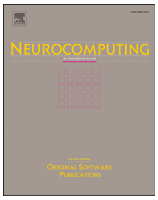
Neurocomputing 707 (2027) 135117

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom



SO-Mamba: Sub-task oriented Mamba-centric network for change detection of remote sensing image

Chunyan Yu, Zhenhao Zhang, Qiang Zhang*, Yulei Wang, Haoyang Yu

Center of Hyperspectral Imaging in Remote Sensing (CHIRS), Information Science and Technology College, Dalian Maritime University, Dalian, China

ARTICLE INFO

Communicated by V. Conti

Keywords:
Remote sensing image
Change detection
Mamba-centric
Task-specific characteristics
Sub-task oriented Mamba

ABSTRACT

Remote sensing image (RSI) change detection (CD) identifies surface changes by analyzing multi-temporal high-resolution images, requiring deep learning models with efficient long-range representation capabilities to handle the discrete distribution of targets. While the Mamba architecture has attracted growing attention in CD owing to its linear computational complexity, current Mamba-based methods remain largely architecture-oriented, concentrating on improving the general capability of Mamba without fully exploiting the task-specific characteristics of CD at different processing stages. To bridge this gap, we propose the sub-task oriented Mamba-centric network (SO-Mamba), which reframes the design philosophy from enhancing Mamba to aligning Mamba with the intrinsic requirements of CD. Specifically, SO-Mamba decomposes the CD task into three complementary subtasks corresponding to the progressive stages, and establishes dedicated Mamba-based modules that adapt the sequential modeling capacity of Mamba to the distinct functional requirements of each stage. The Mamba-centric interaction module constructs dual-branch asymmetric residual structures within the Mamba framework to guide global interactions of dual-temporal features. The Mamba-centric edge-focused difference module incorporates the proposed frequency-enhanced Mamba and edge operators to extract difference information enriched with edge characteristics. The Mamba-centric attention reconstruction module integrates dual-attention mechanisms with the long-sequential representation mechanism to dynamically reconstruct targets with refined feature details. Experiments on three benchmark datasets demonstrate that SO-Mamba achieves significant performance improvements over existing methods, confirming that task-level specialization represents a more effective paradigm than general architectural enhancement for Mamba-based change detection.

1. Introduction

Remote sensing image (RSI) change detection (CD) identifies the dynamic changes in surface objects by analyzing differential features in multi-temporal high-resolution RSIs, holding irreplaceable value in urban planning, disaster assessment, and ecological monitoring. With the development of deep learning technology, the CD task is transitioning from artificial feature engineering [1,2] to data-driven learning [3,4]. In the context of data-driven learning detection paradigms, convolutional neural networks (CNNs) [5,6] and Transformers [7,8] have emerged as the prominent architectures for local feature extraction and global dependency representation for CD task, respectively.

Traditional CNNs have emerged as the cornerstone architecture for local feature capture for CD task [9,10]. Typically, Daudt et al. [11] first applied the fully convolutional networks [12] to the CD task, addressing the bottlenecks in accuracy and efficiency of traditional CD, thereby positioning CNNs as a critical component in CD networks. A deeply supervised image fusion network [13] proposed hierarchical feature extraction with dual-branch CNNs and deeply supervised CNN training mechanism, further exploring the application of CNNs in CD task. Nevertheless, constrained by the local receptive field characteristics of CNNs, CNN-based CD models exhibit limited performance in global contextual information representation.

Transformer-based architectures [14–18] have achieved significant global context-awareness through the self-attention mechanism. Typically, ChangeFormer [19] transcended the limitations of traditional fully convolutional networks by unifying hierarchical Transformer encoders with MLP decoders. BIT [20] proposed a dual-temporal image Transformer, which achieved effective representation of contextual information in the spatial-temporal domain, and further explored the application of Transformer models in CD tasks. Due to the quadratic computational complexity of self-attention mechanism, transformer-based

* Corresponding author.
Email address: qzhang95@dlmu.edu.cn (Q. Zhang).

https://doi.org/10.1016/j.neucom.2026.135117
Received 20 October 2025; Received in revised form 6 June 2026; Accepted 12 September 2026
Available online 14 September 2026

CD architectures encounter substantial processing costs in handling high-resolution RSIs.

Recently, the Mamba architecture [19] based on state space model (SSM) [21] offers a new paradigm for parsing high-resolution RSIs, as it maintains linear computational efficiency while capturing long-range dependencies. RSI change targets typically exhibit discrete distribution characteristics. The selective state space mechanism of Mamba adapts state transition parameters related to each pixel in the sequence, and processes the dual-temporal fusion sequence using internal hidden states as dynamic spatiotemporal memory. In this way, the Mamba-based model autonomously determines whether to preserve genuine change signals or mitigate interference caused by false changes. Besides, CD demands high edge integrity and contour continuity in targets. The long-sequence scanning mechanism of Mamba effectively mimics continuous pixel-level modeling and improves the accuracy of edge detection in RSI change detection tasks. ChangeMamba [22] first introduced the Mamba framework into the CD task and designed a spatial-temporal joint structure to facilitate cross-temporal feature interaction. ConMamba [23] initially explored the effectiveness and prospects of a hybrid CNN-SSM architecture in the CD task. Although current Mamba-based CD methods [24–28] have extensively explored architectural refinements, they tend to lack sufficient attention to the distinctive characteristics of CD tasks.

In this paper, we identify distinct task characteristics across different processing stages of CD and decompose the CD task into three sub-tasks, including dual-temporal feature interaction, difference information extraction, and target detail reconstruction. Accordingly, we develop the sub-task oriented Mamba-centric modules to improve the adaptability of Mamba to characteristics of CD task. Firstly, the proposed Mamba-centric interaction module (MIM) develops dual-branch asymmetric residual blocks by extending selective interaction of Mamba, achieving multistage profound dual-temporal feature interaction. Secondly, the proposed Mamba-centric edge-focused difference module (MEDM) unifies the frequency-enhanced Mamba with the edge-constrained difference block, mitigating contour blurring in difference feature extraction. Thirdly, the proposed Mamba-centric attention reconstruction module (MARM) integrates spatial and channel attention within sequential representation framework of Mamba, dynamically refining detail reconstruction. Based on the three sub-task oriented Mamba-centric modules, we construct the U-shaped architecture, forming the SO-Mamba network. The main innovations are listed as follows:

- Architecturally, we propose a novel sub-task oriented Mamba-centric network named SO-Mamba to prioritize the task-specific characteristics of CD. Specifically, we decompose the CD task into three sub-tasks including dual-temporal feature interaction, difference information extraction, and target detail reconstruction. Accordingly, we present three sub-task oriented Mamba modules to align with the traits of each sub-task. Besides, we build the U-shaped architecture that integrates the three modules into a cohesive network to enhance the accuracy of the CD.
- The presented Mamba-centric interaction module establishes a multi-stage interaction architecture through dual-branch asymmetric residual blocks by incorporating Mamba block. In the MIM, we extract temporally correlated information with designed Siamese network. Importantly, long-range interaction information is connected through the selective Mamba part to achieve profound global interaction between dual-temporal RSIs. Besides, we refine global features guided by temporally independent structure which is beneficial for change information extraction.
- The presented Mamba-centric edge-focused difference module constructs a frequency-enhanced Mamba structure with double-order edge operators to extract edge-constrained difference features. The edge-aware mechanism fully develops the ability of Mamba for capturing detailed information, thereby mitigating the loss of edge details during the difference extraction.
- The presented Mamba-centric attention reconstruction module architects dual-attention mechanisms within Mamba to refine detail reconstruction across spatial and channel dimensions. The dual-dimensional refinement explicitly expands the receptive dimension of sequence-based representation mechanism of Mamba, thereby enhancing the accuracy of target change generation significantly.

2. Related work

2.1. CNN-based & transformer-based CD models

For local feature extraction in CD tasks, CNN-based architectures represent the dominant methodology. Among CNN-based approaches, the fully convolutional Siamese network (FC-EF) models [11] constituted the pioneering application of deep learning to change detection tasks. SEIFNet [29] systematically resolved challenges in detecting change omissions and edge ambiguity for conventional CNN methods within complex scenarios. The method employed spatiotemporal feature enhancement and cross-level feature fusion. To address the problem of spatial information loss due to continuous downsampling, SNUNet-CD [30] resolved edge recognition ambiguity via dense skip-level concatenations, preserving high-resolution details in the UNet encoder–decoder architecture [31]. Nevertheless, traditional CNNs are mainly optimized for local feature extraction, ignoring the need for overall scene context understanding in the CD task.

For long-range feature representation in CD tasks, Transformer-based methods constitute the prevailing technique. Consistent with the trend, ChangeFormer [17] transcended the limitations of traditional fully convolutional networks by synergizing hierarchical Transformer encoders with MLP decoders. SwinSUNet [32] adopted the Transformer structure and designed the double U-shaped architecture to enhance the capture of the global field of view. DMINET [33] demonstrated the potential of multi-scale feature fusion strategies in change detection (CD) tasks by integrating self-attention and cross-attention mechanisms to suppress interference and focus on change regions. Nevertheless, these pure Transformer models pay insufficient attention to local information.

CNN-based methods focus on local features, while Transformer-based methods address long-range dependencies, leading to the emergence of hybrid CNN-Transformer architectures [34,35]. BIT [20] integrated the global modeling capabilities of Transformers with the local feature extraction advantages of CNNs, which systematically resolved limitations in spatial-temporal feature modeling within CNN architectures and boundary ambiguity in Transformer architectures. ACABFNet [36] achieved feature representation from dual dimensions through a bidirectional fusion approach that integrated local and global image information from CNN and Transformer branches. Nevertheless, the CNN-Transformer hybrid architectures face a performance bottleneck due to the inherent secondary computational complexity of Transformers.

2.2. Mamba-based CD models

Recent studies have shown that Mamba exhibits high efficiency in long distance modeling and maintains linear computational complexity. Due to its ability to capture long-range dependencies and the mentioned advantages of Mamba in the Introduction part, the application of Mamba model in CD of RSI has achieved significant progress. ChangeMamba [22] introduced the Mamba model into remote sensing image CD task for the first time, and demonstrated the efficiency and adaptability of the Mamba model in analyzing the changes in time-series RSIs. RS-Mamba [25] proposed the combination forward and backward scanning mechanism and successfully improved the performance of the Mamba model in CD. CDMamba [26] effectively integrated global and local information based on Mamba and addressed the lack of local cues in Mamba when handling dense prediction tasks.

Existing Mamba-based CD methods including ChangeMamba, RS-Mamba, and CDMamba, share a common design philosophy that can be

characterized as architecture-oriented. They focus on enhancing the general modeling capabilities of Mamba (e.g., scanning strategies, hybrid architectures, global-local fusion) and apply the improved architecture to the CD task. To further unleash the advantages of Mamba in CD tasks and enhance the adaptability of Mamba to CD tasks, this paper proposes an innovative sub-task oriented Mamba architecture, SO-Mamba, which systematically decomposes the CD task into complementary sub-tasks and designs corresponding Mamba-centric modules. The task-aligned optimization strategy enhances the capacity of Mamba to analyze RSIs with high-resolution spatial details, thereby optimizing the performance of Mamba-based networks in CD.

3. Methodology

3.1. Architecture overview

Fig. 1 illustrates the flowchart of the proposed SO-Mamba network. Firstly, as shown in the left part of Fig. 1, we decompose the CD task into three preliminary sub-tasks including dual-temporal feature interaction, difference information extraction, and target detail reconstruction. Correspondingly, three specified Mamba-centric modules are established to be associated with the mentioned sub-tasks separately in the right part of Fig. 1. Based on the three presented modules, we propose the three-layer U-shaped SO-Mamba network to coherently unify the task-driven Mamba-centric modules to implement the CD task.

3.2. Mamba-centric interaction module

For the first sub-task, we present the MIM with the detailed structure illustrated in Fig. 2. In MIM, the weight sharing blocks first roughly extract the correlated features in the dual-temporal RSIs. Subsequently, the weight independent blocks inject the globally interacted information back into each temporal branch. Crucially, the Mamba block (MB) is built upon the selective state space block of VMamba [37] and performs cross-temporal feature mixing through hidden-state propagation of the SSM over the concatenated dual-temporal sequence.

Initially, fed with the dual-temporal images T_a and T_b, the weight sharing blocks extract features with correlated characteristics efficiently. In the implementation, the blocks consist of a normalization layer (LN), a linear layer (Linear), and a convolutional layer (Conv). Within the blocks, the residual connections are employed to yield refined features $\bar{T}_a$ and $\bar{T}_b$. To aggregate the refined features from both branches, the concatenation block (CB), as detailed in Fig. 3, functions as a parameter-efficient bottleneck module. Specifically, the CB executes a four-fold channel reduction followed by a two-fold expansion. The bottleneck structure mitigates model complexity and ensures the effective aggregation of cross-temporal features. The mathematical description of the fusion process is as follows:

$$\bar{T}^1 = \mathrm{CB}(\mathrm{Cat}(\bar{T}_a, \bar{T}_b)) \tag{1}$$

where $\bar{T}^1$ denotes the fused features, and $\bar{T}_a$, $\bar{T}_b$ represent the refined features from the dual branches. The notations $\mathrm{Cat}(\cdot)$ and $\mathrm{CB}(\cdot)$ signify the channel-wise concatenation operation and the bottleneck transformation performed by the CB, respectively.

Next, in the MB, the $\bar{T}^1$ is initialized through the operations of LN, Linear and depthwise separable convolution (DWConv). In the cross-scan layer, we adopt the four conventional scanning paths along the horizontal forward, horizontal backward, vertical forward, and vertical backward directions to generate four complementary one-dimensional sequences, with each sequence denoted by a distinct x. Crucially, the selective state space block (S6) [38] dynamically projects the x onto the matrices Δ, B, C via the linear layer, enabling the state space to achieve an input-dependent representation. The dynamic mechanism imparts context-awareness to the state information and enables y to capture relevant features in x over long ranges. Afterwards, the cross-merge layer [37] dynamically integrates y derived from diverse scan paths. The integration ensures that the global representation $\bar{\bar{T}}$ preserves spatial

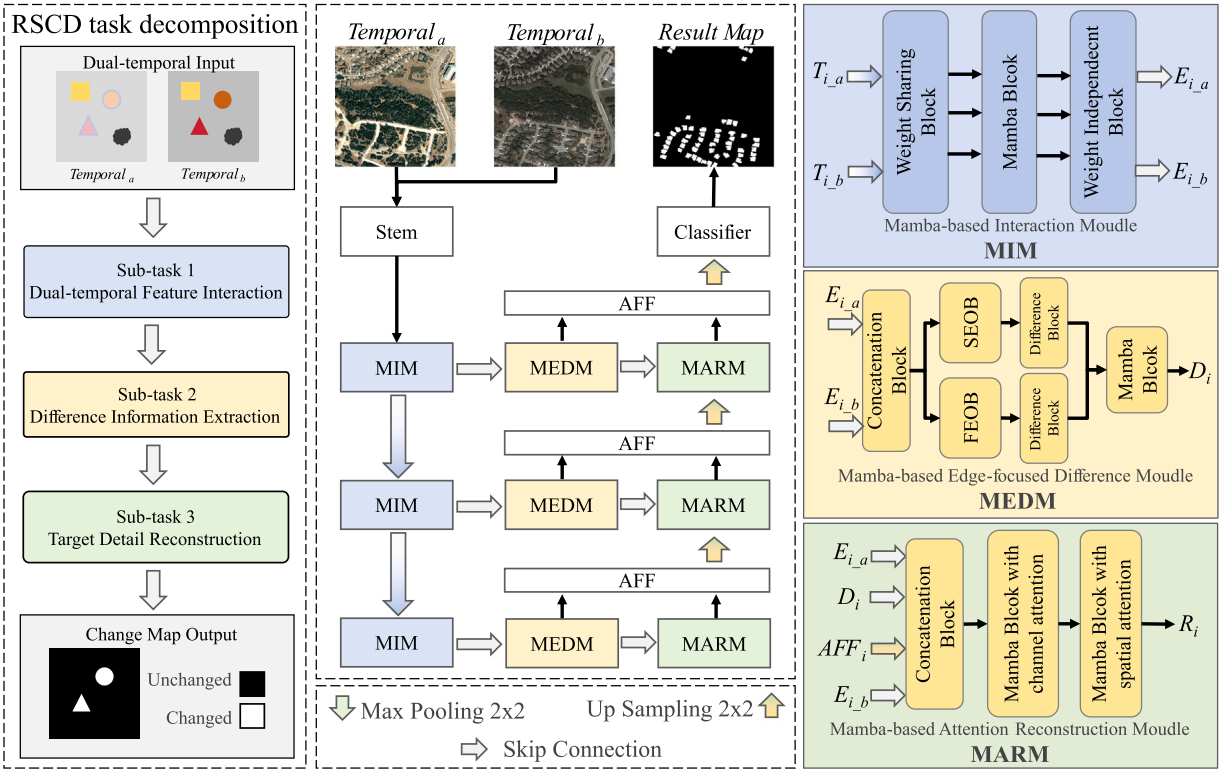


Fig. 1. Schematic of the structure of SO-Mamba. The left panel shows the task decomposition, the center panel presents the model framework, the right panel details the corresponding task modules, and the bottom contains the figure annotations.

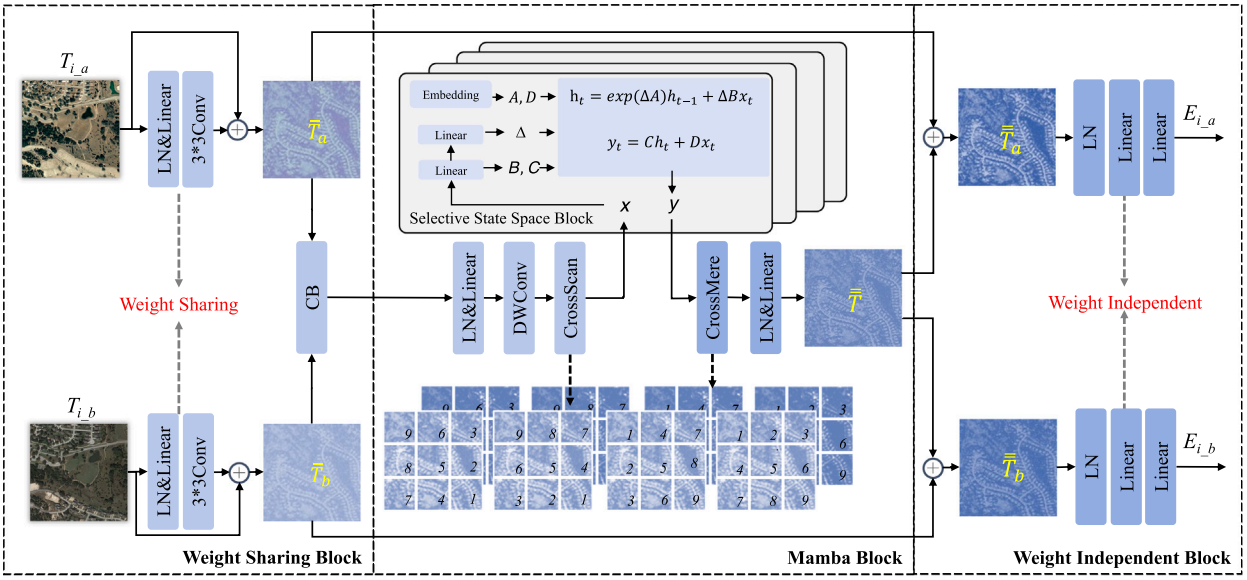


Fig. 2. Schematic of the structure of MIM. The weight sharing block aims to extract the temporal features separately and the weight independent block handles the deep fusion of the acquired features. Significantly, the MB block employs the selective scanning mechanism to facilitate deep global interactions between dual-temporal features over long distances.

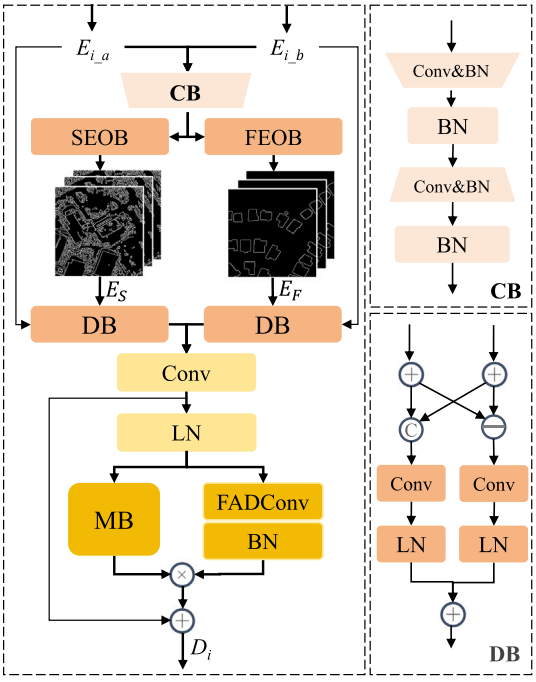


Fig. 3. Schematic of the structure of MEDM. The SEOB and FEOB are responsible for edge extraction, while the MB part globally optimizes the difference features obtained from the DB to strengthen the edge continuity of the changed regions.

proximity of object features in RSIs. The mathematical description of the processing in MB is given as follows:

$$\bar{T}_i^2 = \text{CrossScan}(\text{DWConv}(\text{Linear}(\text{LN}(\bar{T}^1)))) \tag{2}$$

$$\bar{T}_i^3 = \text{S6}\left(\bar{T}_i^2\right) \tag{3}$$

where $\bar{T}_i^2$ and $\bar{T}_i^3$ denote the feature sequences generated by the cross-scanning and selective state space modeling processes, respectively. The index $i \in \{1, 2, 3, 4\}$ identifies the four complementary scanning paths. The operators LN(·), Linear(·), and DWConv(·) signify layer normalization, linear projection, and depthwise separable convolution. Furthermore, CrossScan(·) and S6(·) represent the spatial-to-sequence scanning operation and the selective state space modeling execution.

$$\overline{\overline{T}} = \text{Linear}\left(\text{LN}\left(\text{CrossMerge}\left(\bar{T}_1^3, \bar{T}_2^3, \bar{T}_3^3, \bar{T}_4^3\right)\right)\right) \tag{4}$$

where $\overline{\overline{T}}$ refers to the reconstructed global 2D feature map. The function CrossMerge(·) performs the sequence-to-spatial aggregation of the four state sequences, followed by the subsequent normalization and projection operations to yield the final representation.

Ultimately, the weight independent blocks connect the $\overline{\overline{T}}$ to the original features $\bar{T}_a$ and $\bar{T}_b$, respectively. The residual connections mitigate feature information loss to produce the temporally independent representations $\overline{\overline{T}}_a$ and $\overline{\overline{T}}_b$. Within the blocks, Linear layers expand channel dimensions to enrich feature semantics. Through the fusion of these features, the MIM synthesizes cross-temporal context with specific temporal information to form a precise representation of correlated features. The overall procedure of MIM is summarized in Algorithm 1.

3.3. Mamba-centric edge-focused difference module

For the second sub-task, we build the MEDM as shown in Fig. 3. In MEDM, we adopt frequency-adaptive dilated convolution (FADConv) [39] to augment parsing ability of MB for high frequency features, while maintaining advantage of modeling long sequences of Mamba. Additionally, the architecture incorporates the first-order edge operator block (FEOB) and the second-order edge operator block (SEOB) to impose constraints on the inputs of the difference block (DB).

Specifically, when fed with the outputs from DB, we adopt MB as the trunk to enable salient information extraction from complementary difference representations. The context-adaptive selection mechanism inherent in Mamba provides distinct advantages for capturing edge continuity and orientation, and consequently enhances capacity of MEDM to extract comprehensive image details of difference features. To enhance the parsing capability of Mamba for high-frequency edge features, we replace the linear layer of Mamba with FADConv. The dynamic adjustment of the receptive field size enables FADConv to capture high-frequency edge features through frequency-aware dilated convolutions.

Algorithm 1 For Mamba-centric interaction module (MIM).

Input: Dual-temporal images $T_a \in \mathbb{R}^{H\times W\times C}$, $T_b \in \mathbb{R}^{H\times W\times C}$

Output: Temporally independent representations $\overline{\overline{T}}_a$, $\overline{\overline{T}}_b$

% Stage 1: Weight Sharing Blocks — Extract correlated features

$\bar{T}_a \leftarrow \text{Conv}(\text{Linear}(\text{LN}(T_a))) + T_a$ ▷ Shared-weight residual branch

$\bar{T}_b \leftarrow \text{Conv}(\text{Linear}(\text{LN}(T_b))) + T_b$ ▷ Shared-weight residual branch

% Stage 2: Concatenation Block (CB) — Bottleneck fusion

$\bar{T}_1 \leftarrow \text{CB}\big(\text{Cat}(\bar{T}_a, \bar{T}_b)\big)$ ▷ 4× reduction → 2× expansion

% Stage 3: Mamba Block (MB) — Selective state space interaction

% 3a. Feature projection and Cross-Scan

for $i = 1, 2, 3, 4$ **do** ▷ Four complementary scanning paths

  $\bar{T}_i^{\,2} \leftarrow \text{CrossScan}_i\big(\text{DWConv}(\text{Linear}(\text{LN}(\bar{T}_1)))\big)$

end for

% 3b. Selective State Space Block (S6) — Input-dependent modeling

$A \leftarrow \text{Embedding}_A$ ▷ Learnable state transition parameter

$D \leftarrow \text{Embedding}_D$ ▷ Learnable skip connection parameter

for $i = 1, 2, 3, 4$ **do**

  $B_i,\ C_i,\ \Delta_i \leftarrow \text{Linear}(\bar{T}_i^{\,2})$ ▷ Input-dependent projection

  $\bar{A}_i \leftarrow \exp(\Delta_i \cdot A)$ ▷ Discretized state transition

  $\bar{B}_i \leftarrow \Delta_i \cdot B_i$ ▷ Discretized input matrix

  $h_t \leftarrow \bar{A}_i\, h_{t-1} + \bar{B}_i\, x_t$ ▷ Recurrent hidden state update

  $y_t \leftarrow C_i\, h_t + D\, x_t$ ▷ Output with skip connection

  $\bar{T}_i^{\,3} \leftarrow y_t$

end for

% 3c. Cross-Merge — Sequence-to-spatial aggregation

$\overline{\overline{T}} \leftarrow \text{Linear}\big(\text{LN}\big(\text{CrossMerge}(\bar{T}_1^{\,3}, \bar{T}_2^{\,3}, \bar{T}_3^{\,3}, \bar{T}_4^{\,3})\big)\big)$

% Stage 4: Weight Independent Blocks — Temporal-specific injection of mixed features

$\overline{\overline{T}}_a \leftarrow \text{Linear}(\text{LN}(\overline{\overline{T}})) + \bar{T}_a$ ▷ Residual with temporal-a features

$\overline{\overline{T}}_b \leftarrow \text{Linear}(\text{LN}(\overline{\overline{T}})) + \bar{T}_b$ ▷ Residual with temporal-b features

return $\overline{\overline{T}}_a$, $\overline{\overline{T}}_b$

Consequently, the proposed integration maintains the long-range modeling efficiency of Mamba while achieving frequency-adaptive feature extraction. The convolution operation for FADConv is mathematically defined as follows:

$$Y(p) = \sum_{i=1}^{K*K} W_i' \, X(p + \Delta p_i \times \hat{D}(p)) \tag{5}$$

where $Y(p)$ denotes the pixel value at position p in the output feature map, K is the convolution kernel size, W_i' is the dynamic frequency-adaptive weight parameter of the convolution kernel, $X(p+\Delta p_i)$ denotes the pixel value at position p offset by Δp_i, and $\hat{D}(p)$ is the dilation rate for each pixel, dynamically adjusting to expand the receptive field.

Besides, we construct FEOB and SEOB to impose boundary constraints on E_S and E_F. Specifically, in the FEOB, we employ the Scharr [40] operators as convolution kernels for first-order derivative edge detection in the convolution layers of FEOB. In the specific implementation, we extend the traditional Scharr operator to four directions: 45°, 90°, 135°, and 180° relative to the vertical and horizontal axes, which improves edge connectivity through orthogonal gradient directions that complement the dominant contour orientations. In the SEOB, the difference of Gaussians (DoG) [41] operator functions as the core mechanism for second-order derivative edge detection. This design facilitates the efficient approximation of second-order derivatives through low-cost first-order operations. Specifically, the module utilizes Eqs. (6) and (7) to initialize two distinct operators and applies Gaussian filters with weights σ_1 and σ_2 for smoothing. The subsequent subtraction of features extracted by these two Gaussian operators yields second-order boundary features sensitive to abrupt image changes. The mathematical

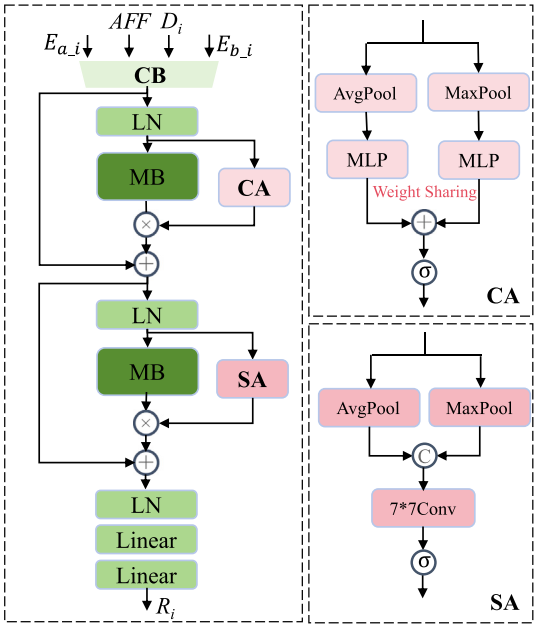


Fig. 4. Schematic of the structure of MARM. At the top of the module are the adaptive feature fusion (AFF) and the CB. The CA and SA constrained MB parts refine the sequence features extracted by Mamba across different dimensions to enhance the details of the reconstructed change regions.

formulation of the DoG operator is defined as follows:

$$(m, n) = \text{meshgrid}([-1, 0, 1], [-1, 0, 1]) \tag{6}$$

$$\text{Gauss}(\sigma) = \frac{1}{2\pi\sigma^2} \mathrm{e}^{-\frac{m^2+n^2}{2\sigma^2}} \tag{7}$$

$$\text{DoG} = \text{Gauss}(\sigma_1) - \text{Gauss}(\sigma_2) \tag{8}$$

where m and n represent the horizontal and vertical coordinate matrices generated by the meshgrid$(\cdot)$ function. The function Gauss$(\cdot)$ defines the Gaussian distribution operation, with σ_1 and σ_2 denoting the Gaussian weights. The DoG signifies the difference of Gaussians operator derived from the subtraction of the two Gaussian operators.

3.4. Mamba-centric attention reconstruction module

For the third sub-task, we present the MARM with the detailed structure illustrated in Fig. 4. As shown in the green block, we build MARM upon the long-range sequential representation mechanism of MB for feature reconstruction. Notably, we employ dual-attention blocks to dynamically and adaptively adjust feature importance across spatial and channel dimensions, which are distinct from the sequence representation perspective of Mamba. With the Mamba block, MARM integrates a dual attention mechanism by adjusting feature weights to enhance the reconstruction of small and weak targets.

The channel attention (CA) branch employs a parallel pooling strategy. MaxPool extracts prominent local features, while AvgPool encodes global contextual information. Subsequently, a shared multi-layer perceptron (MLP) processes the pooled features. The CA module sums the MLP outputs and applies a Sigmoid function to generate normalized weights. Finally, the generated channel weights modulate the input features from the MB through element-wise multiplication. The cross-channel refinement enhances discriminative representation of MB via dynamic channel-weight recalibration. Unlike CA, the spatial attention (SA) branch integrates MaxPool and AvgPool along the channel dimension to obtain spatial attention weights through pooling, yielding two single-channel features. The SA employs a large-kernel

convolutional operation to fuse the two spatial weights into a single-channel attention map. Following multiplication of the spatial weights with the reconstructed features from the second-stage MB, the spatial weights achieve pixel-level emphasis on regions with significant temporal changes while suppressing background noise. The mathematical formulation of CA and SA is defined as follows:

$$W_{CA} = \text{Sigmoid}\left(\text{MLP}\left(\text{AvgPool}\left(F_1\right)\right) + \text{MLP}\left(\text{MaxPool}\left(F_1\right)\right)\right) \tag{9}$$

where F_1 denotes the input features of the first stage, and $W_{CA} \in \mathbb{R}^{C\times1\times1}$ represents the generated channel attention weights. The operators AvgPool(·) and MaxPool(·) correspond to average pooling and max pooling, respectively. MLP(·) signifies the multi-layer perceptron processing, and Sigmoid(·) denotes the activation function used for weight normalization.

$$W_{SA} = \text{Sigmoid}\left(\text{Conv}\left(\text{Cat}\left(\text{AvgPool}\left(F_2\right) + \text{MaxPool}\left(F_2\right)\right)\right)\right) \tag{10}$$

where F_2 denotes the input features of the second stage, $W_{SA} \in \mathbb{R}^{1\times H\times W}$ represents the spatial attention weights, and H, W denote the height and width of the input features.

3.5. Network supplement

3.5.1. Attention feature fusion

To align the features of the difference information extraction task and the target detail reconstruction task, we integrate the AFF as the dynamic fusion module to fuse the outputs of MEDM and MARM. The specific structure of the AFF is shown in Fig. 5. Specifically, the output information of MEDM and MARM is preliminarily fused by two convolutional layers to enhance the feature representation. To enhance the nonlinearity of the features, the tanh activation function with an output distribution of $[-1, 1]$ and zero symmetry is applied to the fused features. The fused features are processed by asymmetric modulation gates $+1$ and -1, respectively. Subsequently, the complementary modulation weights are element-wise multiplied by the original input features. The architecture establishes a bidirectional adaptive modulation mechanism.

3.5.2. Architecture details

In the implementation, the MIM module consists of three hierarchical stages characterized by channel configurations of 64, 128, and 256, with corresponding MB counts of 2, 2, and 10 per stage. The patch embedding layer is constructed from two successive convolutional layers with a kernel size of 3 and a stride of 2 to achieve initial feature extraction and a fourfold reduction in spatial resolution. Each MB incorporates a selective state space module and dual parallel MLP branches with an expansion ratio of 4. The internal configuration of the selective state space module includes an input projection layer with a channel expansion factor of 2, a depthwise convolutional layer with a kernel size of 3, and a SiLU activation function. A fusion component with a kernel size of 1 is integrated within the SSM to facilitate deep dual-temporal interaction. Besides, the downsampling between consecutive stages is performed by convolutional layers with a kernel size of 3 and a stride of 2 to ensure feature stability during the resolution reduction process. In the MEDM and MARM modules, the channel dimensions and feature map sizes are aligned with those of the MIM module. The layer counts for MEDM and MARM are set to 1 for each stage, while the channel scaling factor for the CB module is fixed at 0.25 to reduce computational complexity. The specific configuration supports the concatenation of convolutional layers for the integration of dual temporal feature information.

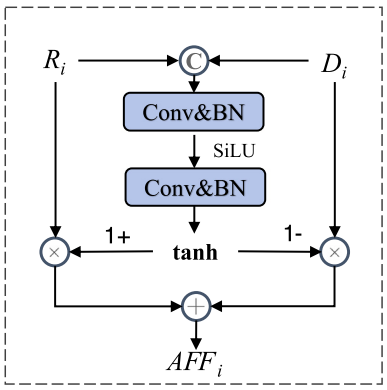


Fig. 5. Schematic of the structure of AFF.

3.5.3. Loss function

In this paper, we employ cross-entropy loss and Lovasz-softmax loss [42]. Lovasz-softmax loss directly optimizes the evaluation metric of intersection-over-union (IoU) by combining Lovász and Softmax, which effectively addresses the issue of pixel imbalance in the CD task. The joint loss is defined as below:

$$L_{\text{total}} = L_{\text{ce}} + \lambda * L_{\text{lov}} \tag{11}$$

where L_{lov} is the Lovasz-softmax loss, λ is the joint coefficient. In our paper, we set the parameter to 0.75 based on ablation experiment comparisons (4). The mathematical expression for cross-entropy loss (L_{ce}) follows:

$$L_{\text{ce}} = -\frac{1}{N}\sum_{i=1}^{N} y_i \log\left(\hat{y}_i\right) \tag{12}$$

where y_i denotes the truth value of the ith pixel, $\hat{y}_i$ denotes the predicted probability, and N denotes the total number of pixels.

4. Experimental results and analysis

4.1. Description of the datasets

LEVIR-CD [43] contains 637 pairs of bitemporal high-resolution images, which cover periods ranging from 5 to 14 years between acquisitions. Each image has a ground sample distance of 0.5 meters per pixel and dimensions of 1024 × 1024 pixels. The dataset contains a wide range of building types, such as villa houses, high-rise apartments, small garages, and large warehouses. The dataset exhibits pronounced seasonal variations due to changes in acquisition times and lighting conditions, which create interference that poses a challenge for change detection. In our paper, we crop the LEVIR-CD images to 256 × 256 pixels and randomly divide the images into training set (7120 images), validation set (1024 images) and test set (2048 images).

WHU-CD [44] provides the detailed information on many buildings in Christchurch, New Zealand from 2012 to 2016. The image resolution is 0.075 m/pixel and the image size is 32,507 × 15,354 pixels. The WHU-CD dataset primarily focuses on building changes such as demolition and reconstruction over relatively short time periods, featuring relatively simple scenes. In our paper, the dataset was partitioned into 8189 image blocks of 256 × 256 pixels and split into a training set (6096 images), validation set (762 images), and test set (762 images).

GZ-CD [45] covers the detailed information on Guangzhou, China, from 2006 to 2019. Comprising 19 pairs of high-resolution images, the dataset offers a spatial resolution of 0.55 meters per pixel. Unlike the previous two datasets, the GZ-CD dataset features lower image resolution, thereby enabling the evaluation of the capability of network to parse blurry images. In our paper, the GZ-CD dataset was preprocessed to 256 × 256 pixels and the images were divided into training set (2834 images), validation set (400 images) and test set (325 images).

The information on these three datasets is summarized in Table 1.

Table 1
The information is summarized from three datasets.

Dataset	Number of image pairs	Size (pixels)	Resolution (m)	Experimental size (pixels)	Number of experimental sets (train:val:test)
LEVIR-CD	637	1024 × 1024	0.5	256 × 256	7120:1024:2024
WHU-CD	1	32,508 × 15,354	0.3	256 × 256	6096:762:762
GZ-CD	19	256 × 256	0.55	256 × 256	2834:400:325

4.2. Experimental setup

4.2.1. Experimental parameter settings

We conducted all experiments on an Ubuntu 20.04 system equipped with an NVIDIA RTX 5080 GPU, using CUDA 12.8 and PyTorch 2.8.0 to ensure sufficient computational resources. For the basic setup, we set the number of training epochs to 200 and the batch size to 16. For training optimization, we employed the AdamW optimizer and a learning rate decay strategy, initializing the learning rate at 0.0001. The learning rate decreased to 0.1 times its current value when the F1 metric failed to improve after 15 consecutive epochs of training.

4.2.2. Evaluation metrics

We select five core evaluation metrics for performance evaluation, including precision (P), recall (R), F1-score (F1), overall accuracy (OA) and intersection over union (IOU). These metrics are combined and applied to provide the comprehensive assessment of the capabilities of the SO-Mamba model, with F1 and IOU as the most important measures of the capabilities of model. The mathematical definitions of the relevant assessment metrics are as follows:

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

$$\text{Recall} = \frac{TP}{TP + FN} \tag{14}$$

$$F1 = \frac{2TP}{2TP + FP + FN} \tag{15}$$

$$OA = \frac{TP + TN}{TP + TN + FP + FN} \tag{16}$$

$$IOU = \frac{TP}{FP + FN + TP} \tag{17}$$

where TP stands for true positives, TN stands for true negatives, FP stands for false positives, and FN stands for false negatives.

4.3. Performance comparison

To validate the effectiveness of SO-Mamba for the change detection task, we selected benchmark methods for comparison. The methods included CNN-based methods: IFNet [13], SNUNet [30], SEIFNet [29], Transformer-based methods: ChangeFormer [17], SwinSUNet [32], BIT [20], DMINet [33] and Mamba-based methods: ChangeMamba [22], RS-Mamba [25], CDMamba [26], CDLamba [46]. To ensure a fair comparison, we strictly adhered to the original parameter settings reported in the papers when reproducing the baseline methods.

4.3.1. LEVIR-CD dataset

The quantitative analysis of all methods on the LEVIR-CD dataset is presented in Table 2. It is observed that Transformer-based and Mamba-based models, which capture long-range dependencies, consistently outperform CNN-based models in overall performance. SO-Mamba achieved the best performance in three key evaluation metrics (F1, IoU, and OA), with an F1 score of 91.51% and an IoU of 84.35%. Compared to the initial Mamba-based network ChangeMamba, SO-Mamba enhances the F1 score by 0.98% and the IoU by 1.66%. CDLamba achieves the precision of 92.80% with a recall of 89.45%, while RS-Mamba reaches a recall of 92.57% and a precision of 89.84%. In comparison, our proposed SO-Mamba offers a precision of 92.34% and a recall of 90.70%. Among the Mamba-based methods, CDLamba also delivers competitive results with an F1 of 91.09% and the highest precision across all compared methods, yet SO-Mamba surpasses it by 0.42% in F1 and 0.70% in IoU, indicating that the sub-task-oriented design provides a more balanced trade-off between precision and recall.

Table 2
Quantitative comparison on LEVIR-CD dataset. All scores are presented as percentages (%). The top three results are highlighted in red, blue, and green.

Architecture	Method	Pre	Rec	F1	IoU	OA
CNN-based	IFNet	86.28	90.03	88.12	78.76	98.76
	SNUNet	90.87	88.83	89.84	81.55	98.98
	SEIFNet	91.10	88.89	89.99	81.80	98.99
Transformer-based	ChangeFormer	89.18	91.25	90.21	82.16	98.99
	SwinSUNet	91.64	86.62	89.06	80.28	98.91
	BIT	91.09	87.61	89.32	80.69	98.93
	DMINet	92.18	88.98	90.95	82.73	99.05
Mamba-based	ChangeMamba	90.24	90.81	90.53	82.69	99.03
	RS-Mamba	89.84	92.57	91.18	83.79	99.09
	CDMamba	89.08	92.62	90.81	83.17	99.05
	CDLamba	92.80	89.45	91.09	83.65	99.10
	SO-Mamba	92.34	90.70	91.51	84.35	99.14

In cases (a) and (b) of Fig. 6, the similarity in color between buildings and the ground causes some methods to exhibit widespread missed detections and edge-related misdetections of buildings. By contrast, SO-Mamba identifies the edge details of buildings more accurately, with detection results much closer to the GT. In cases (c) and (d), seasonal variations lead to color differences between the images A and B, while the small proportion and sparse distribution of buildings cause all methods to have certain degrees of missed detections. Conversely, SO-Mamba filters out color difference interference and effectively detects sparse change targets, which prominently demonstrates its exceptional capabilities in processing edge details and small targets.

4.3.2. WHU-CD dataset

As shown in Table 3, SO-Mamba achieves optimal performance in three evaluation metrics and sub-optimal performance in the remaining two metrics. Compared to the leading CNN-based method SNUNet, SO-Mamba achieves F1 and IoU improvements of 3.11% and 5.28%, respectively. Compared to the leading Transformer-based method SwinSUNet, SO-Mamba achieves a 1.23% improvement in IoU performance. CDMamba produces the highest precision at 94.40%, accompanied by a recall of 90.84%, while SwinSUNet leads in recall at 92.79% with a precision of 91.49%. Notably, our proposed SO-Mamba ranks second in both precision and recall, with impressive values of 93.96% for precision and 91.76% for recall. Also, the SO-Mamba yields the highest F1 score of 92.84%, alongside these best IoU of 86.65%; the results demonstrate that our method achieves an impressive balance of metrics.

In case (a) of Fig. 7, other methods exhibit substantial missed detections of building removal change targets. By contrast, SO-Mamba demonstrates distinct advantages, highlighting the capability to deeply interact with dual-temporal features and identify ground objects. In example (b), regular non-change targets such as vehicles appear around buildings. Compared with other methods, SO-Mamba accurately recognizes non-change objects, reducing false detection phenomena. In examples (c) and (d), shadow interference caused by lighting conditions

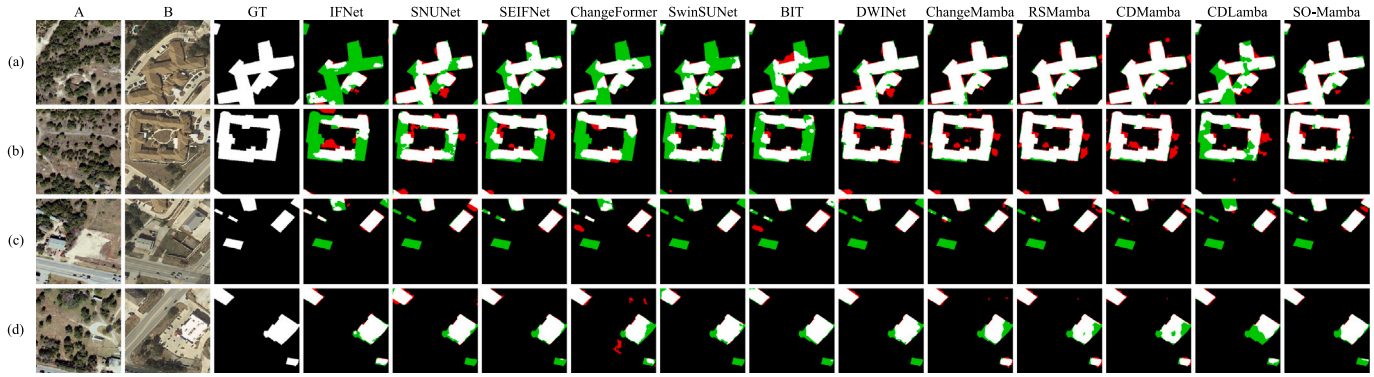


Fig. 6. Visualization results for different methods on the LEVIR-CD test set. (White represents a true positive, black is a true negative, red indicates a false positive, and green stands as a false negative).

Table 3
Quantitative comparison on WHU-CD dataset. All scores are presented as percentages (%). The top three results are highlighted in red, blue, and green.

Architecture	Method	Pre	Rec	F1	IoU	OA
CNN-based	IFNet	90.71	77.86	83.80	72.11	98.80
	SNUNet	90.58	88.88	89.73	81.37	99.19
	SEIFNet	86.68	90.97	88.78	79.82	99.09
Transformer-based	ChangeFormer	93.75	88.80	90.79	83.13	99.29
	SwinSUNet	91.49	92.79	92.14	85.42	99.37
	BIT	87.58	89.80	88.68	79.66	99.09
	DMINet	93.81	86.30	89.90	81.65	99.23
Mamba-based	ChangeMamba	92.65	91.45	92.05	85.26	99.37
	RS-Mamba	92.79	90.59	91.68	84.64	99.35
	CDMamba	94.40	90.84	92.59	86.20	99.42
	CDLamba	93.16	91.08	92.11	85.37	99.38
	SO-Mamba	93.96	91.76	92.84	86.65	99.44

on building roofs leads to confusion between buildings and the ground. Conversely, SO-Mamba shows remarkable advantages in identifying building structures and edges, unequivocally demonstrating the detail-processing capabilities under complex dual-temporal feature conditions.

4.3.3. GZ-CD dataset

As shown in Table 4, SO-Mamba achieved optimal performance in three evaluation metrics, with the F1 score reaching 87.26% and IoU reaching 77.39%. Compared with the sub-optimal method SEIFNet, SO-Mamba improved the F1 and IoU metrics by 0.25% and 0.39%, respectively. Among the Mamba-based methods, CDLamba obtains a notably high recall of 81.46% and Precision of 92.15%, yet its F1 of 86.48% and IoU of 76.18% fall below those of SO-Mamba by 0.78% and 1.21%, indicating that the sub-task-oriented decomposition strategy enables more precise boundary-level discrimination. Although SO-Mamba achieves the highest F1 and IoU, the recall metric (83.75%) is lower than that of SEIFNet (85.68%), indicating that convolutional networks retain advantages in capturing fine-grained local features within lower-resolution datasets. Notably, the WHU-CD dataset achieves a finer spatial resolution of 0.2 m compared to 0.55 m for GZ-CD. Meanwhile, the WHU-CD comprises regularly structured buildings with nearly no pseudo-changes, while GZ-CD depicts suburban landscapes filled with scattered small buildings and extensive pseudo-changes. In our method, the 2D row-wise and column-wise sequential scanning may destroy native local spatial correlations, leading to unsatisfactory edge detection for densely clustered buildings in GZ-CD compared with WHU-CD. Hence, in terms of recall, SEIFNet achieves 85.68% and SwinSUNet reaches 88.31% on GZ-CD, both of which surpass our method. Although both methods have high recall values, the precision is lower than that of our model. Typically, our approach delivers the best F1 score, which is the most valuable metric for CD task. All the results demonstrate that the proposed SO-Mamba achieves an optimal balance between false positives and missed detections.

In case (a) of Fig. 8, the buildings exhibit special colors similar to the surface, causing all methods to have substantial missed detections. By contrast, SO-Mamba has fewer missed detection areas than other methods, highlighting its superior image parsing ability. In cases (b)–(d), the overall low image clarity and shadow issues at building edges lead many methods to exhibit missed detections and misdetections at building boundaries. Compared with alternative methods, SO-Mamba precisely

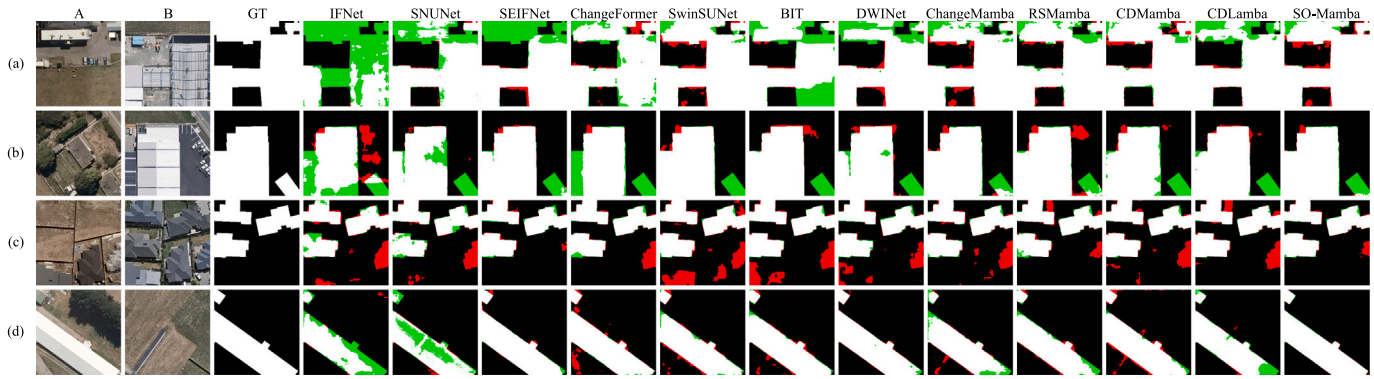


Fig. 7. Visualization results of different methods on the WHU-CD test set. (White represents a true positive, black is a true negative, red indicates a false positive, and green stands for a false negative).

Table 4
Quantitative comparison on GZ-CD dataset. All scores are presented as percentages (%). The top three results are highlighted in red, blue, and green.

Architecture	Method	Pre	Rec	F1	IoU	OA
CNN-based	IFNet	85.08	80.67	82.82	70.67	96.64
	SNUNet	85.12	84.75	84.93	73.81	96.98
	SEIFNet	88.39	85.68	87.01	77.00	97.43
Transformer-based	ChangeFormer	85.51	85.41	85.46	74.18	97.08
	SwinSUNet	84.41	88.31	86.31	75.93	97.18
	BIT	84.63	79.40	84.63	73.35	97.10
	DMINet	89.25	84.69	86.91	76.85	97.44
Mamba-based	ChangeMamba	92.59	79.00	85.26	74.31	97.26
	RSMamba	87.46	81.65	84.46	73.10	96.98
	CDMamba	87.61	81.94	84.68	73.43	97.02
	CDLamba	92.15	81.46	86.48	76.18	97.44
	SO-Mamba	91.06	83.75	87.26	77.39	97.54

identifies building edge details and reduces errors, demonstrating robust processing capabilities for image data and edge information under complex interference conditions.

4.4. Ablation experiments

To validate the feasibility of the task-driven design paradigm and the effectiveness of sub-task modules, we performed ablation experiments on the three sub-task modules. Each ablation variant was comparatively analyzed on three datasets.

4.4.1. The effects of different MIM designs

To validate the structural effectiveness of MIM from the perspective of dual-temporal feature interaction patterns, we removed the dual-temporal interaction structure in MIM and decomposed it into two same modules. We further proposed variant 1 with weight-sharing modules and variant 2 with independent-weight modules, respectively. In addition, to validate the contributions of auxiliary components and the MB, we retained the MB block while removing all augmentation parts within the MIM and defined this implementation as Variant 3 in this subsection, which also corresponds to the vanilla MIM excluding the asymmetric structure. The quantitative comparison of the ablation experiments is shown in Table 5. Variant 1 (employing the Siamese design concept) shows lower F1 and IoU metrics than MIM, illustrating the effectiveness of bi-temporal interactive modeling of MIM. Notably, the variant 2 demonstrates significant metric declines across all three datasets, confirming that weakening bi-temporal interaction directly impairs the detection performance of model. Variant 3 yields F1 scores of 88.70%, 90.36% and 84.35% on the LEVIR-CD, WHU-CD and GZ-CD datasets respectively, while the approach with the proposed MIM obtains 91.51%, 92.84% and 87.26% respectively. Such significant performance degradation verifies that asymmetric structure is indispensable for cross-temporal feature interaction.

In the results (1) and (2) of Fig. 9, the red boxes mark selected contrast regions, demonstrating that both variants recognize building details less effectively than MIM does. In particular, within result (1) of the Siamese variant, the loss of image details exhibits the greatest severity. Consequently, relying exclusively on the Siamese mechanism to achieve dual-temporal interaction proves insufficient to support all network architectures. The comparison verifies the effectiveness of MIM in enhancing the detection performance of SO-Mamba.

4.4.2. The effects of different MEDM designs

To explore the contribution of MEDM to SO-Mamba, we decomposed MEDM into three core components: First-order operator, Second-order operator, and FADConv, and combined them into four variants. To verify that the observed gain is not merely a consequence of additional depth, we introduce variant 5, which strips all edge-specific operators while preserving the feature dimensionality. The quantitative comparison is shown in Table 6. The variants 1–3 (without edge constraint operators) exhibit significant performance degradation, particularly variant 3 (without both second-order edge constraints), which shows severe performance decline. The results confirm the critical role of edge constraints in detection performance. After removing FADConv in variant 4, the metrics on all three datasets decrease, demonstrating contribution of FADConv to enhancing ability of Mamba to parse difference features containing edge information. Compared with the proposed MEDM, the F1 obtained by variant 5 drops by 1.35%, 1.39% and 2.07% on the LEVIR-CD, WHU-CD and GZ-CD, respectively. For IoU, the values decline 2.26%, 2.39% and 3.20% respectively. The results demonstrate that the auxiliary parts including FEOB, SEOB, and FADConv boost the accuracy of our model especially on the GZ-CD dataset. The results validate the effectiveness of introducing edge-based feature constraints in the MEDM module to improve difference capturing capabilities.

In Fig. 10, we show the detection results of the four variants and MEDM. In results (1)–(3), the contrast regions marked by red boxes clearly indicate that variants without edge constraints exhibit inferior capabilities in capturing building edge information compared to MEDM, showing issues such as incomplete boundaries and poor linearity. These issues validate the effectiveness of edge operators in enhancing performance of SO-Mamba in capturing edge-focused difference details. In result (4), the model obviously misdetects interference information, demonstrating the significant role of the FADConv module in processing difference features.

4.4.3. The effects of different MARM designs

To investigate the role of MARM in SO-Mamba, we decomposed the Mamba-centric attention enhancement mechanisms SA and CA, yielding

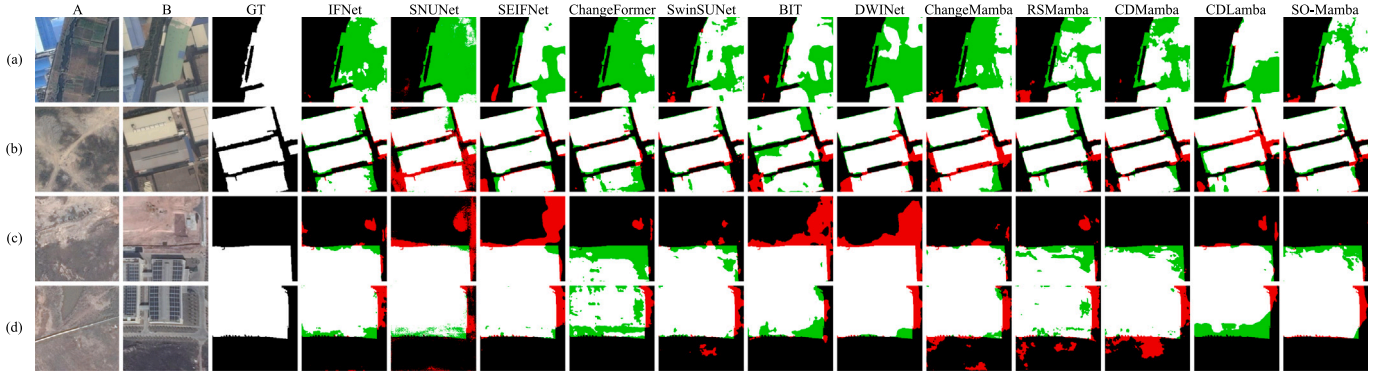


Fig. 8. Visualization results of different methods on the GZ-CD test set. (White represents a true positive, black is a true negative, red indicates a false positive, and green stands for a false negative).

Table 5
Ablation study of the MIM on three datasets in terms of F1 and IoU.

MIM variant	Shallow-layers weight-sharing	Deep-layers weight-sharing	LEVIR-CD		WHU-CD		GZ-CD	
			F1	IoU	F1	IoU	F1	IoU
1	✓	✓	91.25	83.91	92.13	85.41	86.04	75.50
2			90.88	83.28	92.04	85.26	86.22	75.78
	Asymmetric structure							
3			88.70	79.71	90.36	82.42	84.35	72.94
4	✓		91.51	84.35	92.84	86.65	87.26	77.39

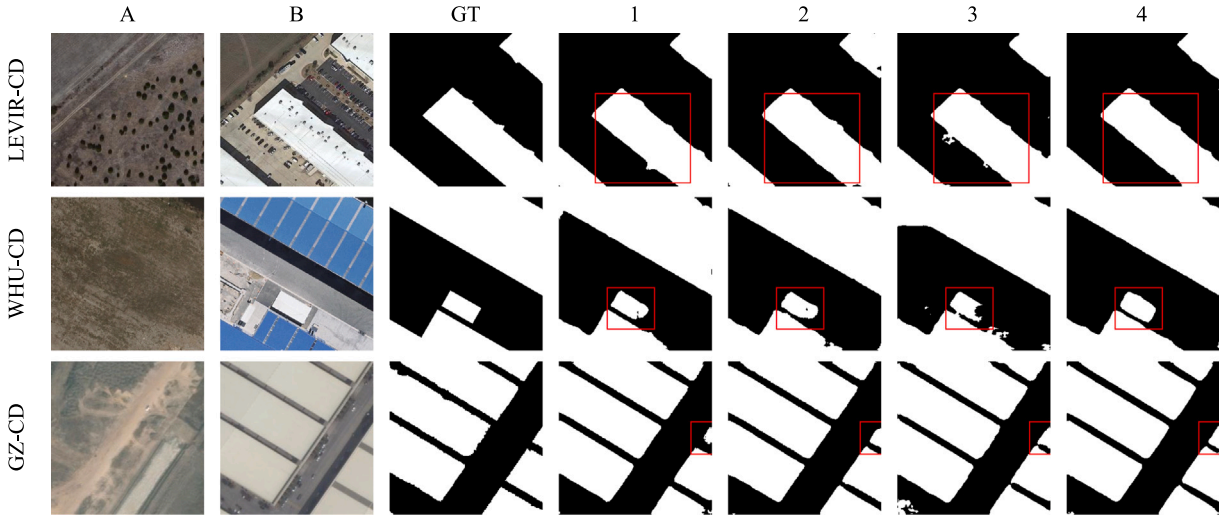


Fig. 9. Visualization results of the MIM ablation. (1)–(4) represent three different groups of MIM ablation experiments respectively.

Table 6
Ablation study of the MEDM on three datasets in terms of F1 and IoU.

MEDM variant	First-order	Second-order	FADConv	LEVIR-CD		WHU-CD		GZ-CD	
				F1	IoU	F1	IoU	F1	IoU
1		✓	✓	91.00	83.49	92.21	85.54	86.98	76.97
2	✓		✓	90.99	83.47	92.09	85.34	86.63	76.41
3			✓	90.29	82.30	91.39	84.16	85.81	75.15
4	✓	✓		91.39	83.15	92.21	86.55	86.63	76.42
5				90.16	82.09	91.45	84.26	85.19	74.19
6	✓	✓	✓	91.51	84.35	92.84	86.65	87.26	77.39

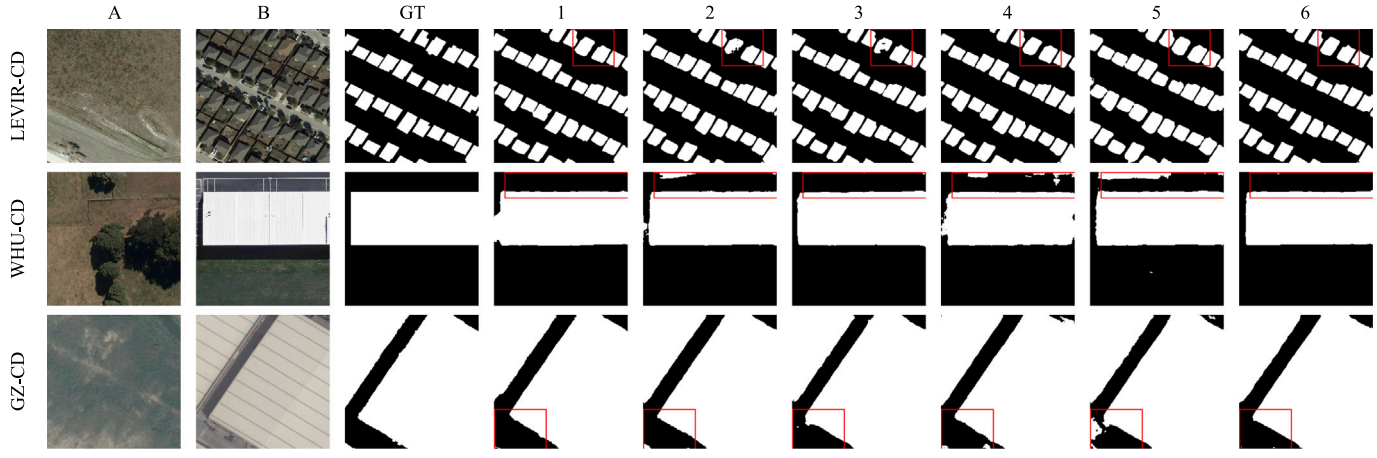


Fig. 10. Visualization results of the MEDM ablation. (1)–(6) represent five different groups of MEDM ablation experiments respectively.

Table 7
Ablation study of the MARM on three datasets in terms of F1 and IoU.

MARM variant	CA	SA	LEVIR-CD		WHU-CD		GZ-CD	
			F1	IoU	F1	IoU	F1	IoU
1	✓		90.99	83.49	92.33	85.75	86.58	76.34
2		✓	91.21	83.84	92.40	85.88	87.01	77.01
3			90.67	82.94	90.30	82.31	86.28	75.87
4	✓	✓	91.51	84.35	92.84	86.65	87.26	77.39

four variants. The quantitative comparison is shown in Table 7. Variant 1 (without spatial attention) exhibits significant performance degradation, while variant 2 (without channel attention) also shows metric declines across all three datasets. With channel and spatial attention removed, variant 3 achieves F1 scores of 90.40%, 90.03% and 84.71% across LEVIR-CD, WHU-CD and GZ-CD, which are consistently lower than the 91.51%, 92.84% and 87.26% yielded by the original MARM. The performance drops reveal the positive contribution of attention-guided feature redistribution for change-region identification, thereby validating the critical role of MARM in enhancing the target reconstruction performance of SO-Mamba in CD tasks.

In Fig. 11, we display the detection results of the three variants and MARM. In the results (1) and (3), the contrast regions marked by red boxes show that buildings exhibit certain deformation, incomplete detection, and blurred edges in the variants. The results confirm that the attention-based enhancement mechanism effectively improves building reconstruction capabilities of SO-Mamba. Case (2) represents MARM without channel attention. Detection performance degradation is less severe in the variant than in MARM without spatial attention. The comparison indicates that increasing spatial attention proportion may represent an effective optimization strategy for SO-Mamba.

4.4.4. The effects of different loss function designs

To investigate the impact of the Lovasz-softmax loss introduced in this work on detection performance, we assigned different values to the weight λ of the Lovasz-softmax loss and tested the corresponding performance of SO-Mamba on the three datasets. As shown in Fig. 12, we set λ to 0, 0.25, 0.5, 0.75, and 1, and recorded the corresponding F1 metrics. It is evident from the results that as λ increases, the metrics significantly improve and reach their peak when λ is 0.75. Notably, compared to the worst performance, the F1 metrics with λ set to 0.75 achieve improvements of 0.7%, 0.73%, and 0.78% on the LEVIR-CD, WHU-CD, and GZ-CD datasets, respectively.

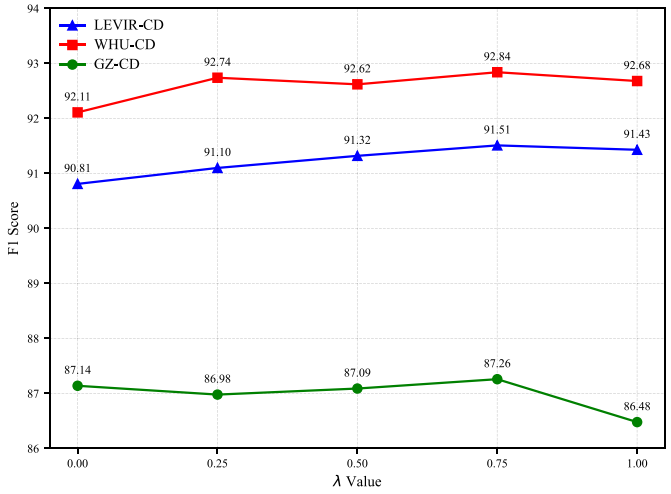


Fig. 12. Visualization scores of F1 for the loss function ablation. The abscissa represents different values of the parameter λ in the loss function L_{total}, and the ordinate represents the corresponding quantitative performance of SO-Mamba.

To further investigate the focus of Lovasz-softmax loss on the IOU metric, we varied the weight λ and evaluated the corresponding IOU scores of SO-Mamba across the three datasets. As illustrated in Fig. 13, the IOU metrics show a significant improvement as the weight of Lovasz-softmax loss increases, peaking when λ is set to 0.75. Notably, compared to the scenario without Lovasz-softmax loss (λ set to 0), the IOU metrics achieve improvements of 1.18% on LEVIR-CD, 1.28% on WHU-CD, and 1.21% on GZ-CD, respectively. The results demonstrate that introducing the Lovasz-softmax loss can significantly enhance the detection performance of SO-Mamba.

4.5. Efficiency analysis

The rigorous comparative analysis of computational efficiency of SO-Mamba is shown in Table 8. We select Params and GFLOPS metrics to compare the model efficiency among methods. The multi-module architecture of SO-Mamba enhances network depth while maintaining parameters at 31.3 M. Notably, SO-Mamba maintains mid-tier efficiency. Compared with IFNet, ChangeFormer, and SwinSUNet, SO-Mamba significantly reduces model complexity while demonstrating

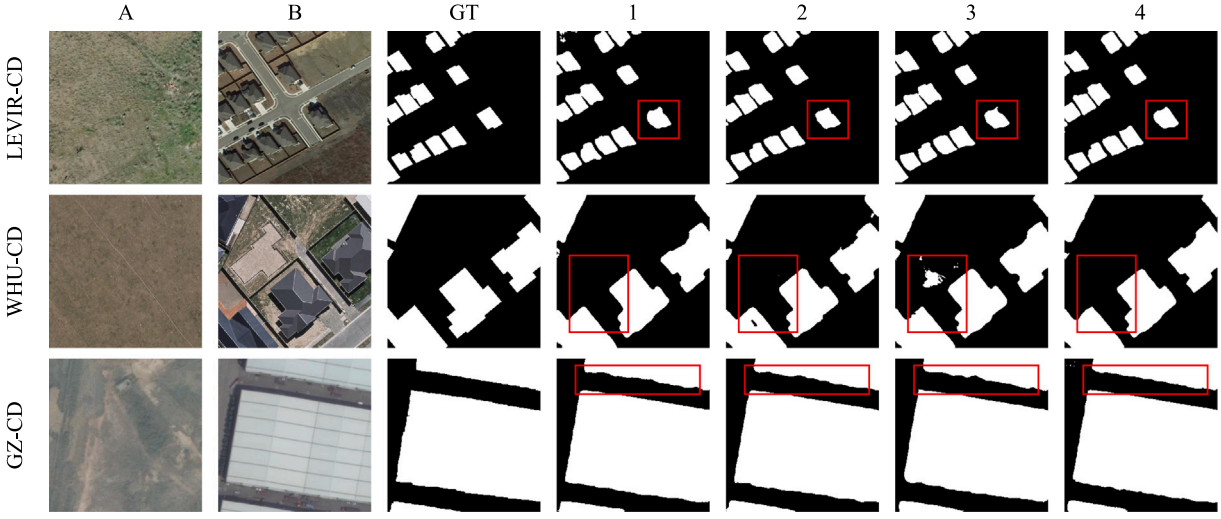


Fig. 11. Visualization results of the MARM ablation. (1)–(4) represent four different groups of MARM ablation experiments respectively.

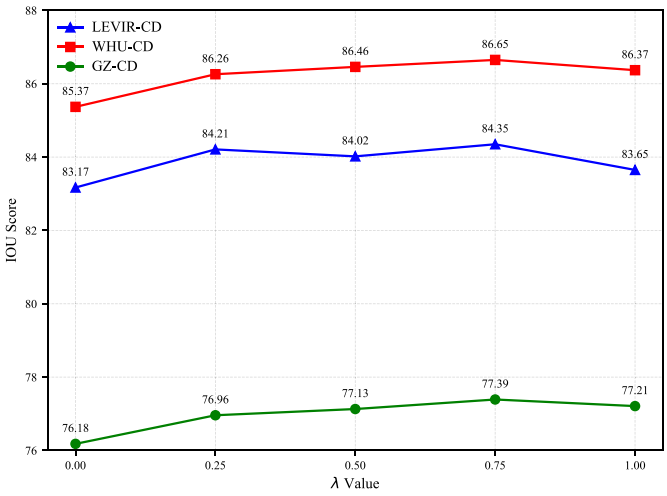


Fig. 13. Visualization scores of IOU for the loss function ablation.

Table 8
Quantitative comparison of method efficiency.

Type	Method	Params(M)	GFLOPs
CNN	IFNet	50.71	82.35
	SNUNet	12.03	54.88
	SEIFNet	27.91	8.37
Transformer	ChangeFormer	41.03	138.80
	SwinSUNet	39.28	43.50
	BIT	3.55	10.59
	DMINet	6.24	14.55
Mamba	ChangeMamba	17.13	45.74
	RSMamba	51.19	22.82
	CDMamba	11.90	49.26
	CDLamba	28.74	15.26
	SO-Mamba	31.30	17.67

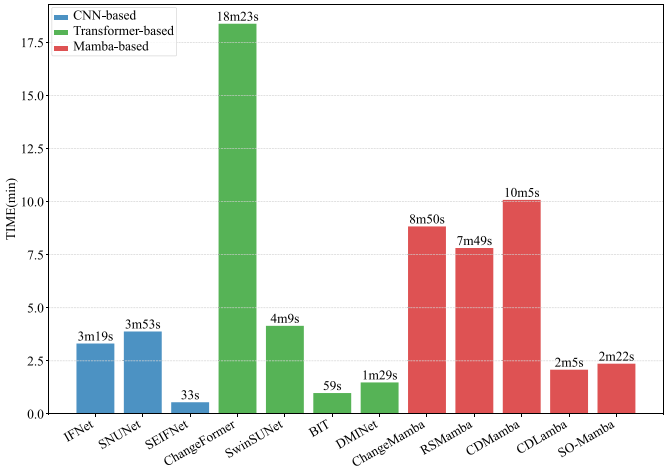


Fig. 14. Visualization results of the training efficiency of the model. The ordinate represents the time required to train one epoch on the LEVIR-CD dataset, with the unit being minutes. The abscissa represents twelve comparative methods.

superior detection performance. When compared to models with smaller Params such as BIT and DMINet, SO-Mamba exhibits comparable GFLOPS. The benefit stems from the remarkable advantage of Mamba in enhancing computational efficiency while maintaining low floating-point operations, thereby validating the applicability of the Mamba architecture in CD tasks requiring high computational performance.

Fig. 14 displays the training time of all methods, with one epoch of training on the LEVIR-CD dataset serving as the comparison benchmark. It is evident from the figure that training time of SO-Mamba is significantly shorter than that of ChangeFormer and ChangeMamba models. The efficiency gain stems from the proposed three-layer network architecture. The architecture reduces image downsampling frequency and feature dimension expansion factors, significantly improving computational efficiency while maintaining high detection accuracy across all three datasets. This result further validates the advantage of SO-Mamba in CD tasks.

5. Conclusion

In this work, we present the novel task-oriented architecture SO-Mamba, which prioritizes the task-specific challenges of CD and decomposes the CD task into three sub-tasks. Firstly, we propose MIM to solve the dual-temporal feature interaction sub-task. The MIM architects temporally corresponding residual blocks to assist the selective interaction of Mamba, achieving profound feature interaction within dual-temporal images. Secondly, we propose MEDM to solve the difference information extraction sub-task. The MEDM constructs edge operators within the frequency-enhanced Mamba architecture to enhance image parsing capability by extracting edge-focused differential information and preserving inherent spatial details. Thirdly, we propose MARM to solve the target detail reconstruction sub-task. The MARM architects dual-attention mechanisms to dynamically adjust reconstruction features within the sequential cross-scanning representation of Mamba, thereby refining target feature details. Experiments on three public datasets show that SO-Mamba outperforms other mainstream architectures in the CD task. Furthermore, ablation experiments conducted across three datasets validate the efficacy of each proposed module. In future work, we will further develop task-optimized Mamba-centric architectures to harness the capabilities of the Mamba framework, advancing change detection performance and computational efficiency.

CRediT authorship contribution statement

Chunyan Yu: Writing – review & editing, Writing – original draft, Methodology, Conceptualization. **Zhenhao Zhang:** Writing – review & editing, Writing – original draft, Software, Methodology. **Qiang Zhang:** Supervision, Conceptualization. **Yulei Wang:** Visualization, Validation. **Haoyang Yu:** Supervision, Resources.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

References

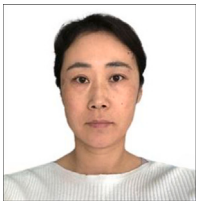
[1] T.J. Glezakos, T.A. Tsiligiridis, L.S. Iliadis, C.P. Yialouris, F.P. Maris, K.P. Ferentinos, Feature extraction for time-series data: an artificial neural network evolutionary training model for the management of mountainous watersheds, Neurocomputing 73 (1) (2009) 49–59. Timely Developments in Applied Neural Computing (EANN 2007)/Some Novel Analysis and Learning Methods for Neural Networks (ISNN 2008)/Pattern Recognition in Graphical Domains.

[2] L. Oneto, F.-M. Schleif, A. Sperduti, Advances in artificial neural networks, machine learning and computational intelligence, Neurocomputing 618 (2025) 129130.

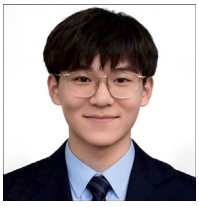
[3] Y. Jiang, P. Zhao, C. Zhao, J. Lin, Towards bandwidth efficient edge–cloud collaborative deep learning with data importance driven compression, Neurocomputing 650 (2025) 130835.

[4] W. Robson Schwartz, V. Hugo Cunha de Melo, H. Pedrini, L.S. Davis, A data-driven detection optimization framework, Neurocomputing 104 (2013) 35–49.
[5] X. Yang, A.S.A. Mohamed, Gaussian-based R-CNN with large selective kernel for rotated object detection in remote sensing images, Neurocomputing 620 (2025) 129248.
[6] D. Yi, J. Su, W.-H. Chen, Probabilistic faster R-CNN with stochastic region proposing: towards object detection and recognition in remote sensing imagery, Neurocomputing 459 (2021) 290–301.
[7] Y. Sun, H. Zhao, J. Zhou, DKETFormer: salient object detection in optical remote sensing images based on discriminative knowledge extraction and transfer, Neurocomputing 625 (2025) 129558.
[8] C. Zhou, Z. He, L. Zou, Y. Li, A. Plaza, HSACT: a hierarchical semantic-aware CNN-transformer for remote sensing image spectral super-resolution, Neurocomputing 636 (2025) 129990.
[9] C. Ma, L. Weng, M. Xia, H. Lin, M. Qian, Y. Zhang, Dual-branch network for change detection of remote sensing image, Eng. Appl. Artif. Intell. 123 (2023) 106324.
[10] Q. Wang, Z. Yuan, Q. Du, X. Li, GETNET: a general end-to-end 2-D CNN framework for hyperspectral image change detection, IEEE Trans. Geosci. Remote Sens. 57 (1) (2018) 3–13.
[11] R.C. Daudt, B. Le Saux, A. Boulch, Fully convolutional Siamese networks for change detection, in: 2018 25th IEEE International Conference on Image Processing (ICIP), IEEE, 2018, pp. 4063–4067.
[12] J. Long, E. Shelhamer, T. Darrell, Fully convolutional networks for semantic segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3431–3440.
[13] C. Zhang, P. Yue, D. Tapete, L. Jiang, B. Shangguan, L. Huang, G. Liu, A deeply supervised image fusion network for change detection in high resolution bi-temporal remote sensing images, ISPRS J. Photogramm. Remote Sens. 166 (2020) 183–200.
[14] Z. Dong, J. Li, Z. Hua, Transformer-based multi-attention hybrid networks for skin lesion segmentation, Expert Syst. Appl. 244 (2024) 123016.
[15] Z. Dong, G. Yuan, Z. Hua, J. Li, Diffusion model-based text-guided enhancement network for medical image segmentation, Expert Syst. Appl. 249 (2024) 123549.
[16] P. Li, T. Si, C. Ye, Q. Guo, Semantic-aware transformer with feature integration for remote sensing change detection, Eng. Appl. Artif. Intell. 135 (2024) 108774.
[17] W.G.C. Bandara, V.M. Patel, A transformer-based Siamese network for change detection, in: IGARSS 2022-2022 IEEE International Geoscience and Remote Sensing Symposium, IEEE, 2022, pp. 207–210.
[18] X. Dong, J. Bao, D. Chen, W. Zhang, N. Yu, L. Yuan, D. Chen, B. Guo, CSWin transformer: a general vision transformer backbone with cross-shaped windows, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 12124–12134.
[19] Y. Feng, H. Xu, J. Jiang, H. Liu, J. Zheng, ICIF-Net: intra-scale cross-interaction and inter-scale feature fusion network for bitemporal remote sensing images change detection, IEEE Trans. Geosci. Remote Sens. 60 (2022) 1–13.
[20] H. Chen, Z. Qi, Z. Shi, Remote sensing image change detection with transformers, IEEE Trans. Geosci. Remote Sens. 60 (2021) 1–14.
[21] A. Gu, K. Goel, A. Gupta, C. Re, On the parameterization and initialization of diagonal state space models, Adv. Neural Inf. Process. Syst. 35 (2022) 35971–35983.
[22] H. Chen, J. Song, C. Han, J. Xia, N. Yokoya, ChangeMamba: remote sensing change detection with spatiotemporal state space model, IEEE Trans. Geosci. Remote Sens. 62 (2024) 1–20.
[23] Z. Dong, G. Yuan, Z. Hua, J. Li, ConMamba: CNN and SSM high-performance hybrid network for remote sensing change detection, IEEE Trans. Geosci. Remote Sens. 62 (2024).
[24] Y. Guo, Y. Xu, G. Tang, Z. Yu, Q. Zhao, Q. Tang, AM-CD: joint attention and Mamba for remote sensing image change detection, Neurocomputing 647 (2025) 130607.
[25] S. Zhao, H. Chen, X. Zhang, P. Xiao, L. Bai, W. Ouyang, RS-Mamba for large remote sensing image dense prediction, IEEE Trans. Geosci. Remote Sens. 62 (2024).
[26] H. Zhang, K. Chen, C. Liu, H. Chen, Z. Zou, Z. Shi, CDMamba: incorporating local clues into Mamba for remote sensing image binary change detection, IEEE Trans. Geosci. Remote Sens. 63 (2025).
[27] T. Xue, J. Zhao, J. Li, C. Chen, K. Zhan, CD-Mamba: cloud detection with long-range spatial dependency modeling, J. Appl. Remote Sens. 19 (3) (2025) 038507.
[28] R. Wu, Y. Liu, G. Ning, P. Liang, Q. Chang, Ultralight VM-UNet: parallel vision Mamba significantly reduces parameters for skin lesion segmentation, Patterns 6 (11) (2025).
[29] Y. Huang, X. Li, Z. Du, H. Shen, Spatiotemporal enhancement and interlevel fusion network for remote sensing images change detection, IEEE Trans. Geosci. Remote Sens. 62 (2024) 1–14.
[30] S. Fang, K. Li, J. Shao, Z. Li, SNUNet-CD: a densely connected Siamese network for change detection of VHR images, IEEE Geosci. Remote Sens. Lett. 19 (2021) 1–5.
[31] Z. Zhou, M.M. Rahman Siddiquee, N. Tajbakhsh, J. Liang, UNet++: a nested U-Net architecture for medical image segmentation, in: International Workshop on Deep Learning in Medical Image Analysis, Springer, 2018, pp. 3–11.
[32] C. Zhang, L. Wang, S. Cheng, Y. Li, SwinSUNet: pure transformer network for remote sensing image change detection, IEEE Trans. Geosci. Remote Sens. 60 (2022) 1–13.
[33] Y. Feng, J. Jiang, H. Xu, J. Zheng, Change detection on remote sensing images using dual-branch multilevel intertemporal network, IEEE Trans. Geosci. Remote Sens. 61 (2023) 1–15.
[34] F. Yang, M. Li, W. Shu, A. Qin, T. Song, C. Gao, G.-S. Xia, ConvFormer-CD: hybrid CNN-transformer with temporal attention for detecting changes in remote sensing imagery, IEEE Trans. Geosci. Remote Sens. 63 (2025).
[35] Q. Li, R. Zhong, X. Du, Y. Du, TransUNetCD: a hybrid transformer network for change detection in optical remote-sensing images, IEEE Trans. Geosci. Remote Sens. 60 (2022) 1–19.
[36] L. Song, M. Xia, L. Weng, H. Lin, M. Qian, B. Chen, Axial cross attention meets CNN: bibranch fusion network for change detection, IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 16 (2022) 21–32.
[37] Y. Liu, Y. Tian, Y. Zhao, H. Yu, L. Xie, Y. Wang, Q. Ye, J. Jiao, Y. Liu, VMamba: visual state space model, Adv. Neural Inf. Process. Syst. 37 (2024) 103031–103063.
[38] A. Gu, T. Dao, Mamba: linear-time sequence modeling with selective state spaces, arXiv preprint arXiv:2312.00752, 2023.
[39] L. Chen, L. Gu, D. Zheng, Y. Fu, Frequency-adaptive dilated convolution for semantic segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 3414–3425.
[40] S. Zhang, Z. Wei, W. Xu, L. Zhang, Y. Wang, J. Zhang, J. Liu, Edge aware depth inference for large-scale aerial building multi-view stereo, ISPRS J. Photogramm. Remote Sens. 207 (2024) 27–42.
[41] D. Marr, E. Hildreth, Theory of edge detection, Proc. R. Soc. Lond, B. Biol. Sci. 207 (1167) (1980) 187–217.
[42] M. Berman, A.R. Triki, M.B. Blaschko, The Lovász-softmax loss: a tractable surrogate for the optimization of the intersection-over-union measure in neural networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 4413–4421.
[43] H. Chen, Z. Shi, A spatial-temporal attention-based method and a new dataset for remote sensing image change detection, Remote Sens. 12 (10) (2020) 1662.
[44] S. Ji, S. Wei, M. Lu, Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set, IEEE Trans. Geosci. Remote Sens. 57 (1) (2018) 574–586.
[45] D. Peng, L. Bruzzone, Y. Zhang, H. Guan, H. Ding, X. Huang, SemiCDNet: a semisupervised convolutional neural network for change detection in high resolution remote-sensing images, IEEE Trans. Geosci. Remote Sens. 59 (7) (2020) 5891–5906.
[46] Z. Wu, X. Ma, K. Zheng, R. Lian, Y. Chen, Z. Huang, W. Zhang, S. Song, CD-lamba: boosting remote sensing change detection via a cross-temporal locally adaptive state space model, IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 19 (2026) 4028–4044.

Author biography



Chunyan Yu received her Ph.D. degree in environmental engineering from Dalian Maritime University, Dalian, China, in 2012. She is currently a Professor at the Information Science and Technology College, Dalian Maritime University. Her research interests include image segmentation, hyperspectral image classification, and pattern recognition.



Zhenhao Zhang received his bachelor's degree in Software Engineering from Zhengzhou University of Light Industry in 2024. Currently, he is pursuing his master's degree in Computer Science and Technology at Dalian Maritime University, in China. His research interests include remote sensing image processing and deep learning.

Qiang Zhang (Member, IEEE) received the B.E. degree in surveying and mapping engineering and the M.E. and Ph.D. degrees in photogrammetry and remote sensing from Wuhan University, Wuhan, China, in 2017, 2019, and 2022, respectively. He is currently a Xinghai Associate Professor with the Center of Hyperspectral Imaging in Remote Sensing (CHIRS), Information Science and Technology College, Dalian Maritime University, Dalian, China. He has authored more than twenty journal articles in the IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, Earth System Science Data, and ISPRS Journal of Photogrammetry and Remote Sensing. His research interests include remote sensing information processing, computer vision, and machine learning. More details can be found at https://qzhang95.github.io.

Yulei Wang received B.S. and Ph.D. degrees in signal and information processing from Harbin Engineering University, Harbin, China, in 2009 and 2015, respectively. She was awarded by the China Scholarship Council in 2011 as a joint Ph.D. student to study at Remote Sensing Signal and Image Processing Laboratory, University of Maryland, Baltimore, MD, USA, for two years. She is an Associate Professor with the Hyperspectral Imaging in Remote Sensing (CHIRS), at the Information Science and Technology College, Dalian Maritime University, Dalian, China. Her research interests include hyperspectral image processing and vital signs signal processing.

Haoyang Yu (S'16-M'21) received B.S. degree in Information and Computing Science from Northeastern University, Shenyang, China, in 2013, and the Ph.D. degree in cartography and geographic information systems from the Key Laboratory of Digital Earth Science, Aerospace Information Research Institute, Chinese Academy of Sciences (CAS), Beijing, China, in 2019. He is currently a Xing Hai Associate Professor with the Center of Hyperspectral Imaging in Remote Sensing (CHIRS), Information Science and Technology College, Dalian Maritime University. His research focuses on models and algorithms for hyperspectral image processing, analysis, and applications.