Pattern Recognition 172 (2026) 112648

Contents lists available at ScienceDirect

Pattern Recognition

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

Enhancing VMamba for change detection via lightweight feature interaction and selection

Mingwei Zhang [a,b], Yuan Jiang [b], Qi Wang [b,*]

[a] The School of Computer Science, Northwestern Polytechnical University, Xi'an, 710072, China
[b] The School of Artificial Intelligence, Optics and Electronics, Northwestern Polytechnical University, Xi'an, 710072, China

ARTICLE INFO

Keywords:
State space model
Bi-temporal interaction
Feature selection
Binary change detection
Multi-class change detection.

ABSTRACT

Change detection aims to identify areas or objects of interest that have changed between bi-temporal images, which is a hot topic in the remote seeing and pattern recognition community. Recently, the visual state space model (VMamba) has demonstrated impressive and efficient results compared to previous methods based on convolutional neural networks and Transformers. However, existing VMamba-based models ignore the domain gap in bi-temporal images, which limits the potential of VMamba for change detection. To this end, a Siamese VMamba with Interaction (VMI-CD) model is developed. Specifically, a parameter-free bi-temporal feature interaction module (BFIM) is proposed, which is custom-designed for VMamba during the feature encoding stage to enhance the perception of the model in domain differences. Besides, a channel-spatial selection module (CS2M) is designed to modulate bi-temporal features, which aims to facilitate the generation of discriminative change representations in difference extraction. Thanks to the lightweight design, only a small number of parameters are added with the introduction of BFIM and CS2M. Experimental results on five remote sensing image change detection datasets with different tasks, MCLC-CD and JL1-CD, which contain multiple change types, LEVIR-CD+ and WHU-CD for building change detection, and SYSU-CD, a category-agnostic binary change detection dataset, demonstrate that VMI-CD surpasses previous state-of-the-art approaches. The code will be available at https://github.com/ptdoge/VMI-CD.

1. Introduction

Change detection aims to monitor the evolution of a given region using bi-temporal images acquired at different times. It is a popular topic in the artificial intelligence and remote sensing communities, with numerous practical applications including urban planning [1], disaster rescue [2], and ecology protection [3]. Therefore, designing an effective and automatic method for change detection is crucial. However, accurate change detection is challenging due to inherent imaging differences between the bi-temporal images, such as lighting changes, seasonal variations, and diverse viewpoints. Recently, researchers have increasingly focused on deep learning-based methods to alleviate the above challenges, achieving impressive results.

The success of deep learning-based change detection models can be largely attributed to the powerful representation capabilities of modern neural networks. In the early stages, many methods primarily relied on CNNs as feature extractors for bi-temporal images. CNNs are particularly effective at extracting local features, capturing low-level patterns through convolution operations, and progressively learning more abstract high-level representations via their multi-layer structure [4]. Daudt et al. [5] first propose a fully convolutional Siamese network for change detection, in which three ways of difference mining are explored, including early image concatenation, late feature concatenation, and subtraction. After that, many works follow it to capture bi-temporal image changes, while various feature enhancement techniques are introduced for improving accuracy. In recent years, Transformers have gained significant attention for change detection tasks. Transformers excel at capturing global context through their self-attention mechanism, which enables them to model long-range dependencies and relationships across the entire image. Bandara et al. [6] develop the first Siamese network based on the Transformer, which achieves impressive performance. Transformer architectures such as SWinTransformer and SegFormer have become common backbone networks for bi-temporal feature extraction [7].

In addition to pure CNN and Transformer models for change detection, some studies explore the hybrid architectures of CNN and Transformer. For example, Chen et al. [8] develop a bi-temporal transformer model, which employs CNN for bi-temporal image encoding, and

* Corresponding author.
E-mail addresses: dlaizmw@gmail.com (M. Zhang), jl4564562022@163.com (Y. Jiang), crabwq@gmail.com (Q. Wang).

https://doi.org/10.1016/j.patcog.2025.112648
Received 5 March 2025; Received in revised form 1 October 2025; Accepted 21 October 2025
Available online 27 October 2025

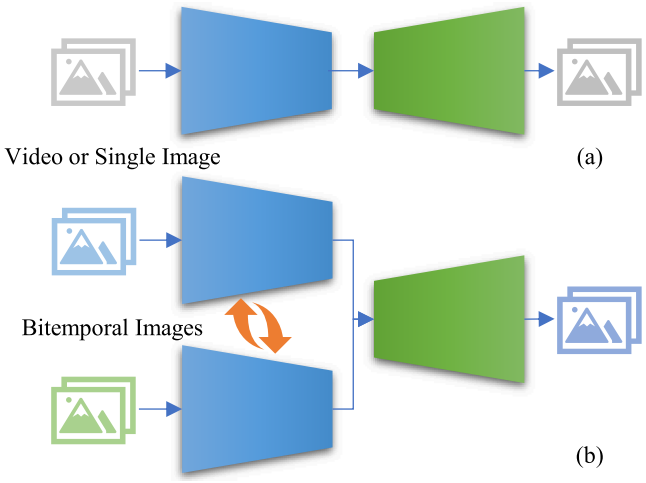

Fig. 1. (a) The schematic diagram of the framework for single image or video processing based on VMamba. (b) The framework diagram for processing bi-temporal images through Siamese VMamba Interaction (VMI-CD).

introduces a Transformer for bi-temporal feature interaction and difference representation. These methods achieve excellent performance in change detection. However, CNNs require deeper layers or larger kernels to fully capture long-range context information. Meanwhile, the Transformer requires more computational resources.

The VMamba, newly proposed by Liu et al. [9], effectively addresses the limitations of both CNNs and Transformers. In detail, it unfolds the two-dimensional image as a one-dimensional sequence and then utilizes the selective state models (SSM) to capture global context knowledge through a compressed hidden state. Therefore, its complexity is linear. Besides, a cross-scan strategy is specially designed for it. This strategy aims to ensure that each location in the feature embedding captures information from all other pixels in different directions, thereby achieving long-range modeling. Its effectiveness and superiority have been proven in video and image analysis. Fig. 1(a) shows the general framework based on VMamba for video or single image processing. For example, VideoMamba [10] introduces an innovative solution for video understanding by overcoming the limitations of 3D CNNs and video transformers, offering scalability, long-term modeling, and multi-modal compatibility. Xiao et al. [11] demonstrate the potential of state-space models for super-resolution. Meanwhile, the characteristics of VMamba make it well-suited for change detection, as it has been shown that the task can benefit from both long-range contextual modeling and efficient processing of multi-temporal images. For example, Chen et al. [12] propose a spatiotemporal state space model. Paranjape et al. [13] design a Mamba-based Siamese Network. However, they apply VMamba to independently extract representations from bi-temporal images, overlooking the impact of the discrete nature of the imaging process across the two-time points. We argue that the full potential of VMamba in change detection has yet to be fully exploited.

Unlike video or single-image analysis, bi-temporal images often exhibit style differences caused by variations in imaging conditions. These differences, referred to as pseudo-changes, include variations in land cover color or light intensity. Based on this observation, as shown in Fig. 1(b), we explore the bi-temporal feature interaction to harness VMamba for change detection more effectively. A model called VMI-CD is proposed. First, a bi-temporal feature interaction module (BFIM) tailored for Siamese VMamba is developed to enhance the ability of Siamese VMamba in bi-temporal domain difference perception. Specifically, BFIM contrastively scales the bi-temporal VMamba features by computing pixel-wise feature distributions. Compared to separate bi-temporal feature extraction, the inclusion of BFIM helps alleviate feature shifts caused by imaging differences, thereby boosting the immunity of the change detector to pseudo-changes. Second, to facilitate the generation of discriminative change representations, a channel-spatial selection module (CS2M) is developed. This module aims to enlarge differences between the bi-temporal features by selectively focusing on the most salient temporal instances in the channel and spatial dimensions, thereby improving sensitivity to changes. In summary, the contributions of this manuscript are as follows,

- We propose a novel change detector VMI-CD. It introduces a well-designed interaction strategy tailored for the Siamese Vmamba, effectively mitigating the impact of pseudo-changes induced by domain differences. In particular, no additional parameters are introduced in the interaction process.
- We develop a channel-spatial selection module that adaptively amplifies and suppresses bi-temporal features across both channel and spatial dimensions to enhance the differences in bi-temporal representations, thereby improving sensitivity to subtle changes. Meanwhile, the module follows a lightweight design concept.
- Our VMI-CD performs state-of-the-art on five datasets with distinct change detection tasks, including two multi-class change detection datasets, two building change detection datasets, and one class-agnostic dataset. Its superiority is widely proven.

2. Related work

2.1. State space model

State Space Models (SSMs) have shown adequate effectiveness of the state space transformation in modeling the interdependency of language sequences. Gu et al. [14] first propose a structured state-space sequence (S4) model with linear complexity to capture context information for long sequences. Subsequently, its various variants are designed including GSS [15], S5 [16], and so on. Gu et al. [17] further develop the Mamba based on their work S4 by introducing the input-dependent SSM layer and selection mechanism. Compared to Transformers, SSMs effectively capture long-range dependency with linear complexity. Due to the demonstrated superiority and effectiveness of State-Space Models (SSMs) in language modeling, SSMs are rapidly introduced into the visual domain and extensively employed.

The Visual State-Space (VSS) model, known as VMamba, is first proposed in [9]. It divides an image into patch sequences and processes them through four stages, with each stage consisting of several VSS blocks. The Vision Mamba [18] is introduced almost simultaneously with VMamba, and is developed based on bidirectional Mamba blocks. It is not as popular as VMamba. As a general backbone network, VMamba is widely employed across various downstream tasks. In natural image processing, Yang et al. [19] develop the ReMamber for referring image segmentation by leveraging the advancements of Mamba in

efficient training and inference. Li et al. [10] propose the VideoMamba for efficient video understanding. They investigate multiple different scan methods for video sequences and indicate that only stacking the spatial tokens frame by frame is effective enough. Additionally, various X-Mamba models have been developed for different tasks, such as MTMamba [20]. In the remote sensing community, Mamba has been employed widely. Liu et al. [21] design a CaMa layer made of spatial difference-guided SSM and temporal traveling SSM, thus constructing a RsCaMa for change caption. Xiao et al. [11] propose an FMSR for super-resolution, which introduces a frequency state space model. In this work, we further delve into the potential of VMamba for change detection.

2.2. Change detection

Change detection is a key area of focus in the remote sensing community. Currently, the mainstream methods for change detection are almost based on deep learning. Based on how bi-temporal features are extracted, these methods are classified into CNN-based, Transformer-based, Mamba-based, or hybrid approaches. For instance, Daudt et al. [5] are the first to propose a fully convolutional Siamese network for change detection. Bandara et al. [6] introduce the first Transformer-based Siamese network for change detection. Additionally, Chen et al. [12] pioneer the application of the state space model in the change detection domain.

Feature interaction is a powerful approach for enhancing performance in multi-input image analysis [22,23]. Thus, in addition to distinguishing feature extraction, existing studies explore various feature interaction strategies to better leverage bi-temporal features for improved change detection. Li et al. [24] introduce Changer, which investigates an interaction mechanism that facilitates the exchange of bi-temporal features across the channel and spatial dimensions. BAN [25] explores a change detection paradigm based on the foundational model, which introduces the interaction between the learnable side representation and the frozen pretrained feature. Meanwhile, the diverse variants of the multi-scale feature fusion and the attention mechanism have been extensively explored, enabling the model to more effectively focus on the changed regions [26,27]. Zheng et al. [28] propose a high-frequency attention-guided model for building change detection, which focuses on enhancing the edge details of buildings. TCRPN [29] introduces a temporal saliency attention to highlight changed areas. RFL-CDNet [30] boosts change detection performance by learning richer features through deep supervision and multi-scale prediction. CLAFA [31] employs a multiplicative channel attention and an additive gated attention to fuse cross-level features, thus enhancing change detection. Besides, some studies explore the domain gap alleviation strategy for change detection. Liu et al. [32] propose a supervised domain adaptation framework to address cross-domain change detection. Inspired by the above studies, this work develops lightweight interaction and selection methods to boost VMamba.

3. Methodology

3.1. Network pipeline

This work proposes a new change detector **VMI-CD**. Its overall architecture is shown in Fig. 2. It can be divided into three parts: VMamba feature extraction, change representation mapping, and change decoding. Firstly, bi-temporal features at different scales are extracted using a Siamese VMamba with interaction. It includes four stages, and a bi-temporal interaction is conducted at the beginning of the VSS blocks [9] in each stage. Let $f_{t,k}$ $(t \in \{t_1, t_2\}, k \in \{1,2,3,4\})$ represent the feature generated at each stage corresponding to the input image I_t, whose scale is $1/2^{k+1}$ of the original input. The process of bi-temporal feature extraction can be formulated as,

$$f_{t_1,1}, f_{t_2,1} = \mathcal{M}_1(\mathcal{I}(f_{t_1,0}, f_{t_2,0})), \tag{1}$$

$$f_{t_1,k}, f_{t_2,k} = \mathcal{M}_k(\mathcal{I}(f^{\downarrow 2}_{t_1,k-1}, f^{\downarrow 2}_{t_2,k-1})),\ k = 2,3,4, \tag{2}$$

where $\mathcal{I}(\cdot)$ refers to the interaction module BFIM. $\mathcal{M}_k(\cdot)$ indicates the VMamba feature operator at the stage k, which consists of N_k VSS blocks. After that, the change-aware feature is produced from the generated bi-temporal features, which can be described as,

$$F_k = \mathcal{S}_k(f_{t_1,k}, f_{t_2,k}), \tag{3}$$

where $\mathcal{S}(\cdot)$ denotes the change representation mapping module, which comprises a CS2M and a difference extraction operation. Finally, a vanilla change decoder like that in [12] is employed to decode the multi-level change representations for change classification. The process is formulated as follows,

$$\overline{F}_3 = F_{4\uparrow} + F_3, \tag{4}$$

$$\overline{F}_{k-1} = F_{k-1} + \mathcal{R}_k(\overline{F}_k)_{\uparrow 2},\ k = 2,3, \tag{5}$$

$$L = C_1(\mathcal{R}_k(\overline{F}_1))_{\uparrow 4}, \tag{6}$$

$$CM = \text{argmax}(L), \tag{7}$$

where $\mathcal{R}(\cdot)$ represents the ResNet basic block [33]. $C_1(\cdot)$ refers to a convolution layer with the kernel size 1. L is the predicted logits and CM is the obtained change map.

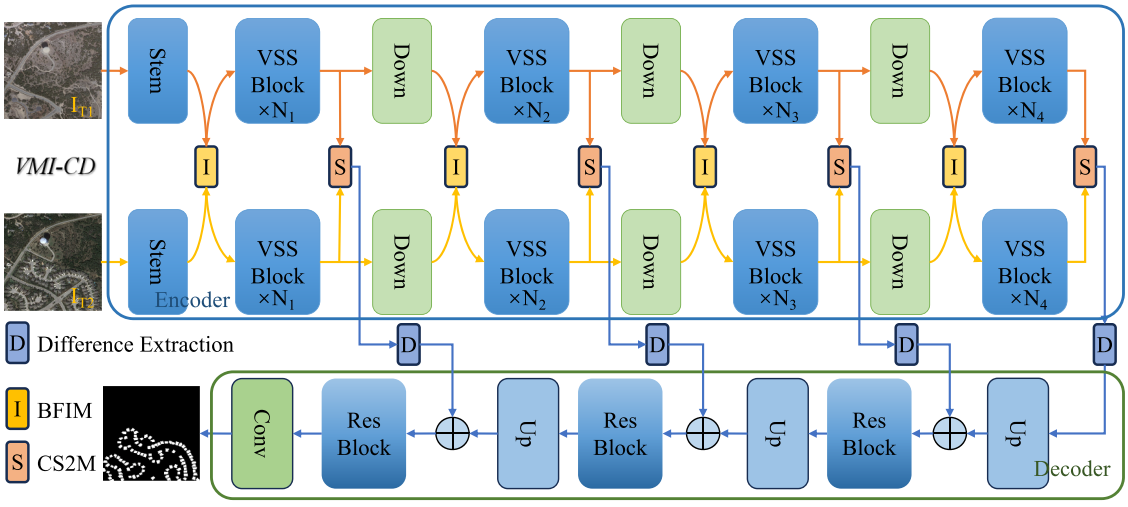

Fig. 2. The pipeline of the proposed VMI-CD, where the BFIM refers to the bi-temporal feature interaction module and the CS2M refers to the channel-spatial selection module. The VSS Block is the visual state space block, and the Res Block is the basic block in ResNet. The "Down" and "Up" indicate the downsample and upsample operations.

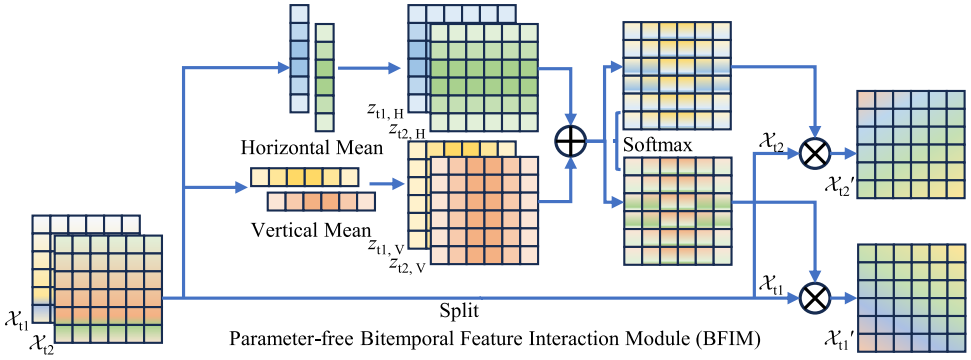

Fig. 3. The schematic diagram of the BFIM.

3.2. Bi-temporal feature interaction

As introduced in the pipeline of our method, we employ the Siamese VMamba for bi-temporal image feature extraction, thus locating change. Significantly different from video or single image inputs, bi-temporal images used for change detection are captured at discrete times, which have different distributions due to light variation or season changes. Therefore, we argue that simply considering VMamba as a feature extractor to extract features from bi-temporal images separately cannot fully unlock its potential for change detection. The domain differences in bi-temporal images can not be well perceived.

With the above perspective, a parameter-free bi-temporal feature interaction module is designed, which can be easily integrated into the Siamese VMamba. Its structure is shown in Fig. 3. Let $\mathcal{X}_{t1} \in \mathbb{R}^{c\times h\times w}, \mathcal{X}_{t2} \in \mathbb{R}^{c\times h\times w}$ represent the bi-temporal features inputted into the BFIM. h and w indicate the height and width of the feature map respectively. First, given a position (m, n), the features along its horizontal and vertical directions are aggregated, which can be formulated as

$$z_{t,V}(m,n) = \frac{1}{hc}\sum_{k=1}^{c}\sum_{i=1}^{h}\mathcal{X}_t(i,n), \tag{8}$$

$$z_{t,H}(m,n) = \frac{1}{wc}\sum_{k=1}^{c}\sum_{j=1}^{w}\mathcal{X}_t(m,j). \tag{9}$$

Then, the interaction operation is introduced. Its computation is denoted as follows,

$$a_t = e^{z_{t,V}+z_{t,H}}, \tag{10}$$

$$\mathcal{X}'_{t1} = \frac{a_{t2}}{a_{t1}+a_{t2}} * \mathcal{X}_{t1}, \mathcal{X}'_{t2} = \frac{a_{t1}}{a_{t1}+a_{t2}} * \mathcal{X}_{t2}, \tag{11}$$

where $\mathcal{X}'_t$ represents the modulated feature. It can be noted that the interplay between bi-temporal features is achieved in Eq. 11. It can contrastively scale the original bi-temporal features to reduce the style distribution gap caused by the difference in imaging conditions (e.g., lighting intensity) of bi-temporal images, thus facilitating subsequent discriminative change representation extraction. Why adopt the aggregated feature along horizontal and vertical directions for bi-temporal interaction? The study in [9] indicates that the effective receptive field (ERF) of VMamba is cross-shaped, which can be attributed to the order limitation of cross-scan paths. To this end, the criss-cross aggregation operation helps BFIM better capture the spatial context information, improving its tuning ability to bi-temporal features. In summary, by performing interactive modulation based on aggregated features along horizontal and vertical directions, BFIM learns to adapt the distribution of bi-temporal representations. Thereby, by embedding BFIM, the potential of the Siamese VMamba can be further unleashed for change detection.

3.3. Channel-Spatial selection

Discriminative difference extraction is fundamental for accurate change detection. It can be achieved by improving consistency in unchanged regions and discrimination in changed objects. As for consistency, integrating the BFIM into the Siamese VMamba can help weaken the impact of pseudo-changes and alleviate the domain gap between the bi-temporal images. This section explores how to highlight the differences in the changed regions. A channel-spatial selection module is developed to refine bi-temporal representations. Its structure is shown in Fig. 4(a). Let f_{t1} and f_{t2} represent features at the pre- and post-temporal in respective. First, the refinement in the channel dimension is implemented as follows,

$$x_t = e^{GAP_s(g_1(f_t))}, \tag{12}$$

$$f'_{t1} = \frac{x_{t1}}{x_{t1}+x_{t2}} * f_{t1}, f'_{t2} = \frac{x_{t2}}{x_{t1}+x_{t2}} * f_{t2}, \tag{13}$$

where $g_1(\cdot)$ refers to a depth-wise convolution with kernel size 3, following a lightweight design idea. GAP_s is a global average pooling operation along the spatial dimension. Next, the refinement in the spatial dimension can be formulated as,

$$y_t = e^{GAP_c(g_2(f'_t))}, \tag{14}$$

$$f^*_{t1} = \frac{y_{t1}}{y_{t1}+y_{t2}} * f'_{t1}, f^*_{t2} = \frac{y_{t2}}{y_{t1}+y_{t2}} * f'_{t2}, \tag{15}$$

where GAP_c is a global average pooling operation along the channel dimension. $g_2(\cdot)$ also is a depth-wise convolution with kernel size 3. Finally, the original feature f_t is added to the modulated embedding f^*_t by a residual connection, obtaining E_t for difference extraction.

According to Eqs. 13 and 15, the channel and spatial selection is reached by a simple softmax function. This process relatively enlarges the bi-temporal feature differences, enhancing the representation discrimination and facilitating focus on changed regions. In contrast, the interaction operation in Eq. 11 is reversed from that in Eqs. 13 and 15 to reduce the impact of domain shift noise in bi-temporal images. From Fig. 2, these two operations are performed in different parts to generate beneficial features for change detection.

3.4. Difference extraction

The difference extraction is implemented based on the refined bi-temporal feature by the CS2M. This work adopts two commonly used difference modeling strategies to generate a fused change representation: absolute feature difference and feature concatenation. The former focuses on binary changes (i.e., unchanged vs. changed), emphasizing the presence of changes by highlighting discrepancies between bi-temporal features. The latter helps to capture changes that depend on temporal order, such as transitions from cropland to buildings or vice versa. By retaining the complete temporal context, feature concatenation enables the model to infer both the direction and the semantic nature of

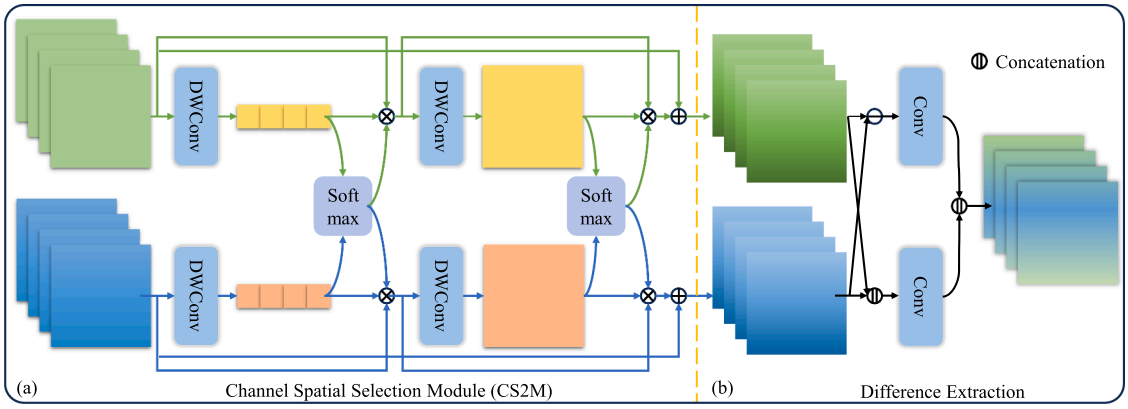

Fig. 4. The schematic diagram of the CS2M.

changes. Without concatenation, the change representation cannot reveal directional information, while without the absolute difference, the representation is prone to temporal-order bias in binary change detection. Therefore, these two strategies complement each other in mapping change representations, ensuring the wide applicability of the proposed model across tasks such as multi-class change detection, class-agnostic binary change detection, and building-specific change detection. In detail, as shown in Fig. 4(b), the difference extraction can be formulated as

$$F_a = C_a(|E_{t1} - E_{t2}|), \tag{16}$$

$$F_b = C_b(E_{t1}||E_{t2}), \tag{17}$$

$$F_c = ReLU(Norm(C_c(F_a||F_b)), \tag{18}$$

where $||$ is the concatenation operation. $C_x, x \in \{a, b, c\}$ represents one convolutional layer with kernel size 1. $Norm$ and $ReLU$ indicate the normalization and the activation function respectively.

3.5. Loss function

In this work, we explore the performance superiority of the proposed model on various change detection datasets, including multi-class change detection, class-specific change detection, and class-agnostic change detection. To maintain consistency, the cross-entropy loss is used to supervise pixel-wise prediction on them, which can be denoted as

$$\mathcal{L}_{ce} = -\frac{1}{|M|}\sum_{i=1}^{M}\sum_{c=1}^{N} y_{i,c} log(p_{i,c}), \tag{19}$$

where N is the number of given classes. For example, $N = 2$ for the class-agnostic change detection datasets, which consist of two categories: unchanged and changed. $y_{i,c}$ is the ground truth label of pixel i. Besides, sample imbalance is a prominent challenge in change detection tasks. To alleviate it, the Lovász-Softmax loss [34] is employed. The total loss is formulated as

$$\mathcal{L}_{total} = \mathcal{L}_{ce} + \lambda\mathcal{L}_{lov}, \tag{20}$$

where λ is a balance coefficient.

4. Experiments

4.1. Datasets

In this work, to fully demonstrate the competitiveness of VMI-CD, five datasets with different tasks are introduced to evaluate its performance. In detail, two multi-class change detection datasets MCLC-CD and JL1-CD, two representative building change detection datasets, and one class-agnostic change detection dataset are adopted. The following introduces them one by one.

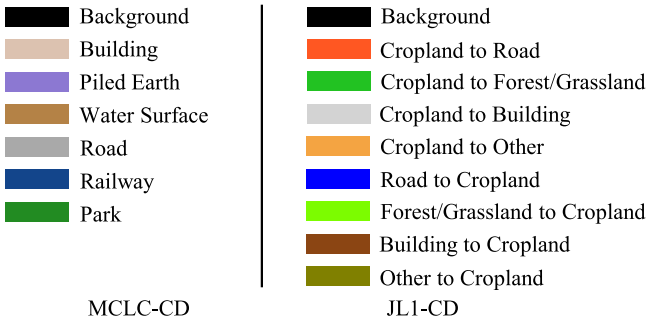

Fig. 5. The legends for the categories in the MCLC-CD and JL1-CD datasets.

MCLC-CD: This is a multi-category land cover (MCLC) change detection dataset with six change types annotated other than the background, which is provided by the ISPRS 2024 TC I Contest on Intelligent Interpretation for Multi-modal Remote Sensing Application. The six types are shown in Fig. 5. The resolution of the images in this dataset is two meters. The total number of images available is 2500, of which 2000 are used for training and 500 are used for testing. Besides, the image size is 512×512. More information can be found at the official website[1].

JL1-CD: This dataset is released at the first Jilin-1 Remote Sensing Application Innovation Competition. It is a multi-class cropland change detection dataset, where the images are captured by the Jilin-1 satellite. The detailed classes are given in Fig. 5. The total number of images available is 6000, and the image size is 256×256. 4200 of them are used for training, 600 for validation, and 1200 for testing. More information can be found on its relevant website[2] and publication [35].

LEVIR-CD+: This dataset [36] is an extended version of the building change detection dataset LEVIR-CD [37], with images of size 1024×1024 pixels. Experiments adhere to the official training and testing data settings provided. In detail, 637 images for training and 348 images for testing.

WHU-CD: This is a high-resolution building change detection dataset [38]. It covers an area affected by an earthquake, where many buildings were either reconstructed or newly built in the subsequent years. The dataset includes a pair of images with a size of 32508×15354 pixels and is split into training and testing sets. Specifically, there are 315 images for training and 165 images for testing.

SYSU-CD: The dataset [39] is designed for class-agnostic change detection and includes aerial imagery of Hong Kong taken between 2007 and 2014. Complex scenarios such as densely packed high-rise buildings and active ports are captured in it. Besides, it consists of 20,000 im-

[1] https://www.gaofen-challenge.com
[2] https://www.jl1mall.com/contest/

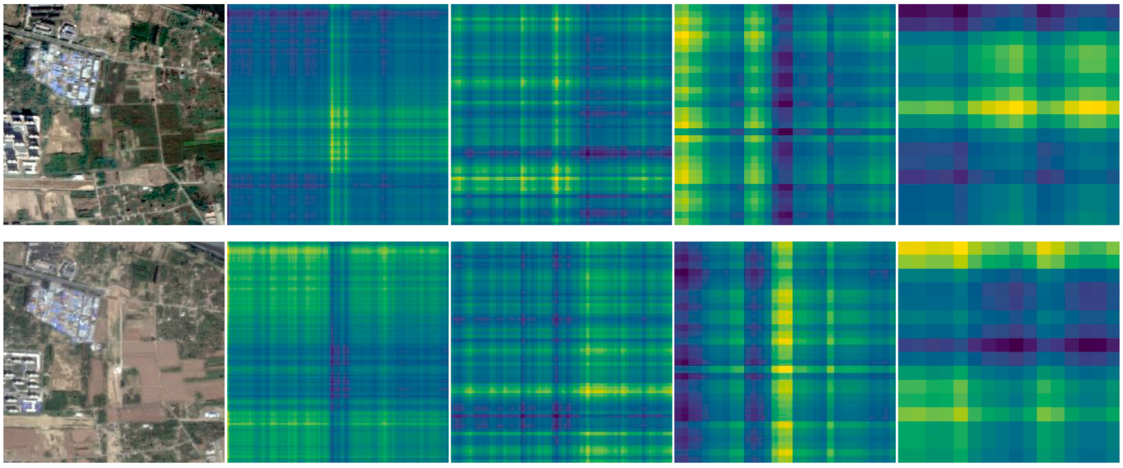

Fig. 6. The visual analysis of BFIM. The interaction maps of four stages in the VMI-CD are shown in turn from left to right.

Table 1
The hyperparameter settings on the five datasets.

Setting	Dataset				
	MCLC	JL1	SYSU	LEVIR	WHU
Batch Size	2	8	8	2	2
Epoch	50	200	50	200	200
Image Size	512	256	256	512	512

Table 2
The ablation studies conducted on the MCLC-CD dataset, with the best values in bold (%).

+CS2M	+BFIM	mIoU	mF1	Δ
		49.35	63.53	–
✓		49.57	63.79	0.05
	✓	50.41	64.56	–
✓	✓	**50.73**	**64.93**	0.05

Table 3
The effectiveness exploration of the interaction operation in BFIM. The best values are in bold (%).

Metrics	Baseline	+BFIM (w/o Eq. 11)	+BFIM (w/ Eq. 11)
mIoU	49.35	49.92	**50.41**
mF1	63.53	64.07	**64.56**

age pairs, each with a resolution of 256 × 256 pixels, and has been augmented with techniques like random flipping and rotation. The dataset is divided following the official publicly available settings.

4.2. Ablation studies and discussion

4.3. Implementation details

The hyperparameter settings for the proposed method are summarized in Table 1. Due to factors such as image size, dataset size, and memory limitations, different datasets require distinct configurations for training epochs and batch size. Additionally, the balanced coefficient in the total loss function is set to 0.75. The Siamese VMamba is built based on the VMamba-S [9]. During training, the AdamW optimizer is used to update the model parameters, with an initial learning rate of 0.0001. A polynomial decay strategy is applied to adjust the learning rate throughout the training process. All experiments are implemented on one NVIDIA 3090 GPU. In addition, several evaluation metrics, including precision (P), recall (R), F1-score (F1), intersection over union (IoU), overall accuracy (OA), mean IoU (mIoU), and mean F1 (mF1), are used to assess the performance of different methods.

The ablation studies are implemented on the MCLC-CD dataset. The last checkpoint is employed for testing to ensure stability. Four combinations are explored for comparison to verify the effectiveness of the proposed CS2M and BFIM modules. The ablation results are shown in Table 2. Compared to the experiment without any modules, the two proposed modules improve accuracy, confirming their effectiveness. Notably, models with only CS2M or BFIM outperform the baseline, demonstrating their ability to identify changes accurately. The best performance is achieved when both modules are combined, as they help reduce pseudo-change impact and enhance sensitivity to subtle changes, thus the performance of the model can be improved. After introducing the BFIM and CS2M, the model parameters only increased by 0.05 %, about 26K.

Figs. 6 and 7 present the visual analysis of BFIM and CS2M, respectively. The interaction maps and spatial selection maps are dynamic and adaptive, adjusting to both spatial regions and feature stages. This adaptability enables the refinement of bi-temporal features, enhancing subsequent change detection. In BFIM, the interaction maps, which result from bidirectional aggregation along both horizontal and vertical directions, exhibit a strip-like structure. This feature distinctly contrasts with the spatial selection maps. The strip-shaped context aggregation strategy aligns seamlessly with the cross-scanning approach in VMamba, effectively aiding in the capture of discriminative features. For CS2M, The selection maps in stages 3 and 4 effectively focus on the more salient regions in bi-temporal images, enlarging the bi-temporal differences. The early selection maps exhibit a weaker effect compared to the later ones, which can be attributed to the weak semantic discrimination ability during the earlier stages. Overall, BFIM and CS2M demonstrate effective modulation capabilities for bi-temporal features, thereby improving the accuracy of change detection.

In addition, the importance of the interaction operation in BFIM is explored. Specifically, the swap modulation expressed in Eq. 11 is replaced by the operation without swap. The latter aims to force the bi-temporal features to separate, rather than narrow their domain differences. Although it is helpful for difference enhancement, the pseudo-change can not be reduced. We believe that it is beneficial to encode bi-temporal features with a weak domain gap. Table 3 reports the comparative results. It can be shown that introducing the swap interaction operation can significantly improve the performance of the model.

Discussion: In this part, we discuss the effectiveness of BFIM in boosting other backbone networks. Table 4 reports the quantitative results. First, the model without being equipped with the BFIM is trained. Then, BFIM is introduced to explore its effect on different backbones. BFIM effectively enhances the performance of the ResNet18 and VMamba-S. However, it produces a negative effect on MiT-b2. We be-

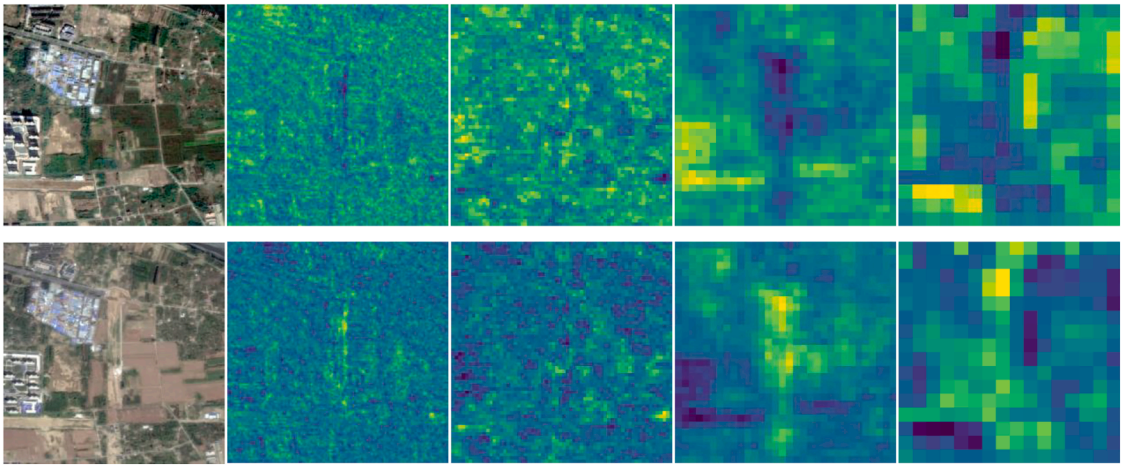

Fig. 7. The visual analysis of CS2M. The spatial selection maps of four stages in the VMI-CD are shown in turn from left to right.

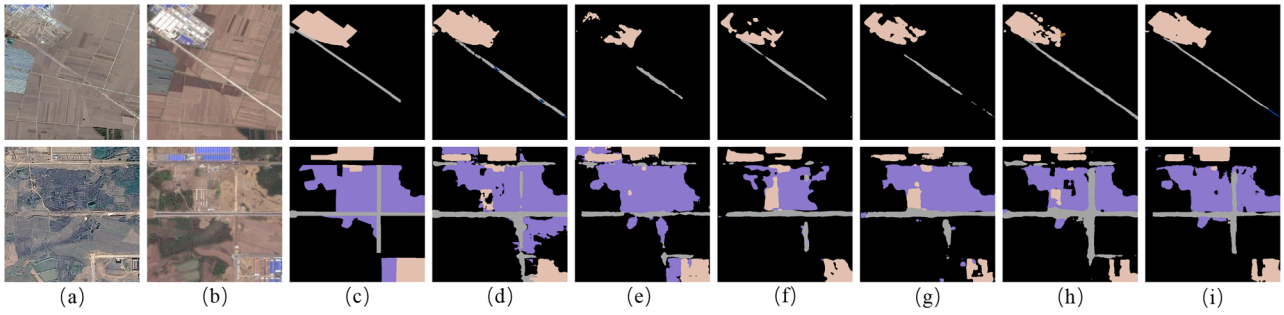

Fig. 8. The qualitative visual comparison on the MCLC-CD dataset. (a) Pre-temporal image. (b) Post-temporal image. (c) Ground Truth. (d) CLAFA. (e) Changer. (f) BAN-BiT. (g) BAN-CF. (h) MambaCD. (g) VMI-CD.

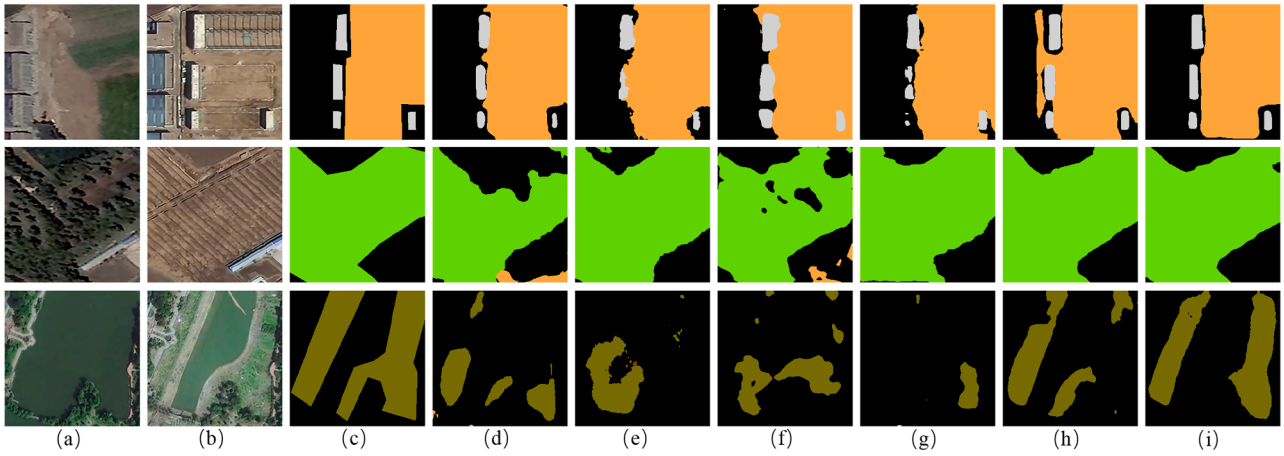

Fig. 9. The qualitative visual comparison on the JL1-CD dataset. (a) Pre-temporal image. (b) Post-temporal image. (c) Ground Truth. (d) CLAFA. (e) Changer. (f) BAN-BiT. (g) BAN-CF. (h) MambaCD. (g) VMI-CD.

Table 4
The effect of BFIM applied to different backbone networks. The best values are in bold (%).

Backbone	+BFIM	mIoU	mF1
ResNet18		40.53	54.19
	✓	41.54	55.38
MiTb2		46.47	60.77
	✓	45.63	59.60
VMamba-S		49.57	63.79
	✓	50.73	64.93

lieve that the phenomenon originates from the customized design of BFIM for VMamba. The cross-shaped aggregation in BFIM could disrupt the feature learning process in MiT-b2, as its architecture and ERF are different from that in VMamba [9]. This mismatch can lead to suboptimal performance. Meanwhile, as the ERF of CNNs is typically local, the introduction of BFIM enhances long-range context perception, leading to improved performance.

4.4. Comparison with the advanced methods

The five advanced and popular state-of-the-art methods are compared with the proposed VMI-CD, including CLAFA [31], Changer[24],

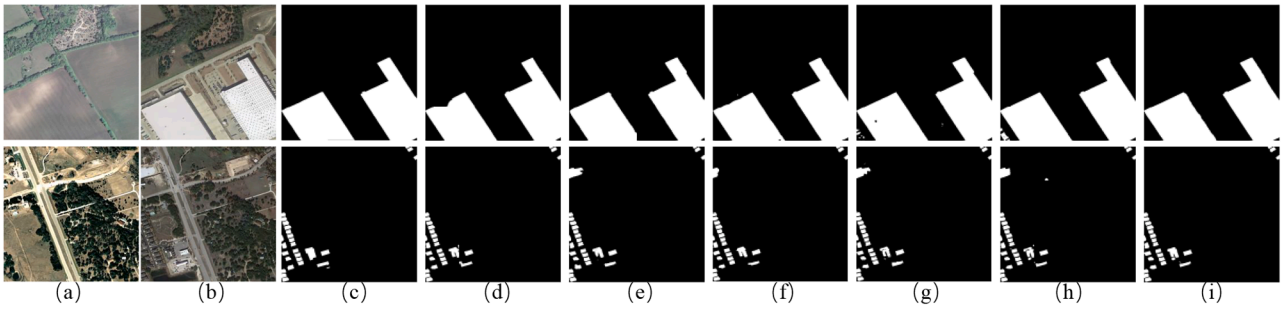

Fig. 10. The qualitative visual comparison on the LEVIR-CD+ dataset. (a) Pre-temporal image. (b) Post-temporal image. (c) Ground Truth. (d) CLAFA. (e) Changer. (f) BAN-BiT. (g) BAN-CF. (h) MambaCD. (g) VMI-CD.

Table 5
The quantitative comparison on the MCLC-CD and JL1-CD datasets, with the best values highlighted in bold (%).

Methods	MCLC-CD		JL1-CD	
	mIoU	mF1	mIoU	mF1
CLAFA	44.73	58.87	50.53	57.88
Changer	42.92	56.83	78.91	88.02
BAN-BiT	44.30	58.36	71.89	83.30
BAN-CF	43.64	57.36	75.42	85.79
MambaCD	49.62	64.28	86.84	92.91
VMI-CD	**50.73**	**64.93**	**88.09**	**93.63**

Table 6
The quantitative comparison on the LEVIR-CD+ dataset, with the best values highlighted in bold (%).

Methods	LEVIR-CD+				
	P	R	F1	IoU	OA
CLAFA	88.22	86.05	87.12	77.18	98.97
Changer	84.32	86.27	85.29	74.35	98.79
BAN-BiT	87.78	84.82	86.28	75.87	98.90
BAN-CF	88.15	85.15	86.62	76.40	98.93
MambaCD	**88.70**	85.85	87.25	77.39	**98.98**
VMI-CD	88.32	**86.29**	**87.30**	**77.46**	**98.98**

Table 7
The quantitative comparison on the WHU-CD dataset, with the best values highlighted in bold (%).

Methods	WHU-CD				
	P	R	F1	IoU	OA
CLAFA	94.61	92.67	93.63	88.02	99.54
Changer	94.82	91.97	93.37	87.56	99.52
BAN-BiT	94.69	90.66	92.63	86.28	99.48
BAN-CF	**96.56**	89.29	92.78	86.54	99.49
MambaCD	94.51	92.73	93.61	87.99	99.54
VMI-CD	95.85	**92.82**	**94.31**	**89.23**	**99.59**

BAN-BiT [25], BAN-CF [25], MambaCD [12]. All compared models are retrained and their best performing models are used for comparison with ours.

Results on the MCLC-CD and the JL1-CD Datasets: Fig. 8 is a visualization comparison of different change detection methods on the MCLC-CD dataset. From the two pairs of images in the given dataset, VMI-CD achieves the best overall prediction results. Although compared methods all can capture long-range context dependency, VMI-CD achieves higher prediction completeness and lower false alarms. In particular, in the second example, the piled earth and the road are mixed in one area and difficult to distinguish. Only VMI-CD accurately identifies corresponding regions. This may benefit from spatial and channel selection of CS2M, which enhances the discrimination of different types of changes. The visualization comparison of the discussed methods on the JL1-CD dataset is shown in Fig. 9. Three sets of visualization results of different land cover changes are selected. The first row shows the changes from cropland to buildings and others. Compared with other methods, VMI-CD can predict the changes of buildings more accurately and has clear boundaries with other types. In the second row of images, VMI-CD does not misjudge the change type. In the third example, the change between the two images is relatively subtle. The prediction results of other change detection methods deviate greatly from the actual changes. In contrast, the prediction of VMI-CD only misses a small fraction, showing superior performance.

Meanwhile, Table 5 report the quantitative comparison on the two datasets. On the MCLC-CD dataset, the accuracy of VMI-CD is the best, and compared with the second-best method MambaCD, the mloU and mF1 prediction accuracy are both improved by more than 0.6 %. On the JL1-CD dataset, VMI-CD still shows superior performance. Compared with the second-best effect, the mloU and mF1 are improved by more than 0.7 %. From the results on the two multi-class change detection datasets, the performance of VMI-CD confirms its superiority.

Results on the LEVIR-CD+ and WHU-CD Datasets: Figs. 10 and 11 show the visual results on the LEVIR-CD+ and WHU-CD datasets, respectively. Several typical examples with obvious differences in light intensity are selected for comparison. In general, our method is the most accurate in identifying the changed regions in the given examples. In the second pair of images on the LEVIR-CD+ dataset, multiple methods misidentify the region in the upper left corner as a changed building. Besides, on the WHU-CD dataset, their prediction completeness is inferior to ours. In contrast, VMI-CD has high edge smoothness and is not affected by shadows. The impressive results of VMI-CD can be attributed to BFIM and CS2M, which can effectively mitigate the impact of domain differences and improve the sensitivity of VMI-CD to changes, enabling VMI-CD to identify buildings that have changed and reduce false positives more accurately. Tables 6 and 7 report the quantitative comparison of different change detection methods on the LEVIR-CD+ and WHU-CD datasets, respectively. In both datasets, the overall performance of VMI-CD is optimal. Although the precision is suboptimal, it balances precision and recall better. Overall, the superiority of VMI-CD used for building change detection is also confirmed.

Results on the SYSU-CD Dataset: This task is more challenging than binary building change detection, as demonstrated by the visual comparison of six methods in Fig. 12. In the first image pair, which illustrates changes in roads and vegetation, BAN-CF produces the least accurate prediction. The second image pair, which shows changes in buildings and hardened ground, reveals that MambaCD misses many positive instances. In the third example, it can be seen that the prediction of CLAFA is disturbed by the shadow. Other advanced methods have large deviations in the recognition of changed areas. Overall, in the given examples

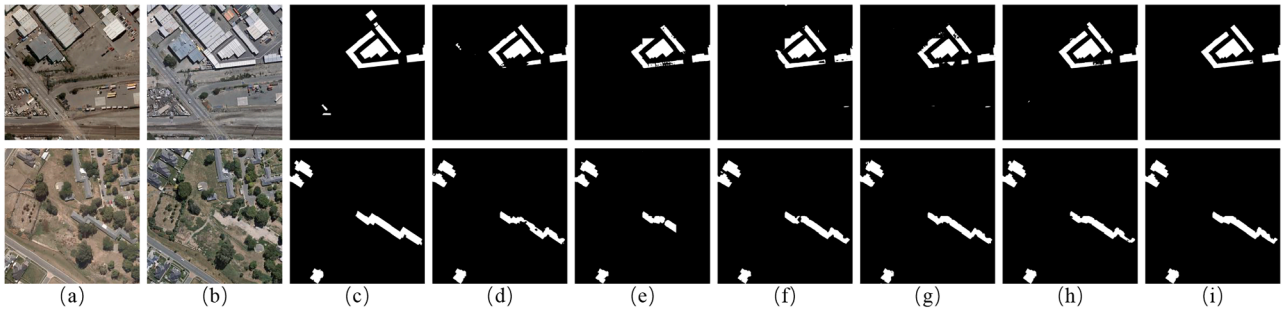

Fig. 11. The qualitative visual comparison on the WHU-CD dataset. (a) Pre-temporal image. (b) Post-temporal image. (c) Ground Truth. (d) CLAFA. (e) Changer. (f) BAN-BiT. (g) BAN-CF. (h) MambaCD. (g) VMI-CD.

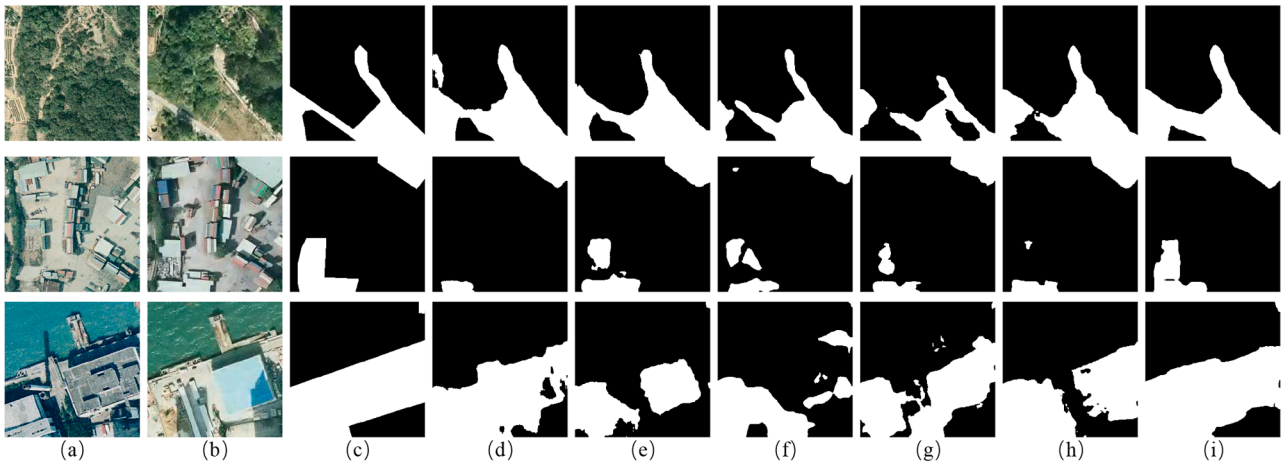

Fig. 12. The qualitative visual comparison on the SYSU-CD dataset. (a) Pre-temporal image. (b) Post-temporal image. (c) Ground Truth. (d) CLAFA. (e) Changer. (f) BAN-BiT. (g) BAN-CF. (h) MambaCD. (g) VMI-CD.

Table 8
The quantitative comparison on the SYSU-CD dataset, with the best values highlighted in bold (%).

Methods	SYSU-CD				
	P	R	F1	IoU	OA
CLAFA	86.79	79.59	83.03	70.98	92.33
Changer	87.44	80.38	83.77	72.07	92.65
BAN-BiT	82.62	79.17	80.85	67.86	91.16
BAN-CF	87.93	76.52	81.83	69.25	91.99
MambaCD	86.31	**80.51**	83.31	71.39	92.39
VMI-CD	**88.64**	79.90	**84.05**	**72.48**	**92.85**

Table 9
Comparison of efficiency among different methods.

Methods	Backbones	Params (M)	FLOPs (G)	mIoU (%)
CLAFA	MobileNetv2	6.71	67.17	44.73
Changer	MiT-b0	3.46	8.53	42.92
BAN-BiT	ViT-B + ResNet18	68.23	65.46	44.30
BAN-CF	ViT-B + MiT-b0	68.89	38.48	43.64
MambaCD	VMamba-S	54.00	114.83	49.62
VMI-CD	VMamba-S	50.98	90.88	50.73

with diverse land cover types and imaging conditions, the performance of VMI-CD is optimal. Besides, Table 8 report the quantative results. Compared with other advanced methods, VMI-CD shows superior performance. Although VMI-CD has a suboptimal recall, the other four metrics are the highest. It is competitive for class-agnostic change detection.

The comparison of efficiency: The evaluation is performed on a pair of 512×512 images. The results are reported in the Table 9. It can be seen that CLAFA and Changer have relatively low parameter counts due to their lightweight backbones, but their accuracy is significantly lower than MambaCD and VMI-CD. Compared with BAN-BiT, BAN-CF, and MambaCD, VMI-CD has the lowest number of parameters, thanks to the lightweight modules introduced on top of the VMamba backbone. However, the FLOPs of VMI-CD are slightly higher than BAN-BiT and BAN-CF. This counter-intuitive result is because these two methods resize the input images to 224×224 to match the pretrained weights of ViT, reducing their computational cost. The full 512×512 resolution is used in VMI-CD and MambaCD, which better preserves accuracy. A fairer comparison shows that VMI-CD reduces FLOPs by about 20G compared with the previous state-of-the-art method MambaCD, while achieving the highest accuracy. These results demonstrate that VMI-CD achieves a favorable balance between efficiency and performance.

In summary, VMI-CD provides satisfactory performance across different change detection tasks and maintains competitive efficiency. Based on the qualitative and quantitative results across the five datasets, it can effectively mitigate the negative impact caused by differences in the bi-temporal image domain and demonstrate strong capability in capturing differences. Compared with the previous MambaCD, the potential of VMamba is further unlocked for change detection.

4.5. Application case

In this section, a practical application of the proposed method is presented in Fig. 13. The case study is from Wangjia Village, Chang'an District, Xi'an, where the pre-temporal image is captured in 2016 and the post-temporal image in 2021. This region exhibits various change types,

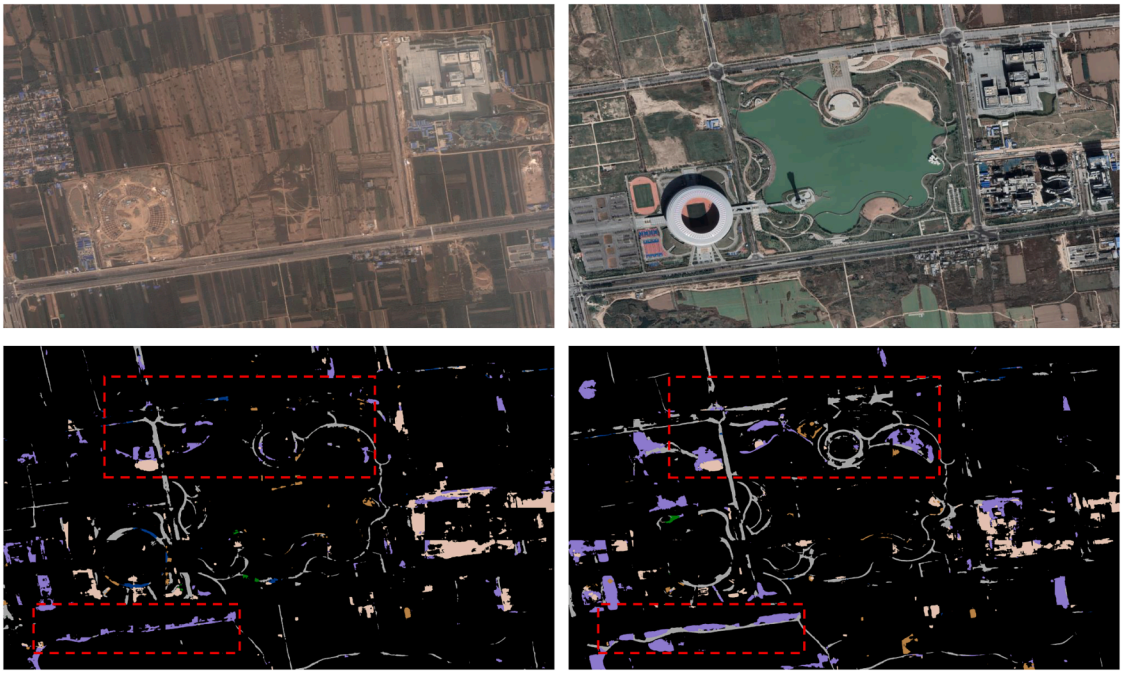

Fig. 13. A practical application case. On the second row, the left is the testing results of MambaCD and the right is that of VMI-CD.

including roads, buildings, water bodies, and piled earth. The model used for testing is trained on the MCLC-CD dataset. The two images are not included in the MCLC-CD dataset, and their resolution is 0.5 m. As shown in Fig. 13, where it can be observed that VMI-CD performs slightly better than MambaCD in the marked regions. Nevertheless, both MambaCD and VMI-CD fail to detect the changed artificial lake, while the pond changed from farmland is effectively distinguished. This discrepancy can be attributed to the resolution difference and the loss of contextual information in large-scale changes caused by image cropping. Meanwhile, the newly built playground and gymnasium are also not reliably detected. These findings point to two promising directions for future research: (1) mitigating the loss of contextual information caused by image patches in large-scale change detection tasks, and (2) improving the ability of the model to generalize in open-world scenarios, where differences in resolution or change categories often result in missed detections.

5. Conclusion

In this work, we investigate the potential of VMamba for change detection by addressing the unique characteristics of non-continuous bi-temporal images, such as light variations, season changes, and so on. In detail, distinct from a single image or continuous video analysis, we develop a tailor-made model VMI-CD for change detection to adapt bi-temporal images with sudden changes in content and style. First, during the feature encoding stage, several parameter-free bi-temporal feature interaction modules are injected into the different levels of Siamese VMamba. It effectively enhances the ability of VMamba to perceive domain differences between bi-temporal images. Notably, these interaction operations are seamlessly aligned with the effective receptive field of VMamba, making them a tailored solution to unleash the potential of VMamba for change detection. Additionally, a feature selection module is designed to selectively focus on key representations within the bi-temporal features from spatial and channel dimensions, emphasizing the changes and thereby helping to enhance the extracted differences. Thanks to the lightweight design of both modules, the interaction and selection mechanisms introduce only a minimal number of additional parameters. Experimental results on five datasets with different change detection tasks illustrate that VMI-CD outperforms the prior state-of-the-art methods. More importantly, this work demonstrates that the full potential of VMamba for change detection remains underexplored. The proposed simple yet effective interaction and selection mechanisms enhance its performance by improving domain difference perception and amplifying change features, offering valuable insights for future research. However, a limitation of our approach is that the interaction mechanism does not effectively enhance the performance of Transformer-based models. In future work, we will further explore unified interaction and selection strategies applicable to different backbone networks.

CRediT authorship contribution statement

Mingwei Zhang: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation; **Yuan Jiang:** Writing – review & editing, Writing – original draft, Visualization, Validation, Methodology, Investigation, Formal analysis, Data curation; **Qi Wang:** Writing – review & editing, Validation, Resources, Project administration, Investigation, Funding acquisition, Data curation, Conceptualization.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grant 62471394 and U21B2041.

Data availability

The data/code link is provided in my manuscript.

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.

References

[1] N. Quarmby, J. Cushnie, Monitoring urban land cover changes at the urban fringe from SPOT HRV imagery in south-east england, Int. J. Remote Sens. 10 (6) (1989) 953–963.

[2] Z. Zheng, Y. Zhong, J. Wang, A. Ma, L. Zhang, Building damage assessment for rapid disaster response with a deep object-based semantic change detection framework: from natural disasters to man-made disasters, Remote Sens. Environ. 265 (2021) 112636.
[3] P. Coppin, M. Bauer, Digital change detection in forest ecosystems with remote sensing imagery, Remote Sens. Rev. 13 (3–4) (1996) 207–234.
[4] J. Zou, W. Zhang, Q. Li, Q. Wang, MOSAIC-Tracker: mutual-enhanced occlusion-aware spatiotemporal adaptive identity consistency network for aerial multi-object tracking, ISPRS J. Photogramm. Remote Sens. 229 (2025) 138–154.
[5] R. Daudt, B. Le, A. Boulch, Fully convolutional siamese networks for change detection, in: Proc. IEEE International Conference on Image Processing, 2018, pp. 4063–4067.
[6] W. Bandara, V. Patel, A transformer-Based siamese network for change detection, in: Proc. IEEE International Geoscience and Remote Sensing Symposium, 2022, pp. 207–210.
[7] 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.
[8] H. Chen, Z. Qi, Z. Shi, Remote sensing image change detection with transformers, IEEE Trans. Geosci. Remote Sens. 60 (2022) 1–14.
[9] Y. Liu, Y. Tian, Y. Zhao, H. Yu, L. Xie, Y. Wang, Q. Ye, J. Jiao, Y. Liu, VMAmba: visual state space model, in: The Thirty-eighth Annual Conference on Neural Information Processing Systems, 2024.
[10] K. Li, X. Li, Y. Wang, Y. He, Y. Wang, L. Wang, Y. Qiao, Videomamba: state space model for efficient video understanding, in: Proc. European Conference on Computer Vision, 2024, pp. 237–255.
[11] Y. Xiao, Q. Yuan, K. Jiang, Y. Chen, Q. Zhang, C.-W. Lin, Frequency-assisted mamba for remote sensing image super-Resolution, IEEE Trans. Multimedia (2024) 1–14.
[12] 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.
[13] J.N. Paranjape, C. de Melo, V.M. Patel, A Mamba-based Siamese Network for Remote Sensing Change Detection, arXiv preprint arXiv:2407.06839 (2024).
[14] A. Gu, K. Goel, C. Re, Efficiently modeling long sequences with structured state spaces, in: International Conference on Learning Representations, 2022, pp. 1–12.
[15] H. Mehta, A. Gupta, A. Cutkosky, B. Neyshabur, Long range language modeling via gated state spaces, in: The Eleventh International Conference on Learning Representations, 2023, pp. 1–16.
[16] J.T.H. Smith, A. Warrington, S. Linderman, Simplified state space layers for sequence modeling, in: The Eleventh International Conference on Learning Representations, 2023, pp. 1–13.
[17] A. Gu, T. Dao, Mamba: linear-Time sequence modeling with selective state spaces, in: First Conference on Language Modeling, 2024, pp. 1–16.
[18] L. Zhu, B. Liao, Q. Zhang, X. Wang, W. Liu, X. Wang, Vision mamba: efficient visual representation learning with bidirectional state space model, in: Icml, 2024, pp. 1–12.
[19] Y. Yang, C. Ma, J. Yao, Z. Zhong, Y. Zhang, Y. Wang, Remamber: referring image segmentation with mamba twister, in: Proc. European Conference on Computer Vision, 2024, pp. 108–126.
[20] B. Lin, W. Jiang, P. Chen, Y. Zhang, S. Liu, Y.-C. Chen, MTMamba: enhancing multi-task dense scene understanding by mamba-Based decoders, in: Proc. European Conference on Computer Vision, 2024, pp. 314–330.
[21] C. Liu, K. Chen, B. Chen, H. Zhang, Z. Zou, Z. Shi, Rscama: remote sensing image change captioning with state space model, IEEE Geosci. Remote Sens. Lett. (2024).
[22] Y. Chen, X. Li, C. Luan, W. Hou, H. Liu, Z. Zhu, L. Xue, J. Zhang, D. Liu, X. Wu, L. Wei, C. Jian, J. Li, Cross-level interaction fusion network-based RGB-T semantic segmentation for distant targets, Pattern Recognit. 161 (2025) 111218.
[23] X. Yang, H. Liu, N. Wang, X. Gao, Bidirectional modality information interaction for visible-infrared person re-identification, Pattern Recognit. 161 (2025) 111301.
[24] S. Fang, K. Li, Z. Li, Changer: feature interaction is what you need for change detection, IEEE Trans. Geosci. Remote Sens. 61 (2023) 1–11.
[25] K. Li, X. Cao, D. Meng, A new learning paradigm for foundation model-Based remote-Sensing change detection, IEEE Trans. Geosci. Remote Sens. 62 (2024) 1–12.
[26] M. Zhang, H. Zheng, M. Gong, Y. Wu, H. Li, X. Jiang, Self-structured pyramid network with parallel spatial-channel attention for change detection in VHR remote sensed imagery, Pattern Recognit. 138 (2023) 109354.
[27] H. Zheng, M. Zhang, M. Gong, A.K. Qin, T. Liu, F. Jiang, Multi-scale hierarchical feature fusion network for change detection, Pattern Recognit. 161 (2025) 111266.
[28] H. Zheng, M. Gong, T. Liu, F. Jiang, T. Zhan, D. Lu, M. Zhang, HFA-Net: high frequency attention siamese network for building change detection in VHR remote sensing images, Pattern Recognit. 129 (2022) 108717.
[29] S. Tian, X. Tan, A. Ma, Z. Zheng, L. Zhang, Y. Zhong, Temporal-agnostic change region proposal for semantic change detection, ISPRS J. Photogramm. Remote Sens. 204 (2023) 306–320.
[30] Y. Gan, W. Xuan, H. Chen, J. Liu, B. Du, RFL-CDNet: towards accurate change detection via richer feature learning, Pattern Recognit. 153 (2024) 110515.
[31] G. Wang, G. Cheng, P. Zhou, J. Han, Cross-Level attentive feature aggregation for change detection, IEEE Trans. Circuits Syst. Video Technol. (2023).
[32] J. Liu, W. Xuan, Y. Gan, Y. Zhan, J. Liu, B. Du, An end-to-end supervised domain adaptation framework for cross-Domain change detection, Pattern Recognit. 132 (2022) 108960.
[33] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proc. IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.
[34] 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: Proc. IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 4413–4421.
[35] Q. Wang, W. Jing, K. Chi, Y. Yuan, Cross-difference semantic consistency network for semantic change detection, IEEE Trans. Geosci. Remote Sens. 62 (2024) 1–12.
[36] L. Shen, Y. Lu, H. Chen, H. Wei, D. Xie, J. Yue, R. Chen, S. Lv, B. Jiang, S2Looking: a satellite side-looking dataset for building change detection, Remote Sens. (Basel) 13 (24) (2021) 5094.
[37] H. Chen, Z. Shi, A spatial-temporal attention-based method and a new dataset for remote sensing image change detection, Remote Sens. (Basel) 12 (10) (2020) 1662.
[38] 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.
[39] Q. Shi, M. Liu, S. Li, X. Liu, F. Wang, L. Zhang, A deeply supervised attention metric-Based network and an open aerial image dataset for remote sensing change detection, IEEE Trans. Geosci. Remote Sens. 60 (2022) 1–16.