selective scan representing the extracted difference features. The difference features are first subjected to layer normalization (LN), then linearly projected back to the original dimension, and finally combined with the original input via a residual connection. Residual connections are employed as the untreated features retain richer semantic information. The combination of treated and untreated features facilitates the integration of differential information across temporal dimensions, while also helping mitigate gradient vanishing and explosion issues.

2.3.2 LDMamba

The LDMamba module, as shown in Fig.3, is situated within the skip connections of the model, which processes the bi-temporal features extracted from the first stage of the encoder and produces two output representations while preserving the original spatial resolution of the features. This operational flow is formally expressed as

$$\widehat{\mathrm{FD}}_{T1}^{k},\widehat{\mathrm{FD}}_{T2}^{k}=\mathrm{LDMamba}(\boldsymbol{F}_{T1}^{k},\boldsymbol{F}_{T2}^{k}) \tag{19}$$

The input features are first processed through a linear projection layer and a DWconv layer, and are then fed into the local difference selective scan (LDSS) module. Within the LDSS module, input features are initially partitioned by the local adaptive difference split (LADS) module into local-related and background-related components. These components are subsequently processed by the DSS module to extract corresponding difference features. Finally, the extracted difference features are merged to the original feature dimension through the local adaptive difference split merge (LADS merge) module. Note that the LADS module consists of three main steps. First, to roughly identify regions rich in locality within the bi-temporal difference features, the absolute difference of the input features is computed and a (1/4, 1/4) average pooling is applied to construct a score window, where Gumbel Softmax is applied to introduce a differentiable approximation for discrete selection when identifying the Top_k windows with the highest scores as

$$\mathrm{Score}_{4\times4}=\sigma(\mathrm{AP}(\mathrm{abs}(\boldsymbol{X}_{T1}-\boldsymbol{X}_{T2}))) \tag{20}$$

where $\sigma(\cdot)$ denotes the Gumbel-Softmax operation, AP$(\cdot)$ the averaging pooling, and $\boldsymbol{X}_{T1}$ and $\boldsymbol{X}_{T2}$ are the input features. Subsequently, this score is utilized to reorganize the input features. Connected components within the Top_k windows are identified and merged to accommodate local change regions of varying shapes and sizes, as expressed by

$$\boldsymbol{W}_{R}=\mathrm{Re}(\boldsymbol{X}_{T},\mathrm{Top\_}k(\mathrm{Score}_{4\times4})) \tag{21}$$

where Re$(\cdot)$ represents the operation that merges connected components into connected windows, and $\boldsymbol{X}_{T}$ the input features. The matrix $\boldsymbol{W}_{R}$ consists of the Top_k windows, renumbered by connected windows, which are assigned values from the set of

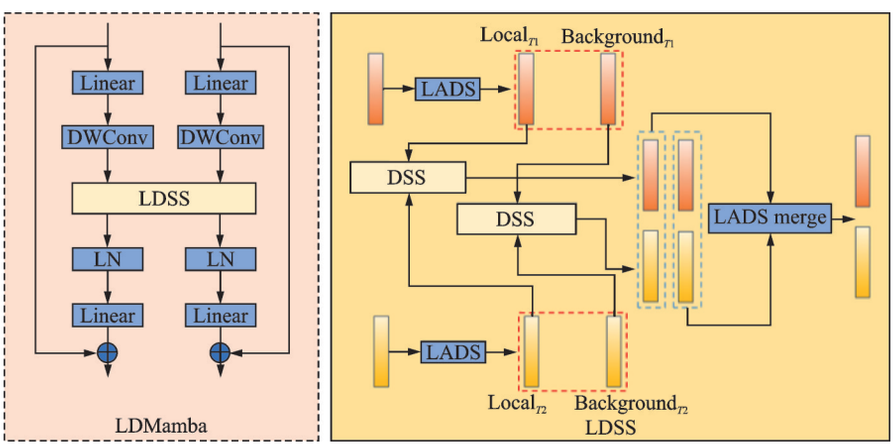

Fig.3 Overall structure of LDMamba

$\{1, 2, \cdots, k\}$. Finally, windows outside the Top_k are treated as a unified set and detached from the matrix to form the background component. Meanwhile, the Top_k windows are rearranged in ascending order based on their values and constitute the local component. This process is formally expressed as

$$F_{\mathrm{L}} = \mathrm{Ar}(W_R) \tag{22}$$

$$F_{\mathrm{B}} = \mathrm{Re}(X_T, \mathrm{NonTop}\_k(\mathrm{Score}_{4\times4})) \tag{23}$$

where F_{L} and F_{B} represent local-related and background-related components; $\mathrm{Ar}(\cdot)$ represents a sorting procedure and NonTop_k the windows outside the Top_k.

2.4 FMamba

To leverage these difference features, we design the FMamba module to integrate spatio-temporal interactions across dimensions. This allows the model to not only learn the complex dynamics of change features holistically but also capture their intrinsic relationships. As shown in Fig.4, the FMamba module takes three features as inputs: The bi-temporal features $\hat{F}_{T1}^{k}$, $\hat{F}_{T2}^{k}$ from the corresponding stage, together with the extracted difference feature $\widehat{\mathrm{FD}}_{D}^{k}$. The input features are first processed through a linear projection layer and a DWconv layer, and are then fed into the fusion selective scan (FSS) module.

This operational flow is formally expressed as

$$\overline{\mathrm{FD}}_{T1}^{k}, \overline{\mathrm{FD}}_{T2}^{k} = \mathrm{FSS}\left(\tilde{F}_{T1}^{k}, \tilde{F}_{T2}^{k}, \widetilde{\mathrm{FD}}_{D}^{k}\right) \tag{24}$$

where $\tilde{F}_{T1}^{k}$, $\tilde{F}_{T2}^{k}$, and $\widetilde{\mathrm{FD}}_{D}^{k}$ represent the processed input features. Subsequently, the feature $\overline{\mathrm{FD}}_{T1}^{k}$, $\overline{\mathrm{FD}}_{T2}^{k} \in \mathbf{R}^{H_k \times W_k \times C_k}$ output by the FSS module, are multiplied with two scaling parameters and concatenated in the channel dimension, forming a combined feature of shape $\mathbf{R}^{H_k \times W_k \times (2 \times C_k)}$. Finally, a linear projection layer is used to reduce the feature shape to $\mathbf{R}^{H_k \times W_k \times C_k}$.

Within the FSS module, the input features are first fused with their corresponding source bi-temporal features to retain essential contextual information. The resulting features are subsequently processed by the SCTMS module. Here, the features are reshaped across spatial and channel dimensions, enabling each pixel to comprehensively integrate information from all others across spatial, channel, and temporal domains. Subsequently, the features are fed into a specially designed VSS block, which learns different aspects of spatio-temporal relationships within change features, capturing the intrinsic connections of change information derived from spatio-temporal sequences. Finally, spatio-temporal features are restored to their original dimensions via SCTMS inverse. This process is described as

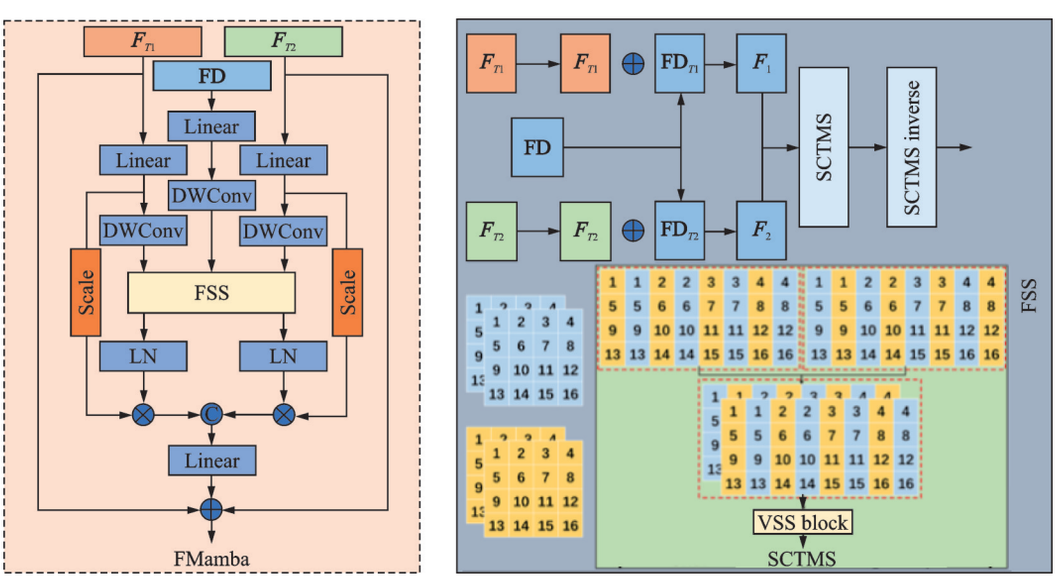

Fig.4 Overall structure of FMamba

$$F_1, F_2 = \mathrm{SCTMS}\left(\tilde{F}_{T1}^{k} + \widetilde{\mathrm{FD}}_{D}^{k}, \tilde{F}_{T2}^{k} + \widetilde{\mathrm{FD}}_{D}^{k}\right) \quad (25)$$

$$\overline{\mathrm{FD}}_{T1}^{k}, \overline{\mathrm{FD}}_{T2}^{k} = \mathrm{Inverse}(F_1, F_2) \quad (26)$$

where Inverse(•) represents the SCTMS inverse operation.

2.5 Loss function

For the binary change detection task, a simple binary cross-entropy loss is adopted to formulate the objective function, which can be expressed as

$$L_{\mathrm{bce}} = -\sum y_{\mathrm{t}} \lg y_{\mathrm{p}} + (1 - y_{\mathrm{t}}) \lg (1 - y_{\mathrm{p}}) \quad (27)$$

where y_{t} denotes the ground-truth change map and y_{p} the predicted change map.

3 Experimental Settings

3.1 Datasets

To verify the effectiveness of our proposed DFFMamba, three CD datasets are employed, namely WHU-CD dataset[31], LEVIR-CD dataset[32], and CLCD dataset[33], the detailed descriptions of these datasets are as follows.

WHU-CD: This dataset focuses on building changes with different scales, it consists of two high-resolution aerial images, with a size of 32 507 pixel × 153 54 pixel and a spatial resolution of 0.3 m. The images were captured in April 2012 and April 2016, covering the same region in Christchurch, New Zealand. To facilitate GPU training, images were cropped into 256 pixel × 256 pixel, which were then randomly partitioned into a training set (5 947 images), a validation set (743 images), and a test set (744 images).

LEVIR-CD: This dataset is a widely used binary change detection dataset containing 637 pairs of Google Earth images with a size of 1 024 pixel × 1 024 pixel and a resolution of 0.5 m. This dataset primarily focuses on changes of building construction and demolition. To facilitate GPU training, images were cropped into 256 pixel×256 pixel, which were randomly divided into a training set (7 120 images), a validation set (1 024 images), and a test set (2 048 images).

CLCD: It is designed for cropland CD and contains 600 pairs of remote sensing images with a size of 512 pixel×512 pixel and a spatial resolution ranging from 0.5 m to 2 m. The images were also cropped into clips of 256 pixel × 256 pixel, which were randomly split into a training set (1 440 images), a validation set (480 images), and a test set (480 images).

3.2 Implementation details

The proposed DFFMamba was implemented using the PyTorch framework powered by an NVIDIA GeForce RTX 4060 Ti 16GB GPU. During training, the AdamW optimizer was used for all three datasets with an initial learning rate of 6×10^{-5} and a weight decay coefficient of 1×10^{-2}. After a linear warm-up phase, the learning rate decays polynomially based on the number of training epochs, shown as

$$\mathrm{lr} = \mathrm{lr}_0 \times (1 - \mathrm{epoch}/150)^{\mathrm{lr\_power}} \quad (28)$$

where lr_0 denotes the initial learning rate, and lr_power the polynomial decay exponent and is set to 0.9. The batch size was uniformly set to 10, and the number of epochs was set to 150. Furthermore, the model random seed is set to 3 407, and the model weights are initialized using the Kaiming normal distribution to prevent gradient explosion or vanishing issues. In terms of data preprocessing, the three-channel data are first normalized using predefined mean and standard deviation values from the ImageNet dataset. Then, data augmentation techniques, such as random cropping, horizontal flipping, and vertical flipping, are applied to enhance the model's generalization capability.

3.3 Comparative methods and evaluation metrics

To demonstrate the superiority of the proposed DFFMamba model, a comparative study was conducted with the following nine classic deep learning-based change detection methods.

(1) SNUNet[17]: SNUNet incorporates a siamese network structure based on the NestedUNet encoder, and employs an integrated channel attention module to aggregate and refine the four output features from the decoder.

(2) DSIFN[18]: DSIFN employs a siamese convolutional encoder and a decoder enhanced with a spatial-channel attention mechanism for feature fu-

sion, while implementing deep supervision on each decoding layer.

(3) Spatial-temporal attention network (STANet)[32]: STANet employs a siamese convolutional network as its encoder. In the decoder, a pyramidal spatio-temporal attention mechanism is incorporated to capture multi-scale spatio-temporal features. Additionally, a metric learning approach is adopted to compute the change map.

(4) Bi-temporal image Transformer (BIT)[23]: BIT employs a siamese convolutional network as its encoder and introduces semantic tokens to leverage Transformer modules for semantically enhancing bi-temporal features.

(5) ChangeFormer[8]: ChangeFormer employs a siamese Transformer encoder to extract deep features from multi-temporal images, while a MLP in the decoder produces the change map.

(6) ChangeViT[34]: ChangeViT employs a plain ViT to extract high-level semantic features, while a detail-capture module extracts low-level detailed information. Subsequently a feature injector is introduced to inject the low-level details into high-level features.

(7) RS-Mamba[12]: RS-Mamba incorporates OSSM, which globally models image context by scanning in multiple directions, thereby capturing large spatial features from diverse orientations.

(8) ChangeMamba[9]: ChangeMamba building upon the VMamba architecture, employs a cross-scanning mechanism to achieve effective modeling of global contextual information of images.

(9) CD-Lamba[15]: CD-Lamba introduces the LASS to overcome the local perception limitations of conventional Mamba, while facilitating bi-temporal feature fusion through a CTSS strategy.

To evaluate the performance of our model, five metrics of precision (Pre), recall (Rec), overall accuracy (OA), F_1-score (F_1), and intersection over union (IoU) are employed, which are defined as

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{29}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{30}$$

$$\text{IoU} = \frac{\text{TP}}{\text{TP} + \text{FN} + \text{FP}} \tag{31}$$

$$\text{OA} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{32}$$

$$F_1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{33}$$

where true positive (TP) refers to the cases that are correctly classified as positive instances of change; false positive (FP) the cases that are incorrectly classified as positive instances of change; true negative (TN) the cases that are correctly classified as negative instances of no change and false negative (FN) the cases that are incorrectly classified as negative instances of change.

4 Experimental Results

4.1 Experimental results on WHU-CD dataset

As shown in Table 1, SNUNet effectively restores fine-grained information and achieves compet-

Table 1 Quantitative comparisons of different CD methods on WHU-CD, LEVIR-CD, and CLCD datasets(The top two optimal values are highlighted in red and blue) %

Type	Method	WHU-CD					LEVIR-CD					CLCD				
		IoU	F_1	OA	Pre	Rec	IoU	F_1	OA	Pre	Rec	IoU	F_1	OA	Pre	Rec
CNN-based	SNUNet[17]	76.95	86.98	98.92	83.47	90.78	79.83	88.79	97.79	89.98	87.63	41.20	58.95	93.74	57.78	58.95
	DSIFN[18]	79.31	88.46	99.13	92.94	84.40	81.18	89.61	97.80	93.30	86.21	44.27	61.37	94.08	59.65	63.19
	STANet[32]	73.61	84.80	98.73	80.97	89.00	78.70	88.10	98.70	85.00	91.40	47.52	64.43	94.58	62.30	65.92
Transformer-based	BIT[23]	69.70	82.15	98.45	75.74	89.74	81.75	89.96	98.89	90.50	89.42	46.29	63.29	94.93	68.62	58.73
	ChangFormer[8]	75.79	86.22	98.95	89.72	82.99	82.48	90.40	99.04	92.05	88.81	41.56	58.72	94.03	60.42	57.11
	ChangeViT[34]	89.66	94.55	99.57	95.61	93.51	84.65	91.69	99.16	92.41	90.98	63.54	77.70	96.77	79.95	75.58
Mamba-based	RS-Mamba[12]	81.79	89.99	99.22	92.16	87.91	82.48	90.40	99.03	91.39	89.42	55.54	71.42	96.02	76.65	66.86
	ChangeMamba(B)[9]	89.99	94.73	99.59	96.25	93.25	84.31	91.49	99.14	92.81	90.20	65.68	79.28	97.01	81.76	76.96
	CD-Lamba[15]	86.49	92.76	99.44	94.38	91.18	81.79	89.98	98.98	89.69	90.28	62.53	76.94	96.68	74.35	76.94
	DFFMamba	90.67	95.11	99.62	95.83	94.39	85.04	91.91	99.19	92.96	90.90	66.56	79.92	97.17	84.79	75.59

itive quantitative results with an F_1-score of 86.98% by leveraging dense connections to propagate features to the decoder. DSIFN further improves detection accuracy by incorporating an attention mechanism to enhance focus on changed regions, reaching an F_1-score of 88.46%. However, it still falls short in capturing global contextual information. To model global context, ChangeViT introduces a plain ViT architecture and achieves a higher F_1-score of 94.55% among the compared methods. Meanwhile, ChangeMamba, built upon the VMamba architecture, incorporates a cross-scanning mechanism that better facilitates global contextual modeling, leading to a further improved F_1-score of 94.73%. Notably, DFFMamba employs an LDMamba module to mitigate the loss of feature locality caused by traditional scanning strategies while retaining global contextual understanding. Consequently, it obtains the best performance with values of 90.67% for IoU, 95.11% for F_1-score, 99.62% for OA, and 94.39% for Recall.

For a qualitative comparison of the CD performance across different methods, several typical scenes are selected, as shown in Fig.5. The first and the second rows in Fig.5 illustrate the complex building change scenario. It can be observed that DFFMamba accurately extracts changes in intricate building structures, whereas other models fail to precisely localize building boundaries, resulting in a significant number of false positives. The third row in Fig.5 illustrates middle-scale building changes within a simple environment. It can be observed that RSMamba, ChangeMamba, and CD-Lamba all deliver visually plausible results with clearly delineated building boundaries. Additionally, SNUNet, with its densely connected architecture, effectively restores fine-grained details, resulting in CD results with relatively refined edges. However, due to the inability to establish global contextual relationships, it exhibits certain missed detections within building interiors. The fourth row in Fig.5 demonstrates the exceptional capability of DFFMamba in detecting minute structural changes, while other methods exhibit noticeable missed detections. This superiority can be attributed to DFFMamba's enhanced extraction of difference features through LDMamba and DMamba. FMamba further leverages these difference representations to holistically model spatio-temporal dependencies, enabling the model to accurately identify change features.

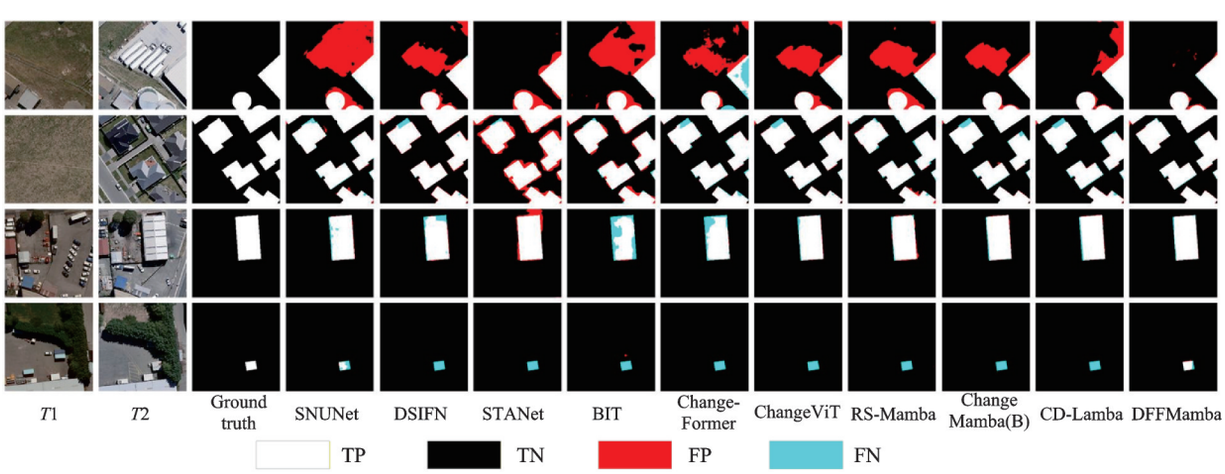

Fig.5 Visual comparison of different CD methods on the WHU-CD dataset

4.2 Experimental results on LEVIR-CD dataset

The quantitative evaluation results of different methods are presented in Table 1. DFFMamba achieved the highest scores with values of 85.04% for IoU, 91.91% for F_1-score and 99.19% for OA. ChangeViT achieves the second-best performance by injecting low-level details into high-level features, thereby enhancing the detection of changes at different scales. Compared to attention-based methods such as DSIFN and STANet, BIT and ChangeFormer, which incorporate Transformer structures into the encoder, can more effectively model global

contextual information, resulting in higher accuracy.

For visual comparisons, several typical scenes are also selected, as shown in Fig.6. Specially, The first and the second rows in Fig. 6 illustrate dense building change scenarios. The results demonstrate that DFFMamba effectively captures clustered structural changes, whereas other models fail to accurately localize building boundaries, resulting in numerous false positive pixels between adjacent structures. The third and the fourth rows in Fig.6 depict large-scale building change scenarios. Compared to other methods, DFFMamba demonstrates stronger local detail extraction capability, more accurately locates building boundaries, and finer edge structures. This improvement can be attributed to the LDMamba module in DFFMamba, which enhances feature locality and mitigates the loss of feature locality caused by conventional scanning mechanisms.

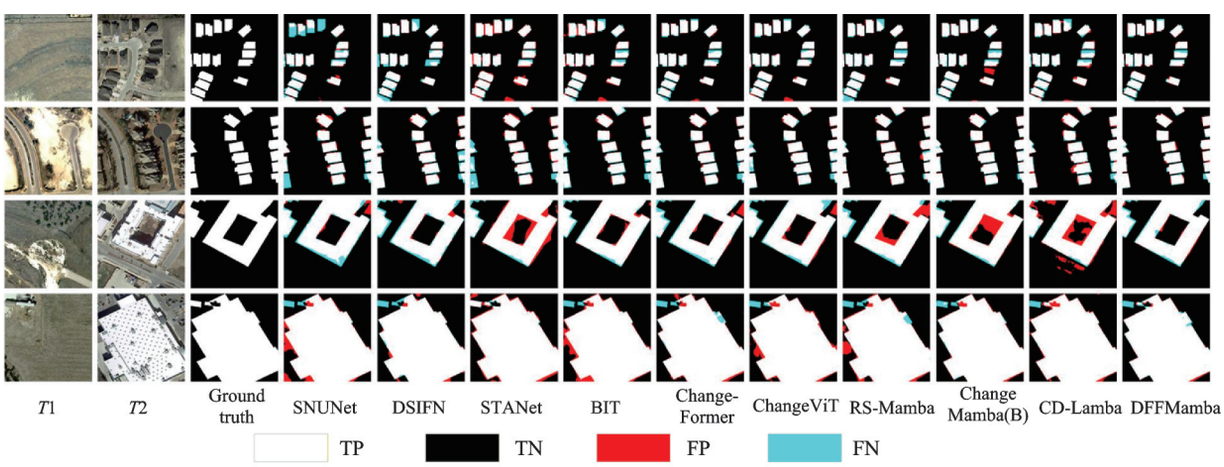

Fig.6 Visual comparison of different CD methods on the LEVIR-CD dataset

4.3 Experimental results on CLCD dataset

The quantitative evaluation results of different CD methods are presented in Table 1. DFFMamba achieved the highest scores across four quantitative metrics, with values of 66.56% for IoU, 79.92% for F_1-score, 97.17% for OA, and 84.79% for Precision. ChangeMamba employs a cross-scanning mechanism to achieve effective modeling of global contextual information, achieving the second-best accuracy among the compared methods. Notably, RS-Mamba utilizes an OSSM module to globally model image context through multi-directional scanning, yielding competitive accuracy. Furthermore, CD-Lamba addresses the loss of feature locality caused by conventional scanning mechanisms by using a LASS module, which enhances local feature representation and achieves superior accuracy compared to RS-Mamba.

For qualitative analysis, four typical sets are selected, as shown in Fig.7. The first and the second rows in Fig.7 depict change scenarios in lakes and bare soil lands, respectively. The results demonstrate that DFFMamba effectively identifies lake and bare soil lands changes, indicating strong generalization capability. As shown in the first row in Fig.7, methods such as SNUNet, DSIFN, BIT, ChangeFormer, RS-Mamba, and CD-Lamba fail to accurately locate lake change regions. In contrast, STANet, ChangeViT, and ChangeMamba suffer from severe missed detections and false alarms. The third and the fourth rows in Fig.7 depict buidings change scenarios. It can be observed that, compared to lake changes, all models have demonstrated improved performance in detecting building changes. As shown in the third row in Fig.7, ChangeMamba achieves relatively refined performance on edges. CD-Lamba mitigates the loss of feature locality by leveraging a LASS module, resulting in finer edge structures. By incorporating LASS to enhance feature locality, along with LDMamba and DMamba to improve the extraction of differential features, and further integrating spatio-temporal interaction mechanisms across different dimensions via FMam-

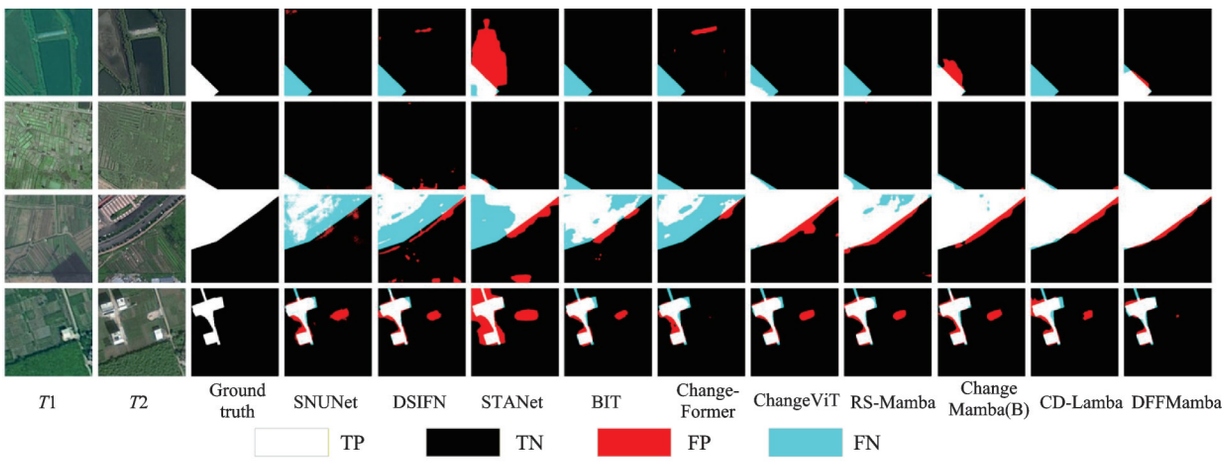

Fig.7 Visual comparison of different CD methods on the CLCD dataset

ba, DFFMamba effectively identifies change characteristics and produces the most precise edge details.

To clarify the limited accuracy of various CD methods on the CLCD dataset, Fig.8 presents four representative samples with poor CD performance. The first and the second rows in Fig.8 illustrate change scenarios in medium-and large-scale bare soil areas, respectively. The results indicate that all methods exhibit substantial omission and commission errors. This challenge arises because bare soil areas, in contrast to well-defined structures with high-contrast textures, often exhibit blurred edges, low contrast, and large homogeneous expanses, making precise feature extraction difficult. In large-scale bare soil scenario, methods such as RS-Mamba, ChangeMamba, and CD-Lamba capitalize on the Mamba architecture to effectively model global context. This allows them to surpass the performance of CNN-based methods, which are constrained by limited receptive fields and thus unable to capture comprehensive contextual information. By integrating multi- dimensional spatiotemporal interactions of change features via FMamba, DFFMamba enhances the model's ability to identify dynamic changes in complex scenarios, thus achieving superior results. The third and the fourth rows in Fig.8 illustrate a change scenario involving roads. As shown, all CD methods fail to detect the target area effectively. This challenge may stem from the agricultural settings of roads in the CLCD dataset, where pronounced seasonal dynamics in background factors are easily captured by models, thereby compromising CD accuracy. Note that DFFMamba improves the extraction of change features through its DMamba and LDMamba modules, while the FMamba module integrates

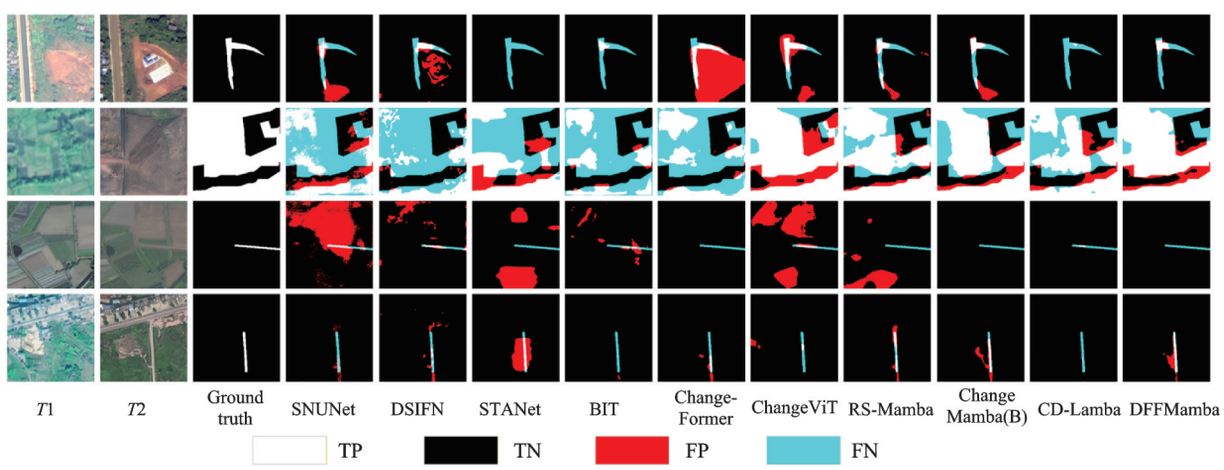

Fig.8 Visual comparison of failure cases for different CD methods on the CLCD dataset

multi-dimensional spatio-temporal interactions among these features. This design mitigates issues like inadequate change feature extraction and background noise interference. Nevertheless, it still fails to fundamentally resolve the core challenges posed by highly dynamic background variations and the intrinsic difficulty in obtaining discriminative change features.

4.4 Ablation study

4.4.1 Impact of different modules in DFFMamba

DFFMamba utilizes LDMamba to enhance feature locality, thereby improving DMamba's ability to extract difference information from bi-temporal features. Furthermore, FMamba leverages these difference representations to integrate spatio-temporal interaction mechanisms across different dimensions, enabling a comprehensive understanding of various aspects of spatio-temporal dependencies in change features. This allows the model to accurately identify change characteristics. To evaluate the impact of the aforementioned modules on CD performance, ablation studies were conducted on the WHU-CD and CLCD datasets, respectively. The results of these experiments for DFFMamba are presented in Table 2. As observed, the introduction of the FMamba module leads to improved accuracy across CD metrics. On the WHU-CD dataset, IoU and F_1-score increased by 2.17% and 1.21%, respectively, while on the CLCD dataset, the corresponding improvements reached 5.61% and 4.21%. These gains can be attributed to FMamba's employment of SCTMS, which enables a comprehensive learning of spatio-temporal dependencies and enhances the fusion of bi-temporal representations. By incorporating either the DMamba or LDMamba module, we observed a comprehensive improvement in all evaluation CD metrics. On the WHU-CD dataset, IoU increased by 1.68% and 1.89%, while the F_1-score increased by 0.94% and 1.06%, respectively. On the CLCD dataset, the standalone use of DMamba led to an increase of 0.96% in IoU and 0.73% in F_1-score, which can be attributed to its enhanced ability to extract change-related features. However, it should be noted that the separate introduction of the LDMamba module resulted in a decrease in model accuracy on the CLCD dataset. This decline is likely due to the low contrast and blurred boundaries characteristic of CLCD data, which hinder LDMamba from effectively capturing localized regions and thus weaken its change feature extraction capability. In addition, the joint integration of the DMamba and LDMamba modules yielded improvements across all change detection metrics, outperforming the use of either module alone. Compared to using only the DMamba module, the IoU increased by 0.38% and 0.31% on the WHU-CD and CLCD datasets, respectively. This improvement stems from the LDMamba module's ability to preserve locality in early-stage features, which in turn strengthens the model's extraction of change-related features. Compared to using only the LDMamba module, the IoU improved by 0.17% and 3.4% on the WHU-CD and CLCD datasets, respectively. This gain stems from the fact that while the LDMamba module may lose locality

Table 2 Ablation studies on WHU-CD and CLCD datasets(The top two optimal values are highlighted in red and blue)

Method			WHU-CD/%					CLCD/%					FLOPs/10^9	Parameter number/10^6
FMamba	DMamba	LDMamba	IoU	F_1	OA	Pre	Rec	IoU	F_1	OA	Pre	Rec		
			88.20	93.73	99.50	93.81	93.65	60.43	75.34	96.56	80.64	70.69	16.66	92.96
√			90.37	94.94	99.60	96.11	93.84	66.04	79.55	97.12	84.46	75.17	18.71	101.97
	√		89.88	94.67	99.58	95.46	93.90	61.39	76.07	96.69	82.20	70.80	17.86	99.74
		√	90.09	94.79	99.59	95.80	93.79	58.30	73.66	96.40	80.94	67.58	17.86	104.85
	√	√	90.26	94.88	99.60	96.19	93.61	61.70	76.31	96.67	80.94	72.19	17.86	104.85
√		√	90.24	94.87	99.60	95.97	93.79	63.84	77.93	96.83	80.92	75.15	20.38	113.86
√	√		90.49	95.01	99.61	95.77	94.26	66.26	79.71	97.15	84.76	75.22	20.38	108.75
√	√	√	90.67	95.11	99.62	95.83	94.39	66.56	79.92	97.17	84.79	75.59	20.38	113.86

when processing deep features, impairing change feature extraction, the DMamba module overcomes this by preserving local details, leading to more effective feature extraction for change detection.

Furthermore, combining the DMamba and LDMamba modules with the FMamba module led to consistent improvements across all change detection metrics. When combined with the FMamba module, DMamba increased IoU by 0.61% (WHU-CD) and 4.87% (CLCD), while LDMamba improved it by 0.15% and 5.54%, respectively, outperforming their standalone use. This improvement stems from FMamba's capacity to effectively model the multi-dimensional spatio-temporal interactions within the difference representations it receives, which in turn enhances the model's recognition of dynamic changes in complex scenes. The integration of all three modules (FMamba, DMamba, and LDMamba) attained optimal performance. This full combination yielded further gains over the two-module baselines: (1) Compared to FMamba+DMamba, adding LDMamba improved IoU by 0.18% and the F_1-score by 0.10% on WHU-CD, and by 0.30% and 0.21% on CLCD. This gain is attributable to the LDMamba module alleviating locality loss in early-stage features and enhancing change-related feature extraction; (2) compared to FMamba+LDMamba, replacing LDMamba with DMamba for deep-feature processing increased IoU by 0.43% and F_1-score by 0.24% on WHU-CD, and by 2.72% and 1.99% on CLCD. This improvement stems from DMamba's superior capacity to strengthen locality in deep features, which mitigates a limitation of LDMamba in this stage. Finally, incorporating the FMamba module on top of the DMamba and LDMamba combination yielded further improvements: IoU increased by 0.41% and the F_1-score by 0.23% on the WHU-CD dataset, and by 4.86% and 3.61% on the CLCD dataset. These gains result from FMamba's role in holistically modeling the spatio-channel interactions of the change features extracted by the preceding modules, which strengthens the model's capacity to identify dynamic changes in complex spatio-temporal scenes. It is worth noting that, with the sequential incorporation of the DMamba, LDMamba, and FMamba modules, the overall parameters and floating-point operations (FLOPs) of the model maintain a moderate growth trend. After integrating all the proposed modules, DFFMamba exhibits an increase of 20.9×10^6 parameters and 3.72×10^9 FLOPs compared to the baseline model, while still operating with high efficiency.

To qualitatively compare the impact of different modules on CD performance, we conducted a visual analysis of the DFFMamba model by incrementally adding modules and configuring different module combinations, as shown in Fig.9. It can be observed that the progressive integration of the proposed modules leads to significant improvements in change detection performance, with a marked reduction in both false positives and missed detections. Specifically, as illustrated in Figs.9(a, d), the full three-module combination achieves the most substantial reduction in both error types across challenging areas. By contrast, the incomplete module combinations exhibit distinct shortcomings: (1) Using only FMamba lacks a dedicated mechanism for extracting change regions, failing to capture discriminative change features effectively; (2) the DMamba and LDMamba combination can extract change features but cannot integrate multi-dimensional change information in a unified manner, limiting its capacity to recognize dynamic changes in complex spatio-temporal scenes; (3) the FMamba and DMamba combination still suffers from locality loss in early-stage features, which weakens feature extraction and hinders the identification of changes in specific regions. As shown in Fig.9(c), compared with the baseline method, adding the DMamba module significantly reduces false positives along building boundaries. This improvement is attributed to DMamba's specialized design for extracting discriminative change features. When further combining DMamba with LDMamba, false positives along building boundaries are greatly reduced, while missed detections still occur inside buildings. When only the FMamba module is added, the model struggles to fully learn multi-dimensional spatio-temporal interactions of change features due to the lack of ac-

curate change representations, leading to missed detections within buildings. However, after progressively integrating DMamba and LDMamba, missed detections inside buildings are markedly reduced. This is because FMamba, when supplied with accurate and reliable change features, can effectively capture dependencies across both spatial and channel dimensions, thereby strengthening the model's ability to identify dynamic changes in complex spatio-temporal scenes.

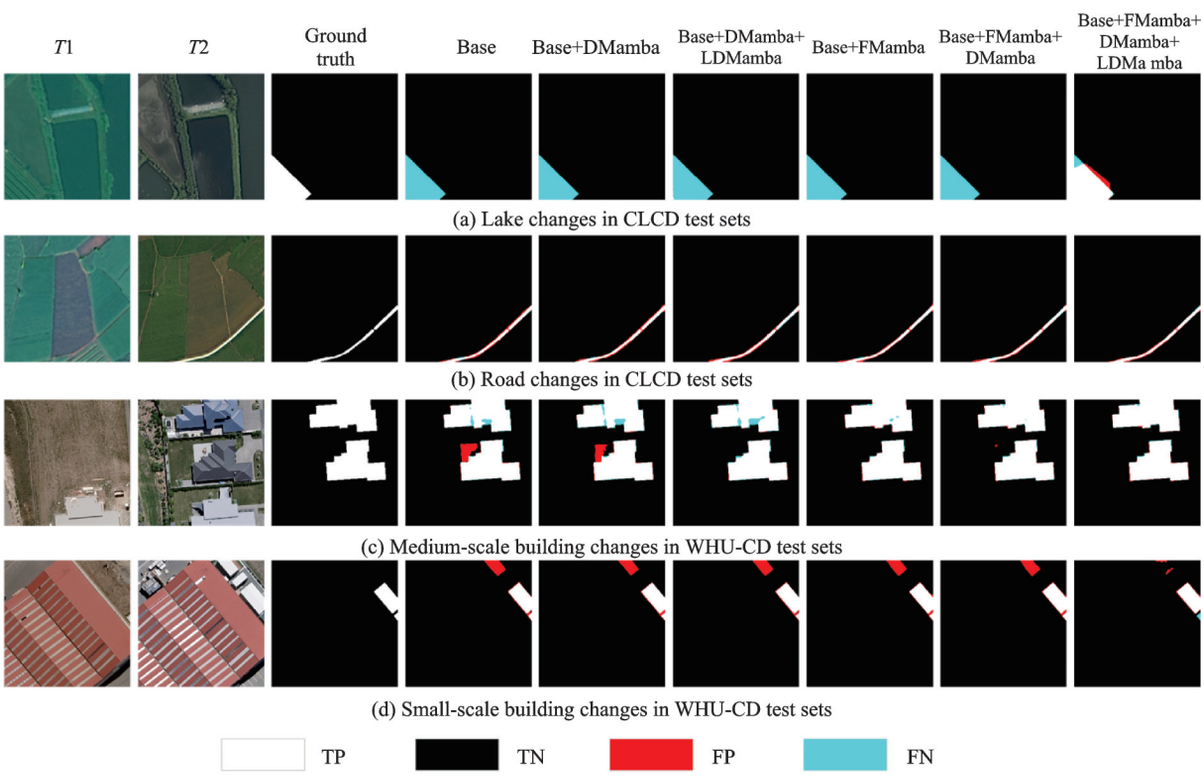

Fig.9 Visualization results of ablation studies on the CLCD and WHU-CD test sets

4.4.2 Impact of different Top_k in LDMamba

To approximately locate regions with strong locality, we performed a preliminary study to determine the key hyperparameter Top_k in the LADS component of LDMamba. Since change regions typically occupy a small image area, we evaluated three candidate values: Top_$k\in\{4, 6, 8\}$. As summarized in Table 3, setting Top_$k=8$ yields the best performance on both datasets: IoU reaches 66.56% and F_1-score 79.92% on CLCD, while on WHU-CD, IoU is 90.67% and F_1-score 95.11%. These results indicate that Top_$k=8$ is the optimal choice for the model.

Table 3 Ablation study of Top_k on WHU-CD and CLCD datasets %

Top_k	WHU-CD					CLCD				
	IoU	F_1	OA	Pre	Rec	IoU	F_1	OA	Pre	Rec
Top_4	90.41	94.96	99.60	95.85	94.09	64.71	78.57	96.99	83.66	74.07
Top_6	90.03	94.76	99.59	95.67	93.86	64.63	78.52	97.02	84.79	73.11
Top_8	90.67	95.11	99.62	95.83	94.39	66.56	79.92	97.17	84.79	75.59

4.4.3 Quantitative visualization

Fig.10 visualizes intermediate feature representations across three critical processing stages: The LDMamba module, DMamba module, and FMamba module. As presented in Fig.10, high-re- ponse regions progressively converge toward change areas after the processing of LDMamba and DMamba, yet exhibit limited spatial precision at change boundaries. It indicates that these modules are capable of effectively focusing model attention on change-rele-

vant regions while lacking pixel-accurate localization capability. Furthermore, subsequent FMamba processing leads distinct high-response activation precisely within change boundaries. This confirms that FMamba improves the model's capability to accurately detect change characteristics by incorporating multi-dimensional spatio-temporal interactions and enabling a holistic capture of multiple aspects of the spatio-temporal dependencies within change features.

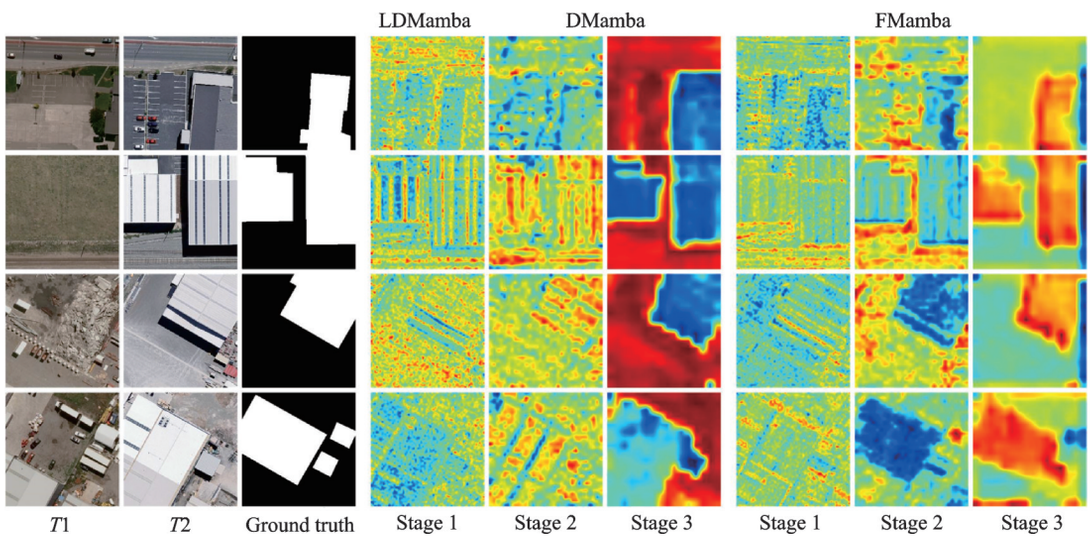

Fig.10 Visualization of the intermediate features on the WHU-CD dataset

4.5 Efficiency analysis

To comprehensively compare the efficiency of different CD methods, we conduct the computational efficiency by calculating the model's parameters, FLOPs and IoU on the WHU-CD dataset. As shown in Table 4, CNN-based models demonstrate higher cost-effectiveness under computational resource constraints due to their relatively smaller parameter counts. For instance, SNUNet achieves an IoU of 76.95% with only 12.03×10^6 parameters by leveraging a densely connected architecture. DSIFN incorporates an attention mechanism at the expense of increased computational overhead, achieving slight performance gain but large increase in model parameters and computational complexity. Differently, Transformer-based models excel in global feature modeling and achieve higher accuracy. ChangeViT, for example, employs a plain vision Transformer to extract high-level semantic features, attaining an IoU of 89.66% with 38.80×10^9 FLOPs and 32.13×10^6 parameters. By contrast, through optimized selective scanning strategy with enhanced global modeling capacity, Mamba-based models strike a balance between performance and efficiency. For instance, RS-Mamba and CD-Lamba achieve an IoU of 81.79% and 86.49%, respectively, with

Table 4 Comparisons of model complexity and accuracy on the WHU-CD dataset

Type	Model	FLOPs/10^9	Parameter number/10^6	IoU/%
CNN-based	SNUNet	54.82	12.03	76.95
	DSIFN	82.26	50.44	79.31
	STANet	13.14	16.93	73.61
Transformer-based	BIT	10.63	3.49	69.70
	ChangeFormer	202.86	41.01	75.79
	ChangeViT	38.80	32.13	89.66
Mamba-based	RS-Mamba	15.7	27.9	81.79
	ChangeMamba(B)	179.32	84.70	89.99
	CD-Lamba	15.26	28.74	86.49
	DFFMamba	20.38	113.86	90.67

only a modest increase in model complexity (15.7×10^9 FLOPs and 15.26×10^9 FLOPs).

Note that DFFMamba achieves new state-of-the-art IoU score while maintaining competitive computational efficiency, although it requires relatively higher parameters due to its use of a powerful yet parameter-dense VMamba V2 backbone.

5 Conclusions

In this paper, we propose DFFMamba, a novel Mamba-based difference feature fusion model for CD. To accurately extract difference information in bi-temporal features, an intermediate module termed DMamba is introduced, where differential information is derived by calculating the disparity between state equation coefficient matrices. To mitigate the loss of feature locality caused by conventional scanning mechanisms, a LASS module is incorporated into DMamba, resulting in the proposed LDMamba module, which specifically enhances locality in early-stage features. Additionally, a FMamba module with the SCTMS unit is proposed to enable a holistic capture of spatio-temporal dependencies in change features, leading to improved capability to identify change features. To validate the effectiveness of the proposed DFFMamba, extensive experiments were conducted on three public datasets. The quantitative results demonstrate that the DFFMamba achieves the best accuracy metrics across all datasets. Visual comparisons indicate that the proposed method effectively captures fine details such as edge structures and small object changes, while significantly reducing both the missed detections and false alarms. Furthermore, ablation studies confirm the contribution of each component in the network. It should be noted that the advanced performance of DFFMamba relies on a large amount of high-quality labeled data. However, acquiring high-quality annotations for change detection is both time-consuming and labor-intensive, which significantly limits the model's practicality in real-world scenarios. In future work, we plan to incorporate semi-supervised learning techniques to reduce the dependency on labeled data while maintaining model accuracy.

References

[1] SINGH A. Review article digital change detection techniques using remotely-sensed data[J]. International Journal of Remote Sensing, 1989, 10(6): 989-1003.

[2] ZHU Q Q, GUO X, DENG W H, et al. Land-use/land-cover change detection based on a Siamese global learning framework for high spatial resolution remote sensing imagery[J]. ISPRS Journal of Photogrammetry and Remote Sensing, 2022, 184: 63-78.

[3] MAHMOUDI S, SARADJIAN M R, ESMAEILY A, et al. Urban expansion monitoring using satellite images by means of decision level fusion of fuzzy change detectors[J]. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 2013, 40: 269-272.

[4] SHI W Z, ZHANG M, KE H F, et al. Landslide recognition by deep convolutional neural network and change detection[J]. IEEE Transactions on Geoscience and Remote Sensing, 2020, 59(6): 4654-4672.

[5] PENG D F, LIU X L, ZHANG Y J, et al. Deep learning change detection techniques for optical remote sensing imagery: Status, perspectives and challenges[J]. International Journal of Applied Earth Observation and Geoinformation, 2025, 136: 104282.

[6] CHENG G L, HUANG Y M, LI X T, et al. Change detection methods for remote sensing in the last decade: A comprehensive review[J]. Remote Sensing, 2024, 16(13): 2355.

[7] DAUDT R C, LE SAUX B, BOULCH A. Fully convolutional siamese networks for change detection[C]//Proceedings of the 2018 25th IEEE international conference on image processing (ICIP). Piscataway, NJ: IEEE, 2018: 4063-4067.

[8] BANDARA W G C, PATEL V M. A Transformer-based siamese network for change detection[C]//Proceedings of the IGARSS 2022—2022 IEEE International Geoscience and Remote Sensing Symposium. Piscataway, NJ: IEEE, 2022: 207-210.

[9] CHEN H R X, SONG J, HAN C X, et al. ChangeMamba: Remote sensing change detection with spatio-temporal state space model[J]. IEEE Transactions on Geoscience and Remote Sensing, 2024, 62: 1-20.

[10] GU A, GOEL K, RÉ C. Efficiently modeling long sequences with structured state spaces[J]. arXiv preprint arXiv:2111.00396, 2021.

[11] GU A, DAO T. Mamba: Linear-time sequence modeling with selective state spaces[J]. arXiv preprint arXiv:2312.00752, 2023.

[12] ZHAO S J, CHEN H, ZHANG X L, et al. RS-Mamba for large remote sensing image dense prediction[J]. IEEE Transactions on Geoscience and Remote Sensing, 2024, 62: 1-14.

[13] LIU Y, TIAN Y J, ZHAO Y Z, et al. Vmamba: Visual state space model[J]. Advances in Neural Information Processing Systems, 2024, 37: 103031-103063.

[14] HUANG T PEI X H, YOU S, et al. LocalMamba: Visual state space model with windowed selective scan[C]//Proceedings of the European Conference on Computer Vision. Cham, Switzerland: Springer, Cham, 2024: 12-22.

[15] WU Z K, MA X W, LIAN R R, et al. CD-Lamba: Boosting remote sensing change detection via a cross-temporal locally adaptive state space model[J]. arXiv preprint arXiv:2501.15455, 2025.

[16] HUANG J Q, YUAN X C, LAM C-T, et al. LCCD-Mamba: Visual state space model for land cover change detection of VHR remote sensing images[J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2025, 18: 5765-5781.

[17] FANG S, LI K Y, SHAO J Y, et al. SNUNet-CD: A densely connected Siamese network for change detection of VHR images[J]. IEEE Geoscience and Remote Sensing Letters, 2021, 19: 1-5.

[18] ZHANG C X, YUE P, TAPETE D, et al. A deeply supervised image fusion network for change detection in high resolution bi-temporal remote sensing images[J]. ISPRS Journal of Photogrammetry and Remote Sensing, 2020, 166: 183-200.

[19] WANG D C, CHEN X N, JIANG M Y, et al. ADS-Net: An attention-based deeply supervised network for remote sensing image change detection[J]. International Journal of Applied Earth Observation and Geoinformation, 2021, 101: 102348.

[20] ZHANG Y, DENG M, HE F, et al. FODA: Building change detection in high-resolution remote sensing images based on feature-output space dual-alignment[J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2021, 14: 8125-8134.

[21] CHEN J, YUAN Z Y, PENG J, et al. DASNet: Dual attentive fully convolutional Siamese networks for change detection in high-resolution satellite images[J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2020, 14: 1194-1206.

[22] DOSOVITSKIY A, BEYER L, KOLESNIKOV A, et al. An image is worth 16×16 words: Transformers for image recognition at scale[J]. arXiv preprint arXiv:2010.11929, 2020.

[23] CHEN H, QI Z P, SHI Z W. Remote sensing image change detection with Transformers[J]. IEEE Transactions on Geoscience and Remote Sensing, 2021, 60: 1-14.

[24] ZHANG C, WANG L J, CHENG S L, et al. SwinSUNet: Pure Transformer network for remote sensing image change detection[J]. IEEE Transactions on Geoscience and Remote Sensing, 2022, 60: 1-13.

[25] LEI T, XU Y T NING H L, et al. Lightweight structure-aware Transformer network for remote sensing image change detection[J]. IEEE Geoscience and Remote Sensing Letters, 2023, 21: 1-5.

[26] CHEN T A, CHEN A T. VisionTwinNet: Gated clarity enhancement paired with light-robust CD Transformers[J]. IEEE Access, 2024, 12: 4544-4560.

[27] ZHANG H T, CHEN K Y, LIU C Y, et al. CD-Mamba: Incorporating local clues into Mamba for remote sensing image binary change detection[J]. IEEE Transactions on Geoscience and Remote Sensing, 2025, 63: 1-16.

[28] DONG Z W, YUAN G J, HUA Z, et al. ConMamba: CNN and SSM high-performance hybrid network for remote sensing change detection[J]. IEEE Transactions on Geoscience and Remote Sensing, 2024, 62: 1-15.

[29] LIU Y C, CHENG G L, SUN Q H, et al. CWMamba: Leveraging CNN-Mamba fusion for enhanced change detection in remote sensing images[J]. IEEE Geoscience and Remote Sensing Letters, 2025, 22: 1-5.

[30] CHEN C W, YU L, MIN S Q, et al. MSVM-UNet: Multi-scale vision mamba unet for medical image segmentation[C]//Proceedings of the 2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). Piscataway, NJ: IEEE, 2024: 3111-3114.

[31] JI S P, WEI S Q, LU M. Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set[J]. IEEE Transactions on Geoscience and Remote Sensing, 2018, 57(1): 574-586.

[32] CHEN H, SHI Z W. A spatial-temporal attention-based method and a new dataset for remote sensing image change detection[J]. Remote Sensing, 2020, 12(10): 1662.

[33] LIU M X, CHAI Z Q, DENG H J, et al. A

CNN-Transformer network with multiscale context aggregation for fine-grained cropland change detection[J]. IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 2022, 15: 4297-4306.

[34] ZHU D W, HUANG X H, HUANG H Y, et al. ChangeVit: Unleashing plain vision Transformers for change detection in remote sensing images[J]. Pattern Recognition, 2026, 172: 112539.

Acknowledgements This work was supported by the National Natural Science Foundation of China (Nos.42371449, 41801386).

Author

The first/corresponding author Dr. **PENG Daifeng** received the Ph.D. degree in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 2017. He is currently an associate professor with School of Remote Sensing & Geomatics Engineering, Nanjing University of Information Science and Technology, Nanjing, China. From 2020 to 2021, he was a post-doctoral fellow with Remote Sensing Laboratory, Department of Information Engineering and Computer Science, University of Trento, Trento, Italy. His research interests include machine learning, deep learning, multi-temporal image analysis, intelligent interpretation and information extraction from high resolution remote-sensing imagery.

Author contributions Dr. **PENG Daifeng** designed the study, complied the models, conducted the analysis, interpreted the results, and wrote the manuscript. Mr. **DONG Fengxu** complied the models, conducted the visualization, and wrote the manuscript. Prof. **GUAN Haiyan** contributed to the discussion and background of the study. All authors commented on the manuscript draft and approved the submission.

Competing interests The authors declare no competing interests.

(Production Editor: SUN Jing)

DFFMamba:一种基于差异特征融合Mamba的新型遥感影像变化检测方法

彭代锋，董峰旭，管海燕

(南京信息工程大学遥感与测绘工程学院,南京 210044,中国)

摘要:变化检测(Change detection, CD)在多个领域发挥着关键作用,其中卷积神经网络(Convolutional neural networks, CNNs)与Transformer均在变化检测任务中展现出卓越性能。然而,CNNs受限于感受野范围,难以捕捉全局特征,而Transformer则受限于高计算复杂度。近年来,基于状态空间模型(State space models, SSMs)的Mamba架构展现了强大的全局建模能力,同时实现了线性计算复杂度。尽管已有研究将Mamba引入变化检测任务,但现有基于Mamba的遥感变化检测方法在展平与扫描遥感影像时,难以有效感知变化区域固有的局部性,限制了变化特征提取能力。为解决上述问题,本文针对传统Mamba扫描方式导致的特征局部性丢失问题,提出一种基于差异特征融合Mamba的新型遥感影像变化检测方法(Difference feature fusion Mamba, DFFMamba)。具体来说,本文设计了两种差异特征提取模块:差异Mamba(Difference Mamba, DMamba)与局部差异Mamba(Local difference Mamba, LDMamba)。其中DMamba通过计算双时相特征状态空间方程中系数矩阵的差值来提取差异特征;在此基础上,LDMamba结合局部自适应状态空间扫描(Locally adaptive state-space scanning, LASS)策略增强了特征局部性,实现了差异特征的精准提取。此外,本文提出融合Mamba(Fusion Mamba, FMamba)模块,该模块采用空间-通道序列建模SSM(Spatial-channel token modeling SSM, SCTMS)机制,整合了变化特征的多维时空交互,捕捉了其在空间与通道维度的依赖关系。为验证DFFMamba的有效性,本文在WHU-CD、LEVIR-CD和CLCD这3个数据集上进行了广泛实验。结果表明,DFFMamba显著优于现有的最优变化检测方法,在3个数据集上的交并比(Intersection over union, IoU)分数分别达到90.67%、85.04%和66.56%。

关键词:变化检测;状态空间模型;变化特征融合;深度学习;差异Mamba;局部差异Mamba;空间-通道序列建模SSM