SCIENCE CHINA
Information Sciences

• RESEARCH PAPER •

November 2026, Vol. 69, Iss. 11, 212303:1–212303:15
https://doi.org/10.1007/s11432-025-4803-4

Mamba-based mutual reinforce diffusion model for unaligned hyperspectral-RGB multimodal change detection

Junying REN, Jiahui QU*, Yunsong LI & Wenqian DONG

State Key Laboratory of Integrated Services Networks, Xidian University, Xi'an 710071, China

Received 13 September 2025/Revised 29 December 2025/Accepted 11 February 2026/Published online 26 October 2026

Abstract Hyperspectral image (HSI) has become an important source of data for analyzing subtle surface changes; however, most existing HSI change detection (CD) methods are restricted to precisely aligned multitemporal homogeneous HSIs, which are difficult to obtain in real scenes. To overcome this limitation, HSI-RGB multimodal CD incorporates RGB images, enabling collaborative detection across different satellite platforms. Although this approach has been shown to have high potential, challenges such as modality differences and spatial deformation limit the direct application of existing HSI-CD methods. This study proposes a Mamba-based mutual reinforce diffusion model for unaligned HSI-RGB multimodal CD, achieving mutual reinforcement between image registration and CD through a conditional diffusion mechanism, thereby effectively identifying changes in terms of spatial deformation and modality difference. Specifically, a spatial-modal alignment diffusion network is designed to implement bidirectional super-resolution and registration, alleviating modality differences and spatial misalignment and enabling the accurate extraction of changed information. By applying a change-aware matching strategy, the registration block utilizes feedback from change maps to correct misalignment. Subsequently, a spectral-spatial-temporal Mamba CD module is proposed, which effectively models spatial, temporal, and spectral correlations of the aligned bitemporal images, achieving accurate CD results. Extensive experiments have demonstrated the superiority of the proposed method relative to existing approaches.

Keywords hyperspectral images, multimodal images, image registration, change detection, diffusion model

Citation Ren J Y, Qu J H, Li Y S, et al. Mamba-based mutual reinforce diffusion model for unaligned hyperspectral-RGB multimodal change detection. Sci China Inf Sci, 2026, 69(11): 212303, https://doi.org/10.1007/s11432-025-4803-4

1 Introduction

Change detection (CD) is a fundamental Earth observation technique for monitoring land surface changes by analyzing multitemporal remote sensing images of the same areas. Due to the growing need for timely change identification, CD has become widely used in agriculture [1], disaster assessment [2], and environmental protection [3]. Hyperspectral images (HSIs), rich in spectral information, enable the precise discrimination of ground objects [4–7]. Consequently, HSI has emerged as the preferred data source for CD applications; however, existing hyperspectral satellites often feature lower temporal resolutions, termed the revisit period [8–10], such as GF5-5days, HJ1-4days, and EnMAP-4days. Unlike HSI-CD, HSI-RGB multimodal CD analyzes changes between HSIs and RGB remote sensing images, effectively alleviating HSI availability limitations and enabling collaborative detection across platforms. It should be emphasized that, in multimodal CD, differences in imaging mechanisms, observation perspectives, and imaging time render critical steps, such as precise image registration and mitigation of modality discrepancies, particularly significant.

To ensure the accuracy of the CD, it is crucial to obtain precise pixel-wise registration between HSI and RGB. Researchers generally consider registration independently and before CD. Image registration methods can be divided into two categories: intensity-based methods and feature-based methods. Intensity-based methods obtain the registration matrix according to the similarity metrics of image intensity values, such as cross-correlation [11] and mutual information methods [12]. Intensity values are easily acquired through such approaches; however, these methods are highly sensitive to changes in illumination and noise. In contrast, feature-based methods have been widely used to solve image registration problems according to the similarity of features, such as scale-invariant feature transform [13] and deep features [14]. Because they are influenced by significant modality gaps and the

* Corresponding author (email: jhqu@xidian.edu.cn)

© Science China Press 2026

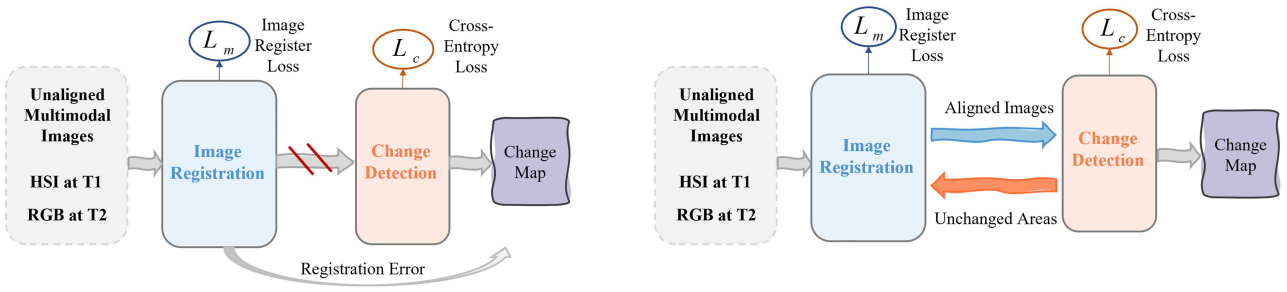

Figure 1 (Color online) Existing multimodal change detection methods typically handle image registration and change detection separately, leading to accumulated registration errors. This study investigates the intrinsic link between registration and detection and proposes a mutually reinforcing registration-detection framework.

presence of change areas, these methods struggle to achieve the precise alignment required for pixel-wise HSI-RGB multimodal CD.

Another key step involves the alleviation of modality differences. In recent years, the quantity of multimodal remote sensing images has increased significantly [15–17]. Researchers address CD by integrating data from diverse satellite platforms [18]. Among these methods, the majority focus on RGB and synthetic aperture radar (SAR) images, assuming that the multimodal images have been accurately registered [19]. Traditional multimodal CD methods distinguish changed areas utilizing either similarity metrics [20] or image regression techniques [21]. These methods define modality-invariant metrics to solve the modality difference problem [22]. However, in scenarios characterized by complexity and significant noise, these methods struggle to accurately compute differences between multimodal images using handcrafted features. Deep learning technology has a powerful feature extraction capability [23–27]. It has been successfully applied to multimodal CD in recent years. Deep learning-based methods can be divided into feature learning-based methods and modality translation-based methods. Feature learning-based methods construct a high-dimensional common feature space, wherein the features of multimodal images can be directly compared. Jiang et al. [28] introduced a commonality autoencoder to identify shared features that effectively differentiate changes between RGB and SAR images. The development of image generation technology has enabled the emergence of modality translation-based methods. Indeed, such methods mitigate the modality gap by translating one image into another modality and detecting changes based on these translated images. Luppino et al. [29] employed cycle-consistent generative adversarial networks (CycleGANs) for image translation. DTCDN [30] used a non-independent component encoding generative adversarial network (NICE-GAN) to model the mapping relationship between multimodal images. Liu et al. [31] proposed a multitask change detection network for CD between RGB and SAR images, integrating the CD network into a generative adversarial network (GAN) to effectively improve robustness. However, the unreliable image translation results obtained, attributed to the instability of GANs, may limit modality translation-based methods in obtaining optimal performance.

Existing methods carried out these two steps separately; however, this can lead to error accumulation and limit detection performance, as shown in Figure 1. The limitations are summarized as follows.

(1) For the task of registering remote sensing images, we can employ geographic coordinates, such as rational polynomial coefficient (RPC) rectification, to align images into a unified coordinate system. However, this approach struggles to achieve pixel-level accuracy. Furthermore, due to the changed regions and modality differences, the accuracy of such registration methods often falls short when applied to bitemporal images with different modalities, let alone achieving pixel-level CD.

(2) HSI-RGB multimodal CD mitigates the limitation imposed by the long revisit period of satellites. However, differences in imaging mechanisms between sensors lead to a modality gap between multimodal images. Consequently, existing HSI-CD methods cannot be directly applied. To address this issue, multimodal CD first needs to implement modal alignment in order to alleviate the modality gap, which also leads to a simplification of the registration process.

It is necessary to address accumulated registration errors in CD task, yet this remains an open challenge. Recently, Zhou et al. [32] integrated registration into a deep learning-based CD framework by sharing the feature extraction model. However, this method focuses only on the implementation of both tasks in a unified network while ignoring the potential for mutual guidance between them. We observed that precise image registration guarantees the consistent spatial location relationships between HSI and RGB images, reducing the pseudo-changes induced by the registration error. Furthermore, mitigating the modality gap and identifying unchanged areas can overcome the challenge of bitemporal image registration. These two tasks are highly related. Can image registration and CD provide mutual reinforcement to simultaneously optimize the parameters for both processes?

In this study, we propose a Mamba-based mutual reinforce diffusion network (M^2DNet) for unregistered HSI-RGB multimodal CD, which jointly optimizes image registration and CD. Specifically, a spatial-modal alignment diffusion network (SMali-Diff) is proposed to address the modality gap and spatial misalignment for the accurate extraction of changed information. This network implements a super-resolution (SR) diffusion process that stably generates bitemporal high-resolution hyperspectral images (HrHSIs) by leveraging the complementary advantages of multimodal data. We then construct a spectral-spatial-temporal Mamba CD network (SST-CDNet) to extract change information between the bitemporal HrHSIs and provide unchanged areas for registration, introducing Mamba with global receptive and linear complexity to jointly obtain spatial-temporal and spectral information while enhancing the performance of the network. The main contributions of our work are summarized as follows.

(1) We propose M^2DNet for unregistered HSI-RGB multimodal CD, which achieves mutual reinforcement between image registration and CD within a unified framework to alleviate spatial misalignment and enhance detection accuracy.

(2) We design SMali-Diff, in which the modality-guided UMamba block is employed to generate a pair of HrHSIs to alleviate modality differences, and the change-aware registration block utilizes feedback from the predicted change map to correct misalignment between the HrHSIs.

(3) For CD, we design SST-CDNet, which integrates the advantages of Mamba in capturing long-range dependencies to model the spatial-temporal and spectral relationships between the registered HrHSIs. This enhances the representational capacity of the network and ensures reliable CD results.

(4) We evaluated the proposed method on three publicly available datasets. Both qualitative and quantitative results have shown that the proposed method outperforms existing state-of-the-art approaches.

The remainder of this paper is organized as follows: Section 2 reviews related work; Section 3 describes the proposed method for multimodal CD; Section 4 analyzes the experimental results; and Section 5 concludes the paper.

2 Related work

2.1 Hyperspectral image change detection

Quantity of HSI CD methods have been proposed in the past few decades, which can be divided into the following four categories: (1) image algebra-based methods; (2) classification-based methods; (3) transformation-based methods; (4) deep learning (DL)-based methods. The image algebra-based methods are relatively simple and straightforward, which include image difference, image ratio, and change vector analysis (CVA) [33]. These methods process the difference information by pixel-wise comparison, are noise-sensitive and have poor performance for bitemporal images with complex spatial texture. Transformation-based methods convert HSIs into a new feature space to reduce the disturbance of noise and highlight the changed information. The representative methods include principal component analysis (PCA) [34], iteratively reweighted MAD [35], and slow feature analysis (SFA) [36]. Transformation-based methods have been shown to improve computing efficiency and reduce noise, but they inevitably disrupt the intrinsic bitemporal relationship during transformation. Classification-based methods consider the classification map images as detection units, including post-classification comparison and compound classification [37]. These methods can provide "from-to" change information but rely heavily on classifier performance, where errors can accumulate and degrade detection results.

Although traditional methods are widely used for CD, they rely on handcrafted features, which limit their ability to represent complex, high-dimensional image changes. Recently, deep learning-based methods have shown strong potential due to their powerful capacity for automatic feature learning [38,39]. Huang et al. [40] designed a feature filtering fusing network for effective CD, which introduces a feature filter and channel information filter to filter out noise from different temporal domains. Due to the advantages of the graph convolutional network (GCN) in processing irregular data, it has earned increasing attention in HSI CD. Qu et al. [41] first proposed a dual-branch difference amplification GCN for HSI-CD task with limited training samples, which applies GCN to extract abundant change information. A sharpening-aware GCN is proposed to achieve HSI-CD, which introduces Laplacian sharpening-based graph convolution to highlight change information between HSIs with complex structures [42]. Transformers, which can catch long sequence dependencies through self-attention and cross-attention mechanisms, have also been successfully applied in HSI-CD [43, 44]. Additionally, Zhou et al. [45] proposed a Kolmogorov-Arnold network-based method to effectively capture change characteristics. Although these DL-based methods have achieved satisfactory performance, they are difficult to apply directly to HSI-RGB CD due to the modality gap. Moreover, they ignore the cumulative error from image registration, limiting their ability to produce reliable

detection results.

2.2 State space model

State space model (SSM) [46] is designed to map a one-dimensional sequential input $x(i)$ into an output $y(i)$ through a hidden state vector $h(i)$. SSM is expressed according to ordinary differential equation (ODE) as follows:

$$\begin{aligned} h'(i) &= \mathbf{A}h(i) + \mathbf{B}x(i), \\ y(i) &= \mathbf{C}h(i) + \mathbf{D}x(i), \end{aligned} \tag{1}$$

where $(\mathbf{A}, \mathbf{B}, \mathbf{C}, \mathbf{D})$ denote the weights of SSMs, and $h'(i)$ is the derivative of the hidden state vector $h(i)$. When the SSMs are applied in the discrete signals, like natural languages and images, discretization is necessary. To achieve discretization, SSMs leverage the zero-order hold (ZOH) rule, which introduces a timescale parameter $\mathbf{\Delta}$ to transform the continuous parameters $\mathbf{A}$ and $\mathbf{B}$ into the discrete parameters $\overline{\mathbf{A}}$ and $\overline{\mathbf{B}}$, which are expressed as follows:

$$\overline{\mathbf{A}} = \exp(\mathbf{\Delta A}), \overline{\mathbf{B}} = (\mathbf{\Delta A})^{-1}(\exp(\mathbf{\Delta A}) - \mathbf{I}) \cdot \mathbf{\Delta B}, \tag{2}$$

where $\mathbf{\Delta}$ is the model parameter, $\overline{\mathbf{A}}$ and $\overline{\mathbf{B}}$ are the discrete form of $\mathbf{A}$ and $\mathbf{B}$. Recently, to enhance the flexibility of SSMs, Mamba [47] replaces the time-invariant parameters with dynamic weights. According to that, the discrete form of (1) can be obtained as follows:

$$\begin{aligned} h_t &= \overline{\mathbf{A}}h_{t-1} + \overline{\mathbf{B}}x_t, \\ y_t &= \mathbf{C}h_t + \mathbf{D}x_t. \end{aligned} \tag{3}$$

Since the parameters are time-varying, dependent on the input, Eq. (3) cannot be reformulated as a convolution. Therefore, Mamba introduces the selective scan mechanism to improve speed with relatively low memory requirement. Finally, the output can be derived as

$$\begin{aligned} \overline{\mathbf{K}} &= \left(\mathbf{C}\overline{\mathbf{B}}, \mathbf{C}\overline{\mathbf{AB}}, \ldots, \mathbf{C}\overline{\mathbf{A}}^{\mathbf{M-1}}\overline{\mathbf{B}}\right), \\ \mathbf{y} &= x * \overline{\mathbf{K}}, \end{aligned} \tag{4}$$

where $\mathbf{M}$ denotes the size of x, and $\overline{\mathbf{K}}$ is a structured convolutional kernel.

3 Proposed method

3.1 Overview

The purpose of the unregistered HSI-RGB multimodal CD is to detect the change regions between the unregistered bitemporal low-resolution HSI (LrHSI) $\mathbf{X}_1$ and high-resolution RGB (HrRGB) $\mathbf{Y}_2$. The LrHSI $\mathbf{X}_1 \in \mathbb{R}^{h\times w\times L}$ and the HrRGB $\mathbf{Y}_2 \in \mathbb{R}^{H\times W\times 3}$ are acquired at times T_1 and T_2 and cover the same region, where h, w, L and $H, W, 3$ ($h \ll H, w \ll W, L \gg 3$) represent the height, width, and number of spectral bands. The problem we are working on is to obtain a change map $\mathbf{M} \in \mathbb{R}^{H\times W}$ with high resolution, which shows the change information between unregistered $\mathbf{X}_1$ and $\mathbf{Y}_2$. Considering the registration error between HSI and RGB, let the HrRGB $\mathbf{Y}_2$ serve as the fixed image and the LrHSI $\mathbf{X}_1$ as the moving image. The final change map $\mathbf{M}$ needs to be spatially consistent with $\mathbf{Y}_2$. Although it is crucial to consider the registration problem in CD tasks, it remains an open question. Existing methods typically require accurately registered multimodal images, necessitating image registration prior to CD. These methods can be formulated as

$$\mathbf{X}_1^{\circ} = f_{\mathrm{reg}}(\mathbf{X}_1, \mathbf{Y}_2), \tag{5}$$

$$\mathbf{M} = f_{\mathrm{CD}}(E(\mathbf{X}_1^{\circ}), E(\mathbf{Y}_2)), \tag{6}$$

where $f_{\mathrm{reg}}(\cdot)$ represents image registration network, $\mathbf{X}_1^{\circ}$ are the registrated image according to $\mathbf{Y}_2$, and the CD network f_{CD} takes the registrated images $\mathbf{X}_1^{\circ}$ and $\mathbf{Y}_2$ as input and generates the change map $\mathbf{M}$. This separated process often leads to the accumulation of registration errors in the detection phase, limiting detection accuracy.

To this end, we propose an M^{2}DNet for unregistered HSI-RGB multimodal CD, which jointly optimizes image registration and change identification in a unified framework. M^{2}DNet consists of two core modules: an SMali-Diff and an SST-CDNet. SMali-Diff aims to transform unregistered bitemporal multimodal inputs into a pair of aligned HRHSIs by jointly modeling super-resolution and registration within a diffusion framework. Based on these aligned

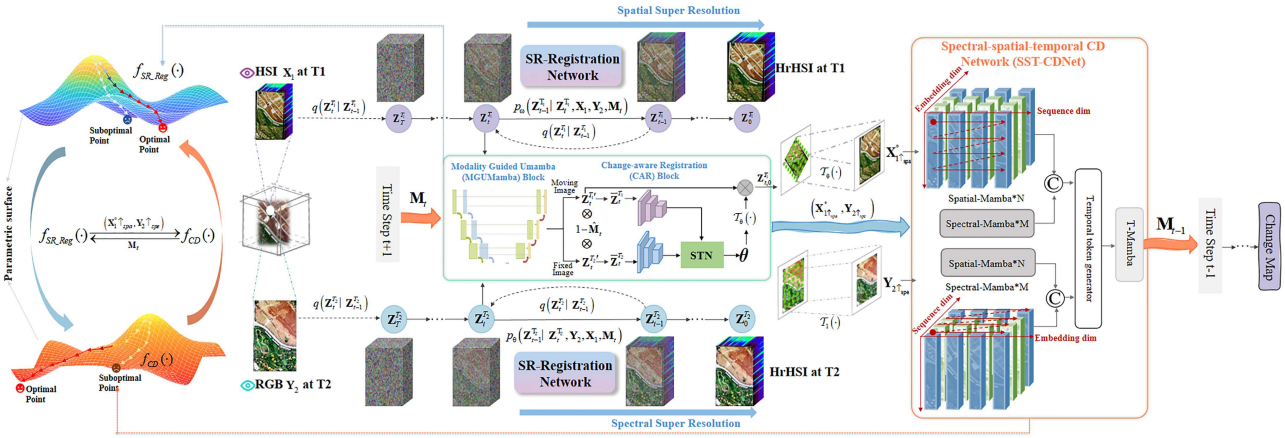

Figure 2 (Color online) Architecture of the proposed M^{2}DNet for unregistered HSI-RGB multimodal change detection. M^{2}DNet consists of an SMali-Diff and an SST-CDNet, which mutually reinforce each other. SMali-Diff generates aligned HrHSI to facilitate change detection, while SST-CDNet produces change map that is fed back to SMali-Diff to further improve registration accuracy.

HRHSIs, SST-CDNet utilizes the Mamba architecture to capture spectral-spatial-temporal features for reliable CD. Subsequently, the change map predicted by SST-CDNet is fed to SMali-Diff via a change-aware matching strategy, where unchanged regions are emphasized to guide the estimation of alignment parameters. Meanwhile, the refined alignment images produced by SMali-Diff suppress pseudo-changes caused by spatial misalignment and modality discrepancies, thereby facilitating more accurate CD in SST-CDNet. Specifically, we propose SMali-Diff based on a conditional diffusion mechanism, in which the SR-Registration network $f_{\mathrm{SR_Reg}}(\cdot)$ implements SR and image registration to alleviate modality differences and correct misalignment. In the reverse denosing process of SMali-Diff, at time step $t-1$, the change-aware matching strategy leverages predicted change map $\mathbf{M}_t$ to improve registration accuracy. Then, a pair of registered bitemporal HrHSIs $(\mathbf{X}_1^{\circ}\uparrow_{\mathrm{spa}}, \mathbf{Y}_2\uparrow_{\mathrm{spe}}) \in \mathbb{R}^{H\times W\times L}$, which is generated by SMali-Diff, is input to SST-CDNet to obtain change map $\mathbf{M}_{t-1}$. SST-CDNet takes the advantages of Mamba to capture the spatial-temporal and spectral information from the registered HrHSIs, thereby producing a precise change map. Formally, the proposed method can be formulated as

$$f_{\mathrm{SR_Reg}}(\cdot) \xrightarrow[\mathbf{M}_t]{(\mathbf{X}_1^{\circ}\uparrow_{\mathrm{spa}}, \mathbf{Y}_2\uparrow_{\mathrm{spe}})} f_{\mathrm{CD}}(\cdot), \tag{7}$$

where $\mathbf{M}_t$ represents the change map at time step t according to SST-CDNet, $\mathbf{X}_1^{\circ}\uparrow_{\mathrm{spa}}$ is the registered HSI $\mathbf{X}_1$ after spatial SR, $\mathbf{Y}_2\uparrow_{\mathrm{spe}}$ is the RGB $\mathbf{Y}_2$ after spectral SR, and $f_{\mathrm{SR_Reg}}(\cdot)$ represents the unified SR and registration network. Finally, the change map $\mathbf{M}$ can be generated by

$$\mathbf{M} = f_{\mathrm{CD}}(\mathbf{X}_1^{\circ}\uparrow_{\mathrm{spa}}, \mathbf{Y}_2\uparrow_{\mathrm{spe}}). \tag{8}$$

3.2 Spatial-modal alignment diffusion network

As shown in Figure 2, the SMali-Diff is composed of the spatial-SR branch and the spectral-SR branch. Specifically, the spatial-SR branch generates the registered HrHSI $\mathbf{X}_1^{\circ}\uparrow_{\mathrm{spa}}$ at time T_1, which is conditioned on HSI $\mathbf{X}_1$. Meanwhile, the spectral-SR branch generates the HrHSI $\mathbf{Y}_2\uparrow_{\mathrm{spe}}$ at time T_2, which is conditioned on the RGB $\mathbf{Y}_2$. Due to the structural similarity of the spatial-SR and spectral-SR branches, except for the image registration network, for simplicity, the details of the spatial-SR branch will be described below.

The spatial-SR branch consists of the forward diffusion process and the reverse denoising process. In the forward diffusion process, noise is gradually added to a clean HrHSI $\mathbf{Z}_0^{\mathrm{T}_1}$ through a Markov chain to generate a standard Gaussian noise $\mathbf{Z}_T \sim \mathcal{N}(0, \mathbf{I})$. Each time step $t \in [0, T]$ of sampling latent variables $\mathbf{Z}_t^{\mathrm{T}_1}$ can be defined as

$$q(\mathbf{Z}_t^{\mathrm{T}_1}|\mathbf{Z}_{t-1}^{\mathrm{T}_1}) = \mathcal{N}(\mathbf{Z}_t^{\mathrm{T}_1}; \sqrt{1-\beta_t}\mathbf{Z}_{t-1}^{\mathrm{T}_1}, \beta_t\mathbf{I}), \tag{9}$$

where $\mathcal{N}(\mathbf{Z}_t^{\mathrm{T}_1}; \sqrt{1-\beta_t}\mathbf{Z}_{t-1}^{\mathrm{T}_1}, \beta_t\mathbf{I})$ represents that $\mathbf{Z}_t$ subjects to the Gaussian distribution with the mean of $\sqrt{1-\beta_t}\mathbf{Z}_{t-1}^{\mathrm{T}_1}$ and the variance of $\beta_t\mathbf{I}$, and $\beta_t \in (0,1)$ is the variance schedule of the noise. Through reparameterization, the distribution of $\mathbf{Z}_t$ can be directly obtained from $\mathbf{Z}_0$

$$q(\mathbf{Z}_t^{\mathrm{T}_1}|\mathbf{Z}_0) = \mathcal{N}(\mathbf{Z}_t^{\mathrm{T}_1}; \sqrt{\alpha_t}\mathbf{Z}_{t-1}^{\mathrm{T}_1}, \alpha_t\mathbf{I}), \tag{10}$$

$$\mathbf{Z}_t^{\mathrm{T}_1} = \sqrt{\alpha_t}\mathbf{Z}_0^{\mathrm{T}_1} + \sqrt{1-\alpha_t}\varepsilon, \tag{11}$$

where $\alpha_t \in \prod_{i=0}^{t}(1-\beta_i)$ and $\varepsilon \sim \mathcal{N}(0, \mathbf{I})$.

For the reverse denoising process, the registered clear HSI $\mathbf{Z}_{t,0}^{\mathrm{T}_1}$ is gradually recovered from the noisy image $\mathbf{Z}_T$ through a neural network with the HSI $\mathbf{X}_1$ as condition, the RGB image $\mathbf{Y}_2$ as spatial guidance, and the previous change map $\mathbf{M}_t$ as the mask. For each time step t, the prediction $\mathbf{Z}_{t,0}^{\mathrm{T}_1}$ can be defined as follows:

$$\mathbf{Z}_{t,0}^{\mathrm{T}_1} = f_{\mathrm{SR_Reg}}(\mathbf{Z}_t^{\mathrm{T}_1}, \mathbf{X}_1, \mathbf{Y}_2, \mathbf{M}_t), \tag{12}$$

where $f_{\mathrm{SR_Reg}}(\mathbf{Z}_t, \mathbf{X}_1, \mathbf{Y}_2, \mathbf{M}_t)$ represents the SR-Registration network. Specifically, the SR-Registration network includes the modality guided UMamba (MGUMamba) block for SR and the change-aware registration block for correct spatial offsets. Then, the input of the next time step can be obtained as follows:

$$\begin{aligned}\mathbf{Z}_{t-1}^{\mathrm{T}_1} =& p_\omega\left(\mathbf{Z}_{t-1}^{\mathrm{T}_1} \mid \mathbf{Z}_t^{\mathrm{T}_1}, \mathbf{X}_1, \mathbf{Y}_2, \mathbf{M}_t\right)\\ =& \frac{\sqrt{\alpha_{t-1}}(1-\alpha_t)}{1-\alpha_t} f_{\mathrm{reg_SR}}(\mathbf{Z}_t^{\mathrm{T}_1}, \mathbf{X}_1, \mathbf{Y}_2, \mathbf{M}_t)\\ &+ \frac{\sqrt{\alpha_t}(1-\alpha_{t-1})}{1-\alpha_t}\mathbf{Z}_t^{\mathrm{T}_1} + \sqrt{\frac{(1-\alpha_{t-1})(1-\alpha_t)}{1-\alpha_t}}\varepsilon.\end{aligned} \tag{13}$$

As the core component of the spatial-SR branch, the SR-Registration network implements image denoising to generate a HrHSI $\mathbf{Z}_{t,0}^{\mathrm{T}_1}$, which is regraded as $\mathbf{X}_1^{\circ}\uparrow_{\mathrm{spa}} \in \mathbb{R}^{H\times W\times L}$. This output preserves semantic consistency with the condition $\mathbf{X}_1$ while achieving spatial alignment with $\mathbf{Y}_2$. In this branch, the function $f_{\mathrm{SR_Reg}}(\mathbf{Z}_t^{\mathrm{T}_1}, \mathbf{X}_1, \mathbf{Y}_2, \mathbf{M}_t)$ incorporates an MGUMamba block to inject spatial information from the RGB image $\mathbf{Y}_2$ and change-aware registration (CAR) block to achieve registration. At time step $t-1$, the spatial-SR branch takes as input the noisy image $\mathbf{Z}_t^{\mathrm{T}_1}$, the spatial guidance image $\mathbf{Y}_2$, and the change map $\mathbf{M}t$ predicted by SST-CDNet at the previous time step.

3.2.1 *Modality guided UMamba block*

Given the noisy image $\mathbf{Z}_t^{\mathrm{T}_1}$ and the RGB image $\mathbf{Y}_2$, the aim of modal alignment is to recover the clean HrHSI $\mathbf{Z}_t^{\mathrm{T}_1'}$ with realistic spatial information, conditioned on the HSI $\mathbf{X}_1$ and guided by the spatial information from the RGB image $\mathbf{Y}_2$. Traditional diffusion models commonly adopt a U-Net backbone, where conditional information is typically injected through simple channel concatenation. Such a design lacks explicit mechanisms for dynamic interactions between the noisy input and the conditional signals, thereby limiting the effective utilization of conditional information. Although existing methods have incorporated conditional information to guide the denoising process through a cross-attention mechanism, these methods are limited by the high complexity of transformers. Recently, Mamba, as an SSM, has shown promising performance across multiple modalities, including language, audio, and images. Mamba demonstrates strong performance akin to Transformer in long sequence modeling with linear complexity. We propose the MGUMamba block, which enhances the U-Net backbone with Mamba to enable more effective conditional injection and generate images with realistic spatial details. Specifically, the MGUMamba block achieves the spatial guidance from $\mathbf{Y}_2$ via cross-modality Mamba, which explores the correlation between different modalities. Furthermore, given the complexity in scale of remote sensing objects, the U-Net architecture is employed to effectively extract hierarchical features. As shown in Figure 3(b), the MGUMamba block consists of an encoder, a bottleneck stage, and a decoder, with skip connections. Its hierarchical encoder-decoder with skip connections enables effective multi-scale feature interaction, retaining the hierarchical spectral semantics of HSI while incorporating the complementary multi-scale spatial structures from RGB, thereby avoiding semantic drift and spatial detail loss. The input $(\mathbf{Z}_t^{\mathrm{T}_1}, \mathbf{Y}_2)$ is encoded by the encoder to obtain the feature $(\mathbf{F}^l_{\mathbf{Z}_t^{\mathrm{T}_1}}, \mathbf{F}^l_{\mathbf{Y}_2}), l = [0, \ldots, N]$, where N denotes the number of layers. During the decoding stage, $\mathbf{F}^l_{\mathbf{Z}_t^{\mathrm{T}_1}}$ is first fed into a residual block to obtain an updated feature representation $\mathbf{F}^{l'}_{\mathbf{Z}_t^{\mathrm{T}_1}}$. Subsequently, the cross-modality Mamba block is introduced to learn accurate cross-modal guidance from $\mathbf{F}^l_{\mathbf{Y}_2}$. Algorithm 1 illustrates the inference procedure of the cross-modality Mamba.

The architecture of the cross-modality Mamba is illustrated in Figure 3(c), the l-layer features $\mathbf{F}^{l'}_{\mathbf{Z}_t^{\mathrm{T}_1}}$ and the auxiliary feature $\mathbf{F}^l_{\mathbf{Y}_2}$ are first projected into token sequences $\mathbf{T}^l_{\mathbf{Z}_t^{\mathrm{T}_1}}$ and $\mathbf{T}^l_{\mathbf{Y}_2}$ through the linear embedding layer. Specifically, the linear embedding layer, structurally instantiated as a fully connected layer, applies a learnable linear

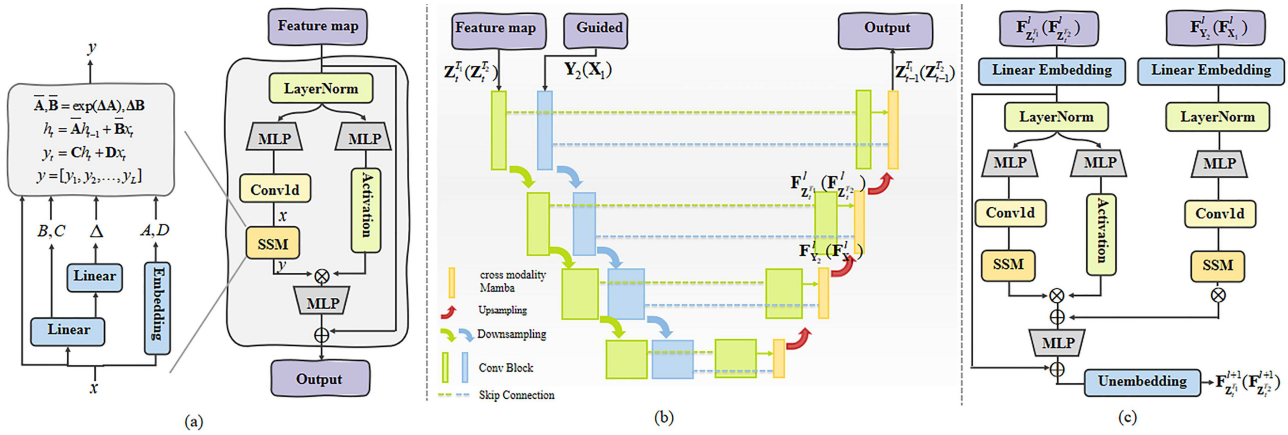

Figure 3 (Color online) Architecture of the proposed MGUMamba block. (a) The architecture of a single Mamba; (b) MGUMamba block utilizes an encoder-decoder architecture; (c) the architecture of the cross-modality Mamba.

Algorithm 1 Cross-modality Mamba block.

Require: Token $\mathbf{T}^{l}_{\mathbf{Z}_t^{\mathrm{T}_1}} : (B, M, D)$, $\mathbf{T}^{l}_{\mathbf{Y}_2} : (B, M, D)$.

Ensure: Token sequence $\mathbf{T}^{l+1}_{\mathbf{Z}_t^{\mathrm{T}_1}} : (B, M, D)$.

for $o \in \{\mathbf{Z}_t^{\mathrm{T}_1}, \mathbf{Y}_2\}$ **do**
  /* Normalize the input sequence $\mathbf{T}^{l}_{o}$ */
  $\mathbf{T}^{l'}_{o} : (B, M, D) \leftarrow \mathrm{Norm}(\mathbf{T}^{l}_{o})$;
  $x_o : (B, M, E) \leftarrow \mathrm{Linear}_o(\mathbf{T}^{l'}_{o})$;
  /* Process with sequence */
  $x'_o : (B, M, E) \leftarrow \mathrm{SiLU}(\mathrm{Conv1d}_o(x_o))$;
  $A_o : (B, M, E, N),\ B_o : (B, M, E, N),\ C_o : (B, M, N) \leftarrow \mathrm{ParametersFunction}_o(x'_o)$;
  $y_o : (B, M, E) \leftarrow \mathrm{SSM}(A_o, B_o, C_o)(x'_o)$;
end for
$z : (B, M, E) \leftarrow \mathrm{Linear}^{z}(\mathbf{T}^{l}_{\mathbf{Z}_t^{\mathrm{T}_1}})$;
/* Get gated y_o */
$y'_{\mathbf{Z}_t^{\mathrm{T}_1}} : (B, M, E) \leftarrow y_{\mathbf{Z}_t} \odot \mathrm{SiLU}(z)$;
$y'_{\mathbf{Y}_2} : (B, M, E) \leftarrow y_{\mathbf{Y}_2} \odot \mathrm{SiLU}(z)$;
/* Residual connection */
$\mathbf{F}^{l}_{\mathbf{Z}_t^{\mathrm{T}_1}} : (B, D, H, W) \leftarrow \mathrm{Reshape}(\mathrm{Linear}^{T}(y'_{\mathbf{Z}_t^{\mathrm{T}_1}} + y'_{\mathbf{Y}_2}) + \mathbf{T}^{l}_{\mathbf{Z}_t^{\mathrm{T}_1}})$;
/* Depth-wise convolution */
$\mathbf{T}^{l+1}_{\mathbf{Z}_t} : (B, M, D) \leftarrow \mathrm{Flatten}(\mathrm{DWConv}(\mathbf{F}^{l}_{\mathbf{Z}_t^{\mathrm{T}_1}}) + \mathbf{F}^{l}_{\mathbf{Z}_t^{\mathrm{T}_1}})$;
return $\mathbf{T}^{l+1}_{\mathbf{Z}_t^{\mathrm{T}_1}}$.

transformation to project the input spatial-spectral feature maps into a token sequence with a designated hidden dimension. By converting the input features into this embedding space, the layer unifies the feature dimensions of the multi-modal data, thereby providing a requisite structural basis for the subsequent layer normalization and cross-modal sequence modeling processes. Subsequently, the token sequences are normalized using layernorm blocks. Then the input sequences are transformed via linear layers, followed by the SiLU activation function. After the parameters A_o, B_o, and C_o are generated by parameters function, we apply SSM:

$$y_o = \mathrm{SSM}(A_o, B_o, C_o)(x'_o), \quad o = [\mathbf{Z}_t^{\mathrm{T}_1}, \mathbf{Y}_2]. \tag{14}$$

Then, the gating operation is applied to achieve accurate cross-modal guidance, where the gating parameter z is obtained from $\mathbf{T}^{l}_{\mathbf{Z}_t^{\mathrm{T}_1}}$

$$z = \mathrm{Linear}(\mathbf{T}^{l}_{\mathbf{Z}_t^{\mathrm{T}_1}}), \tag{15}$$

where z is a gating parameter that contains the correlation of input features. Then, the gating parameter z are used to warp $y_{\mathbf{Z}_t^{\mathrm{T}_1}}$ and $y_{\mathbf{Y}_2}$, after which a residual connection is incorporated

$$\mathbf{F}^{l+1}_{\mathbf{Z}_t} = \mathrm{Reshape}(\mathrm{Linear}(y_{\mathbf{Z}_t^{\mathrm{T}_1}} \otimes \sigma(z) + y_{\mathbf{Y}_2} \otimes \sigma(z)) + \mathbf{T}^{l}_{\mathbf{Z}_t^{\mathrm{T}_1}}), \tag{16}$$

where $\mathbf{F}^{l+1}_{\mathbf{Z}_t^{\mathrm{T}_1}}$ is the $(l+1)$-layer features in UMamba architecture, the $\mathrm{Reshape}(\cdot)$ operation transforms the token

sequence $\mathbf{T}^{l+1}_{\mathbf{Z}_t^{\mathrm{T}_1}}$ into a two-dimensional feature map, $\sigma(\cdot)$ denotes a SiLU activation function, and $\mathbf{T}^{l}_{\mathbf{Z}_t^{\mathrm{T}_1}}$ represents the token sequences from the l-layer input features. By learning the gating parameter exclusively from HSI spectral features, semantic information is explicitly dominated by HSI, while RGB features are controllably weighted to supplement spatial structures. This design effectively preserves spectral semantics while enabling reliable spatial information injection. Finally, the super-resolution HSI $\mathbf{Z}_t^{\mathrm{T}_1'}$, generated by the MGUMamba block at the N-layer, preserves consistent semantics with the condition image $\mathbf{X}_1$ following the super-resolution process. In the spectral-SR branch, the MGUMamba block also generates a HrHSI $\mathbf{Z}_t^{\mathrm{T}_2'}$, which maintains consistent semantics with the condition image $\mathbf{Y}_2$.

3.2.2 *Change-aware registration block*

Given a pair of clean HrHSIs $(\mathbf{Z}_t^{\mathrm{T}_1'}, \mathbf{Z}_t^{\mathrm{T}_2'})$, i.e., $(\mathbf{X}_1\uparrow_{\mathrm{spa}}, \mathbf{Y}_2\uparrow_{\mathrm{spe}})$, generated by the MGUMamba block, the CAR block aims to learn a set of affine parameters to correct misalignment between $\mathbf{Z}_t^{\mathrm{T}_1'}$ and $\mathbf{Z}_t^{\mathrm{T}_2'}$, and transform the HrHSI $\mathbf{Z}_t^{\mathrm{T}_1'}$ into a registered HrHSI $\mathbf{Z}_{t,0}^{\mathrm{T}_1}$, i.e., $\mathbf{X}_1^{\circ}\uparrow_{\mathrm{spa}}$. Here, $\mathbf{Z}_t^{\mathrm{T}_1'}$ is the moving image and $\mathbf{Z}_t^{\mathrm{T}_2'}$ is the fixed image. It is important to note that after processing by the MGUMamba block, the multimodal registration task has been simplified into a uni-modal registration task. Further, to obtain a more accurate registration result, i.e., a set of affine parameters, we need to reduce the influence of changed areas. Therefore, we propose the change-aware matching strategy, which utilizes the change map $\mathbf{M}_t$ to highlight the unchanged regions that contribute much to the affine transformation. In the implementation of the change-aware matching strategy, the change map $\mathbf{M}_t$ are used to weight the image $\mathbf{Z}_t^{\mathrm{T}_1'}$ and $\mathbf{Z}_t^{\mathrm{T}_2'}$, obtaining two weighted images $\overline{\mathbf{Z}}_t^{\mathrm{T}_1}$ and $\overline{\mathbf{Z}}_t^{\mathrm{T}_2}$ as

$$\overline{\mathbf{Z}}_t^{*} = (1 - \mathbf{M}_t) \otimes \mathbf{Z}_t^{*'}, * \in \mathrm{T}_1, \mathrm{T}_2, \tag{17}$$

where $\mathbf{M}_t \in (0, 1)$, 0 represents the unchanged pixels, and 1 represents the changed pixels. In this way, the CAR block will concentrate on the unchanged areas to achieve spatial alignment. Then $\overline{\mathbf{Z}}_t^{\mathrm{T}_1}$ and $\overline{\mathbf{Z}}_t^{\mathrm{T}_2}$ are fed to the affine transformation network to generate affine parameters. Specifically, each input image is first processed by convolutional layers to project the original high-dimensional hyperspectral representations into low-dimensional, geometry-sensitive feature maps. The extracted features are then passed to a classic spatial transformer network (STN) [48], where a lightweight multilayer perceptron (MLP) is employed to regress the parameters of an affine transformation θ of size 2×3. After that, the affine parameters are used to warp $\mathbf{Z}_t^{\mathrm{T}_1'}$ to obtain the aligned image $\mathbf{Z}_{t,0}^{\mathrm{T}_1}$

$$\mathbf{X}_1^{\circ}\uparrow_{\mathrm{spa}} = \mathbf{Z}_{t,0}^{\mathrm{T}_1} = \mathcal{T}_{\theta}(\mathbf{Z}_t^{\mathrm{T}_1'}), \tag{18}$$

where $\mathcal{T}_{\theta}(\cdot)$ represent the affine transformation operation.

The SMali-Diff network achieves the modal alignment and spatial alignment through a step-by-step denoising process, thereby obtaining a pair of registered HrHSIs for CD.

3.3 Spectral-spatial-temporal Mamba change detection network

Given the diverse land cover distributions and extensive spectral bandwidth typical of HSI, capturing global context is indispensable for accurately distinguishing genuine changes between bitemporal images. Traditional CNN-based backbones are limited by the local receptive fields of convolution operations. Although Transformer-based backbones can capture global dependencies via self-attention, they suffer from quadratic computational complexity ($\mathcal{O}\left(N^2\right)$) regarding the input size. Mamba has the ability to model global context with linear complexity, effectively capturing global semantic information. In this paper, SST-CDNet is designed to capture the difference information between the pair of registered HrHSIs, and provide the change map for the subsequent CAR block. SST-CDNet takes advantage of Mamba to model the spatial, spectral, and temporal relationships within bitemporal HrHSIs, thereby extracting discriminative features for accurate changed and unchanged region discrimination. Further, a more accurate change map generated from this process can streamline the subsequent image registration process by providing unchanged areas.

Take the t-th step as an example, the SMali-Diff network will output the potential features $\mathbf{Z}_{t,0}^{\mathrm{T}_1}$ and $\mathbf{Z}_{t,0}^{\mathrm{T}_2}$ in the spatial-SR branch and the spectral-SR branch sparately. Then $(\mathbf{Z}_{t,0}^{\mathrm{T}_1}, \mathbf{Z}_{t,0}^{\mathrm{T}_2})$ will be fed into the SST-CDNet to extract changed information and obtain a change map $\mathbf{M}_t$. The unchanged regions of $\mathbf{M}_t$ will help to overcome the challenge of bitemporal image registration in the $(t-1)$-th step. To reduce the complexity of the network, we exclusively perform CD at designated time steps.

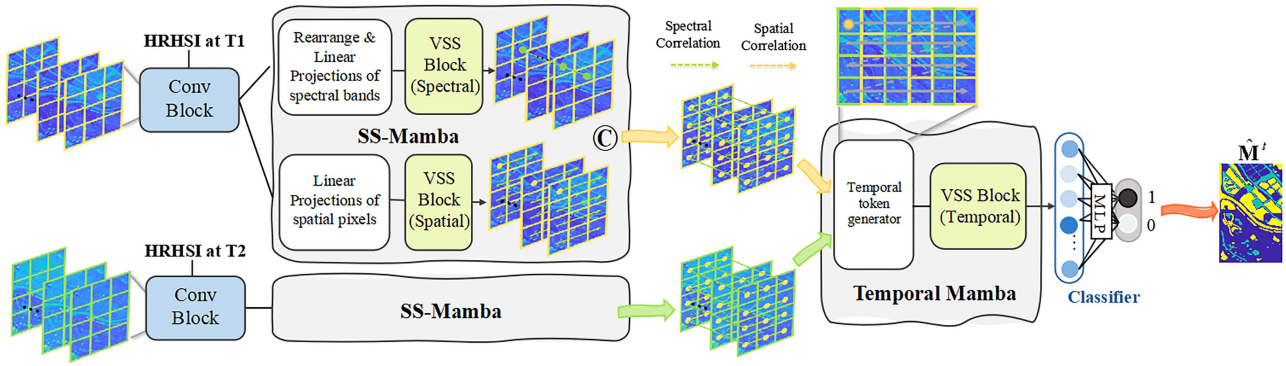

Figure 4 (Color online) Architecture of SST-CDNet. The SS-Mamba block extracts joint spatial-spectral features, while the temporal Mamba block models temporal dependencies and captures discriminative difference features for change detection.

As shown in Figure 4, the proposed SST-CDNet consists of spectral-spatial Mamba (SS-Mamba) and temporal Mamba blocks with temporal tokens generation to highlight the changed information and compress the unchanged information. First, the HrHSIs are divided into many patch pairs and fed into conv block to obtained $(\mathbf{F}_{t,0}^{\mathrm{T}_1}, \mathbf{F}_{t,0}^{\mathrm{T}_2})$. In SS-Mamba, since Mamba is limited to encoding sequential data, it is essential to initially transform 3D HSI data into tokens via scanning. For the generation of spatial tokens, the patch $\mathbf{F}_{t,0}^{\mathrm{T}_1} \in \mathbb{R}^{H\times W\times D}$ is reshaped into a spatial token of shape $\mathbb{R}^{HW\times D}$, then a lightweight linear projection is used to the token for reducing dimensions:

$$\mathbf{T}_{\mathrm{spa}}^{\mathrm{T}_1} = \mathrm{MLP}\left(\mathrm{Reshape}(\mathbf{F}_{t,0}^{\mathrm{T}_1})\right), \quad \mathbf{T}_{\mathrm{spa}}^{\mathrm{T}_1} \in \mathbb{R}^{HW\times D}, \tag{19}$$

where $\mathbf{T}_{\mathrm{spa}}^{\mathrm{T}_1}$ represents the spatial token extracted from $\mathbf{F}_{t,0}^{\mathrm{T}_1}$. The spatial token $\mathbf{T}_{\mathrm{spa}}^{\mathrm{T}_2}$ can be derived from $\mathbf{F}_{t,0}^{\mathrm{T}_2}$ in a similar approach:

$$\mathbf{T}_{\mathrm{spa}}^{\mathrm{T}_2} = \mathrm{MLP}\left(\mathrm{Reshape}(\mathbf{F}_{t,0}^{\mathrm{T}_2})\right), \quad \mathbf{T}_{\mathrm{spa}}^{\mathrm{T}_2} \in \mathbb{R}^{HW\times D}, \tag{20}$$

where $\mathbf{T}_{\mathrm{spa}}^{\mathrm{T}_2}$ represents the spatial token extracted from $\mathbf{F}_{t,0}^{\mathrm{T}_2}$. For the generation of spectral tokens, $\mathbf{T}_{\mathrm{spa}}^{\mathrm{T}_1} \in \mathbb{R}^{HW\times D}$ is reshaped into the spectral token $\mathbf{T}_{\mathrm{spe}}^{\mathrm{T}_1} \in \mathbb{R}^{D\times HW}$. In a similar way, the spectral token $\mathbf{T}_{\mathrm{spe}}^{\mathrm{T}_2} \in \mathbb{R}^{D\times HW}$ can also be obtained. $\mathbf{T}_{\mathrm{spa}}^{\mathrm{T}_1}$, $\mathbf{T}_{\mathrm{spe}}^{\mathrm{T}_1}$ are fed into a series of Mamba blocks to encode the information of the HrHSI at time T_1. The structure of the Mamba block is illustrated in Figure 3(a). The output $\mathbf{F}^{\mathrm{T}_1}$ and $\mathbf{F}^{\mathrm{T}_2}$ from SS-Mamba can be formulated as

$$\mathbf{F}^{*} = \mathrm{Reshape}(\mathrm{Concat}(\mathrm{Mamba}(\mathbf{T}_{\mathrm{spa}}^{*})^{\mathrm{T}}, \mathrm{Mamba}(\mathbf{T}_{\mathrm{spe}}^{*}))), * \in \mathrm{T}_1, \mathrm{T}_2, \tag{21}$$

where $\mathbf{F}^{\mathrm{T}_1}$ and $\mathbf{F}^{\mathrm{T}_2}$ represent the output features from SS-Mamba, respectively, the $\mathrm{Reshape}(\cdot)$ operation unfolds a token from $\mathbb{R}^{D\times HW}$ to a tensor in $\mathbb{R}^{H\times W\times D}$, $\mathrm{Mamba}(\cdot)$ represents Mamba blocks for enhanced long-range dependency modeling.

To extract the difference information between $\mathbf{F}^{\mathrm{T}_1}$ and $\mathbf{F}^{\mathrm{T}_2}$, we consider the features from bitemporal HrHSIs as time series data and design temporal token generation to model these time series for CD. The temporal token generation is proposed to model the temporal relationships between $(\mathbf{F}^{\mathrm{T}_1}, \mathbf{F}^{\mathrm{T}_2})$. First, $(\mathbf{F}^{\mathrm{T}_1}, \mathbf{F}^{\mathrm{T}_2})$ are fused into a temporal feature $\mathbf{F}^{\mathrm{tem}} \in \mathbb{R}^{H\times 2W\times D}$ through an alternating filling strategy. The features of $\mathbf{F}^{\mathrm{T}_1}$ are mapped to the odd-indexed columns (i.e., the $2w$-th columns), while the features of $\mathbf{F}^{\mathrm{T}_2}$ are mapped to the even-indexed columns (i.e., the $(2w+1)$-th columns). Formally, the temporal feature is defined as

$$\mathbf{F}_{B,C,H,2W}^{\mathrm{tem}} = \mathbf{F}^{\mathrm{T}_1}, \quad \mathbf{F}_{B,C,H,2W+1}^{\mathrm{tem}} = \mathbf{F}^{\mathrm{T}_2}, \tag{22}$$

where $\mathbf{F}_{B,C,H,2W}^{\mathrm{tem}}$ and $\mathbf{F}_{B,C,H,2W+1}^{\mathrm{tem}}$ denote the odd-indexed columns and the even-indexed columns of $\mathbf{F}^{\mathrm{tem}}$, respectively. Next, the temporal feature $\mathbf{F}^{\mathrm{tem}}$ is flattened into a sequential token $\mathbf{T}^{\mathrm{tem}}$, which are then processed through temporal Mamba blocks. Following this, the output $\mathbf{T}^{\mathrm{tem}'}$ generated by the temporal Mamba blocks is reshaped into $\mathbf{F}^{\mathrm{tem}'}$. The difference feature $\mathbf{P}$ is computed as

$$\mathbf{P} = \mathbf{F}^{\mathrm{tem}'}{}_{B,C,H,2W} - \mathbf{F}^{\mathrm{tem}'}{}_{B,C,H,2W+1}, \tag{23}$$

where $\mathbf{P}$ denotes the difference feature. Subsequently, a lightweight linear projector is employed as a classifier on $\mathbf{P}$. The final change map at time step t is then obtained as

$$\mathbf{M}_t = \arg\max(\mathrm{Classifier}(\mathbf{P})), \tag{24}$$

where $\mathbf{M}_t$ denotes the change map at time step t. The change map at the time step 0 is regarded as the final result $\mathbf{M}$.

Table 1 Description of the three multimodal datasets.

Dataset	The size of RGB	The size of HSI	Location	Date (T_1, T_2)
Bay Area	$600 \times 500 \times 3$	$200 \times 167 \times 224$	Patterson, California, USA	2013, 2015
Hermiston	$307 \times 241 \times 3$	$102 \times 80 \times 154$	Hermiston, Umatilla County, Oregon, OR, USA	May 2004, May 2007
XDU-Liyukou	$750 \times 375 \times 3$	$250 \times 125 \times 194$	Liyukou, Huyi, Xi'an, Shaanxi, China	November 2022, April 2023

3.4 Training objective

Given a training set $(\mathbf{X}_1, \mathbf{Y}_2, \mathbf{Z}^{\mathrm{T}_1}, \mathbf{Z}^{\mathrm{T}_2}, \mathbf{M}_{gt})$, where $\mathbf{X}_1$, $\mathbf{Y}_2$, $\mathbf{M}$ denote the LrHSI acquired at time T1, the HrRGB acquired at T1 and the ground truth, and $(\mathbf{Z}^{\mathrm{T}_1}, \mathbf{Z}^{\mathrm{T}_2})$ are the reference HrHSIs corresponding to $(\mathbf{X}_1, \mathbf{Y}_2)$. Through the MGUMamba block in the SMali-Diff network, the HrHSIs $(\mathbf{Z}_t^{\mathrm{T}_1'}, \mathbf{Z}_t^{\mathrm{T}_2'})$ are generated, i.e., $(\mathbf{X}_1{\uparrow_{\mathrm{spa}}}, \mathbf{Y}_2{\uparrow_{\mathrm{spe}}})$, and the alignment of multimodal images in terms of modality is achieved. In order to enhance the accuracy of spatial and spectral information during the SR process and enhance the similarity between the input data and the HrHSIs, we choose L_1 loss as the L_{SR} loss function to train the MGUMamba block:

$$L_{\mathrm{SR}} = \left\| \mathbf{Z}^{\mathrm{T}_1} - \mathbf{Z}_t^{\mathrm{T}_1'} \right\|_1 + \left\| \mathbf{Z}^{\mathrm{T}_2} - \mathbf{Z}_t^{\mathrm{T}_2'} \right\|_1, \tag{25}$$

where $\mathbf{Z}_t^{\mathrm{T}_1'}$ and $\mathbf{Z}_t^{\mathrm{T}_2'}$ denote the LrHSI $\mathbf{X}_1$ after spatial SR and the HrRGB $\mathbf{Y}_2$ after spectral SR, $(\mathbf{Z}^{\mathrm{T}_1}, \mathbf{Z}^{\mathrm{T}_2})$ are the reference HrHSIs, and $\|\cdot\|_1$ is the L_1 loss. The registration loss function, designed to constrain the spatial alignment between $\mathbf{Z}_{t,0}^{\mathrm{T}_1} = \mathcal{T}_\theta(\mathbf{Z}_t^{\mathrm{T}_1'})$ and $\mathbf{Z}_{t,0}^{\mathrm{T}_2} = \mathbf{Z}_t^{\mathrm{T}_2'}$, is formally defined as follows:

$$\begin{aligned} L_{\mathrm{Reg}} =& L_{\mathrm{MSE}} \left(T_\theta(\mathbf{Z}_t^{\mathrm{T}_1'}) \otimes (1 - \mathbf{M}_t), \mathbf{Z}_t^{\mathrm{T}_2'} \otimes (1 - \mathbf{M}_t) \right) \\ &+ L_{\mathrm{MSE}} \left(\mathbf{Z}_t^{\mathrm{T}_1'} \otimes (1 - \mathbf{M}_t), {T_\theta}^{-1}(\mathbf{Z}_t^{\mathrm{T}_2'}) \otimes (1 - \mathbf{M}_t) \right), \end{aligned} \tag{26}$$

where L_{MSE} represents the mean squared error loss, which measures the similarity between images, the change map $\mathbf{M}_t$ is initialized as a zero matrix and updated by SST-CDNet at each designated time step t, the first term ensures alignment between the moving image $\mathbf{Z}_t^{\mathrm{T}_1'}$ after registration $\mathbf{Z}_{t,0}^{\mathrm{T}_1}$ and the fixed image $\mathbf{Z}_t^{\mathrm{T}_2'}$, the second term utilizes the registered reversibility, not only maintaining alignment between the registered moving image and the fixed image but also ensuring that the resampling of the fixed image using backward deformation ${\mathcal{T}_\theta}^{-1}(\cdot)$ aligns with the moving image $\mathbf{Z}_t^{\mathrm{T}_1'}$. This dual mechanism enhances the robustness of the registration. Additionally, by incorporating the change map $\mathbf{M}_t$, the CAR block focuses on unchanged areas to achieve spatial alignment.

In addition, we employ the cross-entropy loss to measure the similarity between the prediction $\mathbf{M}$ and the ground-truth of CD $\mathbf{M}_{gt}$:

$$L_{\mathrm{CD}} = -\frac{1}{n} \sum_{i=1}^{n} [m_{gt}^i \log m^i + (1 - m_{gt}^i) \log(1 - m^i)], \tag{27}$$

where m^i is the prediction of i-th sample pair, m_{gt}^i is the corresponding ground-truth, and n is the number of training samples. The L_{SR}, L_{Reg}, and L_{CD} loss are joint optimized to obtain the optimal model.

3.5 Data description

Experiments are conducted on three datasets that capture a range of change conditions. The three multimodal datasets examined are listed in Table 1, which contains synthetic image pairs (e.g., Bay Area and Hermiston) and a real-world image pair (e.g., XDU-Liyukou [49]). Each dataset comprises two images of the same area, which features spatial deformation. To model potential distortions that could arise during the imaging process, an affine transformation is applied to the HSI captured at time T_1. This transformation incorporates a shift of 3 pixels, a shear factor varying between 0.95 and 1.05, and a rotational adjustment within the range of -5 to 5 degrees. We employ six evaluation metrics to assess the performance of the proposed method: overall accuracy (OA), kappa coefficient (KC), F1-score (F1), precision (Pre), recall (Rec) and average accuracy (AA).

3.6 Competing methods and experimental settings

We compare M^{2}DNet with several SOTA multimodal CD methods: GSP [50], INLPG [50], IRG-McS [22], IRMAD [35], HAFF [51], DTCDN [52], SUNet [53] and AFENet [54]. The models were implemented with PyTorch using an NVIDIA RTX 3090 GPU, and the Adam algorithm was applied to optimize the network. We set a batch size

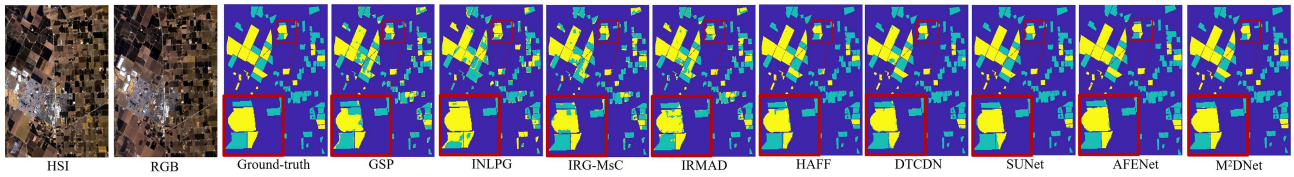

Figure 5 (Color online) Change maps of different methods on the Bay Area dataset.

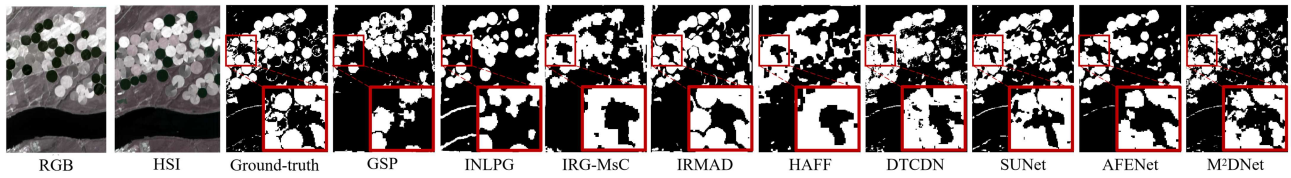

Figure 6 (Color online) Change maps of different methods on the Hermiston dataset.

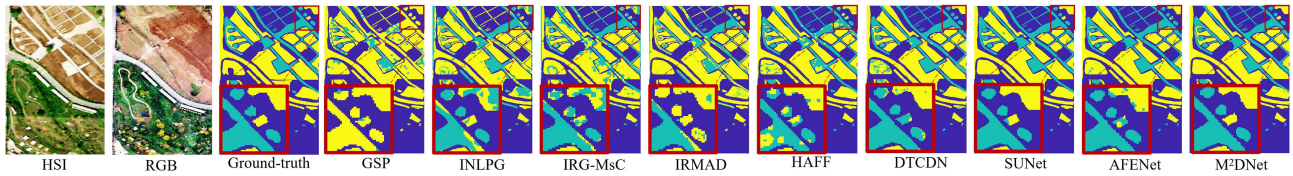

Figure 7 (Color online) Change maps for different methods on the XDU-Liyukou dataset.

of 16, a learning rate of 0.0001, and trained the model for 1000 epochs. SMali-Diff was pre-trained for 60 epochs. The proportion of samples used for training is 10%. In the diffusion model, the time step T was set to 1500, and the hyperparameter sequence $\{\beta_1, \beta_2, \ldots, \beta_n\}$ was defined within the range 0–0.02.

3.7 Experimental results

Qualitative results. The qualitative results of different methods on the three examined datasets are presented in Figures 5–7. For the Bay Area and XDU-Liyukou datasets, yellow, green, and blue denote changed, unchanged, and unlabeled regions, respectively; whereas for the Hermiston dataset, white and black correspond to changed and unchanged regions. In Figure 5, the deep learning-based approaches are shown to exhibit superior performance relative to traditional methods; however, they still fail to capture fine details in small changed areas, as shown in the local magnified views. The proposed method yields results closest to the ground truth.

As shown in Figure 6, the deep learning competing approaches face high false alarm rates and difficulty in detecting small changes due to spatial misalignment in bitemporal images. M^{2}DNet effectively handles spatial misalignment and accurately captures small-scale changes, achieving the best overall performance. As shown in Figure 7, traditional methods contain extensive missed detections in the upper-right region of the XDU-Liyukou dataset. Deep learning methods significantly alleviate the missed detection issue; however, misidentification in some edge regions continues to be visible in the magnified regions. The proposed method outperforms existing state-of-the-art methods in both detection accuracy and generalization ability across different change scenarios.

Quantitative results. As shown in Table 2, M^{2}DNet achieves the best detection performance across the three datasets. Consistent with the qualitative results, these findings indicate that the proposed method effectively suppresses pseudo-changes caused by spatial misalignment and modality discrepancies. Although SUNet attains a higher degree of recall, its low precision leads to significantly reductions in both F1-score and KC, indicating that many unchanged pixels are misclassified as changed due to spatial offsets and modality differences. In contrast, by jointly performing super-resolution and registration through a spatial-modal alignment diffusion network, M^{2}DNet effectively mitigates these issues, resulting in improved precision and F1-score. For the XDU-Liyukou dataset, GSP exhibits lower OA and KC owing to its limited capability in handling complex structural changes. While IRG-McS and IRMAD outperform other traditional methods, their AA values remain below 90%. Overall, deep learning-based methods show clear advantages, and M^{2}DNet consistently achieves the most accurate results across all evaluation metrics.

The proposed method achieves optimal performance in visualization and quantitative results; this can primarily be attributed to the mutual reinforcement between image registration and CD within a unified framework. SMali-Diff reduces modality differences and spatial misalignment via joint super-resolution and change-aware registration

Table 2 Quantitative assessment of different change detection methods on three datasets. The best results are highlighted in bold.

Method	Bay Area						Hermiston						XDU-Liyukou					
	OA%	KC%	F1%	Pre%	Rec%	AA%	OA%	KC%	F1%	Pre%	Rec%	AA%	OA%	KC%	F1%	Pre%	Rec%	AA%
GSP	82.09	64.07	83.03	84.09	82.00	82.00	83.11	47.51	57.86	66.09	51.46	76.41	73.06	26.16	38.50	75.03	25.90	73.92
INLPG	75.07	50.15	75.91	78.50	73.49	75.07	81.83	49.35	61.19	59.01	63.53	74.08	75.22	41.65	59.36	63.70	55.58	71.74
IRG-MsC	83.98	67.83	84.94	85.34	84.54	83.89	86.49	62.25	71.04	68.74	73.50	80.44	79.49	52.52	67.47	69.78	65.32	76.76
IRMAD	82.80	65.56	83.49	85.72	81.36	82.76	87.00	64.02	72.50	69.30	76.01	81.06	81.11	52.50	64.51	83.06	52.74	81.83
HAFF	97.04	94.04	97.30	95.03	99.68	97.32	90.68	75.10	81.23	74.36	89.51	85.56	94.98	88.43	92.11	94.41	89.92	94.82
DTCDN	95.46	90.82	95.90	92.66	99.37	95.94	92.62	79.55	84.37	80.68	88.41	88.61	95.33	89.32	92.78	93.38	92.18	94.81
SUNet	98.59	97.15	98.69	97.53	99.88	98.70	92.90	80.77	85.43	79.44	**92.40**	88.56	97.56	94.55	96.38	93.17	**99.82**	96.54
AFENet	98.81	97.61	98.89	98.75	99.03	98.82	93.45	80.33	84.45	90.86	78.89	92.47	98.24	96.03	97.35	95.60	99.17	97.59
M^{2}DNet	**99.72**	**99.43**	**99.73**	**99.80**	**99.67**	**99.71**	**94.06**	**82.97**	**86.80**	**86.93**	86.68	**91.53**	**99.02**	**97.76**	**98.49**	**98.70**	98.27	**98.93**

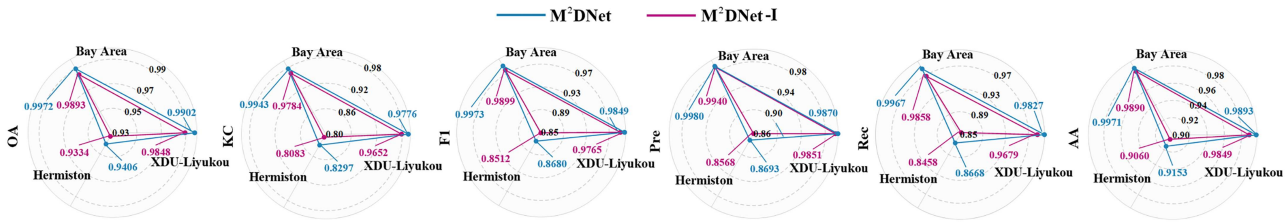

Figure 8 (Color online) Detection performance of M^{2}DNet and its variant removing the CAR block from the SMali-Diff network (M^{2}DNet-I) on three datasets.

in a diffusion framework, providing well-aligned inputs suitable for SST-CDNet. These refined representations enable more accurate modeling of spectral-spatial-temporal correlations, whereas the resulting change maps further guide the registration, forming a closed-loop mechanism that improves the overall performance. Thus, the method effectively mitigates misalignment-induced errors and leverages cross-modal complementary information for more accurate and robust CD.

3.8 Ablation studies

3.8.1 *Effectiveness of CAR*

To validate the effectiveness of the proposed CAR block, we design a variant of M^{2}DNet termed M^{2}DNet-I, which removes the CAR block from the SMali-Diff network and achieves super-resolution only to mitigate modality differences. We compared the performances of M^{2}DNet and M^{2}DNet-I on the three examined datasets, namely the XDU-Liyukou, Bay Area, and Hermiston datasets. As shown in Figure 8, the results obtained clearly show that the variant M^{2}DNet-I yields lower metric values than M^{2}DNet on all three datasets. This confirms that the proposed method effectively mitigates irrelevant unchanged information and enhances the performance of multimodal CD on unaligned images.

3.8.2 *Effectiveness of cross-modality Mamba*

To demonstrate the effectiveness of the proposed MGUMamba block, we design a variant of M^{2}DNet called M^{2}DNet-II, in which the proposed MGUMamba block is replaced by a standard U-Net, completely removing the Mamba components. As shown in Figure 9, M^{2}DNet obtains better performance than its variant M^{2}DNet-II in all quantitative metrics across the three examined datasets, demonstrating that the proposed method with cross-modality Mamba effectively achieves cross-modal guidance and generates realistic HrHSIs, thereby contributing to improved accuracy and robust CD results.

3.8.3 *Comparison with other CD backbones*

To verify the effectiveness of SST-CDNet, we selected representative CNN-based and Transformer-based CD backbones, including ReCNN [55], MSDFNN [56], BIT [57], and MFCEN [58], to substitute SST-CDNet in CD tasks. ReCNN and MSDFNN use CNN as their backbones, whereas BIT and MFCEN are Transformer-based methods. We compared the performance, model size, and computational complexity of different backbones using the XDU-Liyukou dataset, the results of which are shown in Figure 10. The top row shows the number of parameters (in millions) for each model, and the right column compares computational cost in GFLOPs (in millions). It is evident

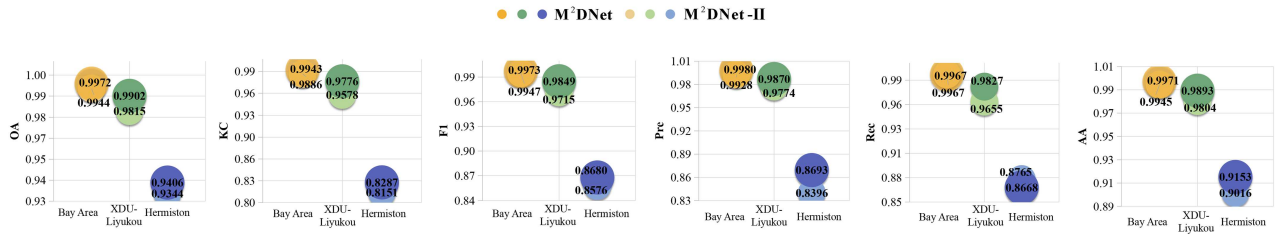

Figure 9 (Color online) Detection performance of M^{2}DNet and its variant omitting the cross-modal guidance provided by cross-modality Mamba (M^{2}DNet-II) on three datasets.

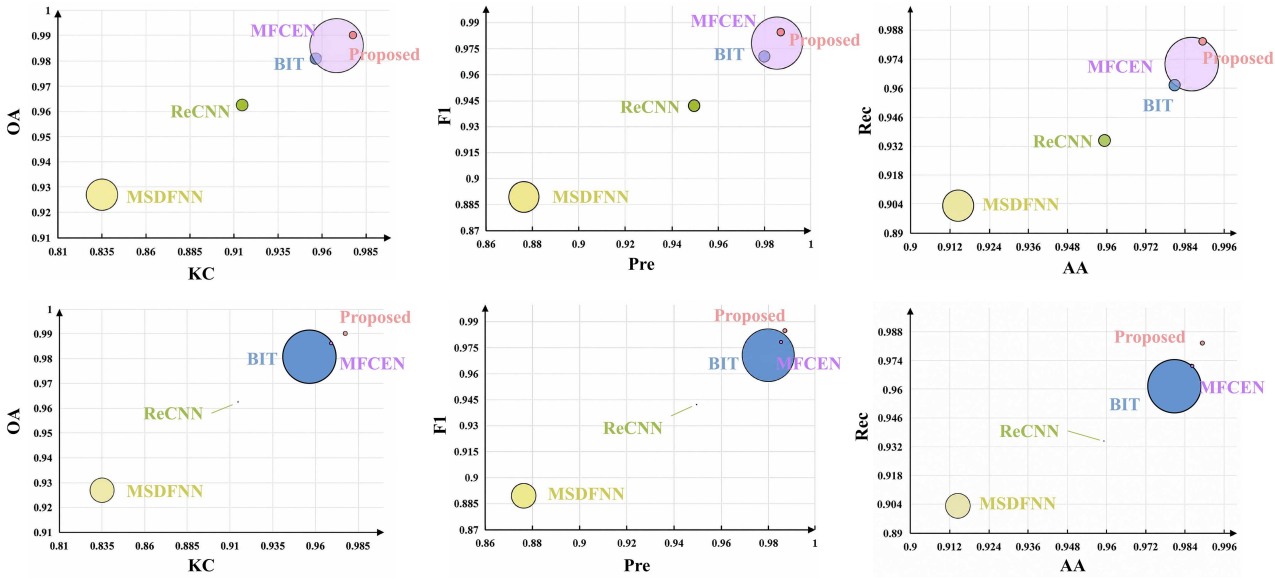

Figure 10 (Color online) Detection performance of M^{2}DNet with different CD backbones on the XDU-Liyukou dataset.

that CNN-based methods, particularly MSDFNN, exhibit lesser performance due to their limitation in capturing only local information, coupled with their inability to model spatial-temporal and spectral relationships between HrHSIs. Transformer-based CD backbones achieve strong performance using self-attention and cross-attention mechanisms; however, such approaches also place high demands on computational resources. SST-CDNet strikes the best balance, achieving state-of-the-art accuracy with model sizes and computational costs that are much lower than their Transformer-based counterparts.

4 Conclusion

M^{2}DNet is proposed for unaligned HSI-RGB multimodal CD, enabling mutual reinforcement between image registration and CD within a conditional diffusion framework. To reduce modality differences and spatial misalignment, the SMali-Diff network performs SR and registration. In particular, the registration block uses a change-aware matching strategy and feedback from change maps to correct misalignment. SST-CDNet processes the aligned homogeneous images to model spatial-temporal and spectral correlations, leading to improved detection accuracy and identifying unchanged areas for registration. Extensive experiments demonstrate that our method achieves superior performance for unaligned HSI-RGB multimodal CD relative to other approaches. This study focuses primarily on binary CD to identify changed regions; future work should further explore semantic CD capable of recognizing both changed areas and their corresponding change categories, thereby extending the applicability of CD models.

Acknowledgements This work was supported in part by National Natural Science Foundation of China (Grant No. 62471359), Young Talent Fund of Association for Science and Technology in Shaanxi (Grant Nos. 20250133, 20230117), Key Research and Development Program of Shaanxi (Grant No. 2025SF-YBXM-513), and Fundamental Research Funds for the Central Universities (Grant No. QTZX25084).

References

1 Li J, Li Y F, He L, et al. Spatio-temporal fusion for remote sensing data: an overview and new benchmark. Sci China Inf Sci, 2020, 63: 140301

2 Gong J Y, Sui H G, Sun K M, et al. Object-level change detection based on full-scale image segmentation and its application to Wenchuan Earthquake. Sci China Ser E-Tech Sci, 2008, 51: 110–122
3 Chen Y X, Zhang Z P, Huang J R, et al. Learning starts from optimizing the composition of temporal information for hyperspectral change detection. IEEE Trans Geosci Remote Sens, 2024, 62: 1–15
4 Mei S H, Geng Y H, Hou J H, et al. Learning hyperspectral images from RGB images via a coarse-to-fine CNN. Sci China Inf Sci, 2022, 65: 152102
5 He J, Qu J H, Zhang X H, et al. Diffusion-driven mutual enhancement of matching and fusion for reference-based hyperspectral image super-resolution. IEEE Trans Geosci Remote Sens, 2026, 1–1
6 Wu J, Cong R M, Fang L Y, et al. Unpaired remote sensing image super-resolution with content-preserving weak supervision neural network. Sci China Inf Sci, 2023, 66: 119105
7 Dong Y N, Zhu B, Yang X C, et al. Deep metric learning based on Brownian covariance representation for few-shot hyperspectral image classification. IEEE Trans Geosci Remote Sens, 2025, 63: 1–13
8 Sun X, Tian Y, Lu W X, et al. From single- to multi-modal remote sensing imagery interpretation: a survey and taxonomy. Sci China Inf Sci, 2023, 66: 140301
9 Li T, Liu T, Li X, et al. Multi-sensor multispectral reconstruction framework based on projection and reconstruction. Sci China Inf Sci, 2024, 67: 132303
10 Xu S F, Xia X F, Li H W, et al. Strengthened residual graph and multiscale gated guided convolutional fusion network for hyperspectral change detection. IEEE Trans Geosci Remote Sens, 2024, 62: 1–14
11 Le Moigne J, Campbell W J, Cromp R F. An automated parallel image registration technique based on the correlation of wavelet features. IEEE Trans Geosci Remote Sens, 2002, 40: 1849–1864
12 Rangarajan A, Chui H, Duncan J S. Rigid point feature registration using mutual information. Med Image Anal, 1999, 3: 425–440
13 Ma J Y, Jiang J J, Zhou H B, et al. Guided locality preserving feature matching for remote sensing image registration. IEEE Trans Geosci Remote Sens, 2018, 56: 4435–4447
14 Wang J, Qu Z G, Kong L S, et al. Learning a perspective-invariant descriptor for remote sensing image matching. IEEE Trans Circ Syst Video Technol, 2025, 35: 12182–12195
15 Yang Y, Qu J, Dong W, et al. Prototype-based meta-prompt tuning: toward rehearsal-free few-shot class-incremental learning for multi-modal remote sensing image. IEEE Trans Image Process, 2026, 35: 434–448
16 Sun W W, Yang G, Huang Y L, et al. First wetland mapping at 10-m spatial resolution in South America using multi-source and multi-feature remote sensing data. Sci China Earth Sci, 2024, 67: 3252–3269
17 Luo F, Chen X, Fu C, et al. HDiff-HIR: hierarchically conditional diffusion model for hyperspectral image reconstruction. IEEE Trans Circ Syst Video Technol, 2026, 36: 777–791
18 Qu J, Zhao J, Dong W, et al. MambaFedCD: spatial-spectral-temporal collaborative Mamba-based active federated hyperspectral change detection. IEEE Trans Image Process, 2026, 35: 1478–1492
19 Zhang Z C, Wang J Y, Wen F R, et al. Large multimodal models evaluation: a survey. Sci China Inf Sci, 2025, 68: 221301
20 Sun Y L, Lei L, Li X, et al. Patch similarity graph matrix-based unsupervised remote sensing change detection with homogeneous and heterogeneous sensors. IEEE Trans Geosci Remote Sens, 2021, 59: 4841–4861
21 Sun Y L, Lei L, Guan D D, et al. Image regression with structure cycle consistency for heterogeneous change detection. IEEE Trans Neural Networks Learn Syst, 2024, 35: 1613–1627
22 Sun Y L, Lei L, Guan D D, et al. Iterative robust graph for unsupervised change detection of heterogeneous remote sensing images. IEEE Trans Image Process, 2021, 30: 6277–6291
23 Gao H M, Feng H, Zhang Y Y, et al. Interactive enhanced network based on multihead self-attention and graph convolution for classification of hyperspectral and LiDAR data. IEEE Trans Geosci Remote Sens, 2024, 62: 1–16
24 Fang L Y, Tang Q, Ouyang L H, et al. Long-tailed object detection of kitchen waste with class-instance balanced detector. Sci China Tech Sci, 2023, 66: 2361–2372
25 Wang Q W, Gu Y F, Yang M, et al. Multi-attribute smooth graph convolutional network for multispectral points classification. Sci China Tech Sci, 2021, 64: 2509–2522
26 Xu S F, Ding X C, Zhang Y Y, et al. Dual-feature attention-based contrastive prototypical clustering for multimodal remote sensing data. IEEE Trans Geosci Remote Sens, 2024, 62: 1–13
27 Zhang T Z, Dong W Q, Li Y S, et al. SAM-prior-guided spatial-spectral-temporal Mamba framework for affine misregistration hyperspectral change detection. IEEE Trans Geosci Remote Sens, 2026, 64: 1–14
28 Jiang X, Li G, Zhang X P, et al. A semisupervised Siamese network for efficient change detection in heterogeneous remote sensing images. IEEE Trans Geosci Remote Sens, 2022, 60: 1–18
29 Luppino L T, Hansen M A, Kampffmeyer M, et al. Code-aligned autoencoders for unsupervised change detection in multimodal remote sensing images. IEEE Trans Neural Networks Learn Syst, 2024, 35: 60–72
30 Li X H, Du Z S, Huang Y Y, et al. A deep translation (GAN) based change detection network for optical and SAR remote sensing images. ISPRS J Photogramm Remote Sens, 2021, 179: 14–34
31 Liu Q, Ren K, Meng X C, et al. Domain adaptive cross reconstruction for change detection of heterogeneous remote sensing images via a feedback guidance mechanism. IEEE Trans Geosci Remote Sens, 2023, 61: 1–16
32 Zhou R F, Quan D, Wang S, et al. A unified deep learning network for remote sensing image registration and change detection. IEEE Trans Geosci Remote Sens, 2024, 62: 1–16
33 Bovolo F, Marchesi S, Bruzzone L. A framework for automatic and unsupervised detection of multiple changes in multitemporal images. IEEE Trans Geosci Remote Sens, 2011, 50: 2196–2212
34 Deng J S, Wang K, Deng Y H, et al. PCA-based land-use change detection and analysis using multitemporal and multisensor satellite data. Int J Remote Sens, 2008, 29: 4823–4838
35 Nielsen A A. The regularized iteratively reweighted MAD method for change detection in multi-and hyperspectral data. IEEE Trans Image Process, 2007, 16: 463–478
36 Wu C, Du B, Zhang L P. Slow feature analysis for change detection in multispectral imagery. IEEE Trans Geosci Remote Sens, 2013, 52: 2858–2874
37 Ahlqvist O. Extending post-classification change detection using semantic similarity metrics to overcome class heterogeneity: a study of 1992 and 2001 U.S. National Land Cover Database changes. Remote Sens Environ, 2008, 112: 1226–1241
38 Cao X, Dong W, Qu J, et al. RMNet: dual-dimensional difference recalibration-guided CNN-VMamba synergistic network for remote sensing image change captioning. IEEE Trans Geosci Remote Sens, 2026, 64: 1–13
39 Dai X Q, Dong Y N, Zhang Y X, et al. AdaptHAD: adaptive one-step hybrid network for hyperspectral anomaly detection. IEEE Trans Geosci Remote Sens, 2025, 63: 1–14
40 Huang J Q, Yuan X C, Lam C T, et al. F3Net: feature filtering fusing network for change detection of remote sensing images. IEEE J Sel Top Appl Earth Observ Remote Sens, 2024, 17: 10621–10635
41 Qu J H, Xu Y S, Dong W Q, et al. Dual-branch difference amplification graph convolutional network for hyperspectral image change detection. IEEE Trans Geosci Remote Sens, 2021, 60: 1–12
42 Yang B, Sun W W, Peng J T. SAGN: sharpening-aware graph network for hyperspectral image change detection. IEEE Trans Geosci Remote Sens, 2024, 62: 1–12
43 Chen Y X, Zhang Z P, Dong L, et al. A joint saliency temporal-spatial-spectral information network for hyperspectral image change

detection. IEEE Trans Geosci Remote Sens, 2024, 62: 1–15
44 Wang Y H, Hong D F, Sha J J, et al. Spectral-spatial-temporal transformers for hyperspectral image change detection. IEEE Trans Geosci Remote Sens, 2022, 60: 1–14
45 Zhou T, Luo F, Fu C, et al. TSCCD: temporal self-construction cross-domain learning for unsupervised hyperspectral change detection. IEEE Trans Image Process, 2026, 35: 830–843
46 Gu A, Goel K, Re C. Efficiently modeling long sequences with structured state spaces. In: Proceedings of the International Conference on Learning Representations, 2022
47 Gu A, Dao T. Mamba: linear-time sequence modeling with selective state spaces. In: Proceedings of the 1st Conference on Language Modeling, 2024
48 Jaderberg M J, Simonyan K, Zisserman A, et al. Spatial transformer networks. In: Proceedings of the 29th International Conference on Neural Information Processing Systems, 2015. 2017–2025
49 Dong W Q, Ren J Y, Xiao S, et al. Cycle translation-based collaborative training for hyperspectral-RGB multimodal change detection. IEEE Trans Image Process, 2025, 34: 6347–6360
50 Sun Y L, Lei L, Guan D D, et al. Graph signal processing for heterogeneous change detection. IEEE Trans Geosci Remote Sens, 2022, 60: 1–23
51 Lv Z Y, Liu J, Sun W W, et al. Hierarchical attention feature fusion-based network for land cover change detection with homogeneous and heterogeneous remote sensing images. IEEE Trans Geosci Remote Sens, 2023, 61: 1–15
52 Li X H, Du Z S, Huang Y Y, et al. A deep translation (GAN) based change detection network for optical and SAR remote sensing images. ISPRS J Photogramm Remote Sens, 2021, 179: 14–34
53 Shao R Z, Du C, Chen H, et al. Sunet: change detection for heterogeneous remote sensing images from satellite and UAV using a dual-channel fully convolution network. Remote Sens, 2021, 13: 3750
54 Pu Y, Gong M G, Liu T F, et al. Adversarial feature equilibrium network for multimodal change detection in heterogeneous remote sensing images. IEEE Trans Geosci Remote Sens, 2024, 62: 1–17
55 Mou L, Bruzzone L, Zhu X X. Learning spectral-spatial-temporal features via a recurrent convolutional neural network for change detection in multispectral imagery. IEEE Trans Geosci Remote Sens, 2019, 57: 924–935
56 Luo F L, Zhou T Y, Liu J M, et al. Multiscale diff-changed feature fusion network for hyperspectral image change detection. IEEE Trans Geosci Remote Sens, 2023, 61: 1–13
57 Chen H, Qi Z P, Shi Z W. Remote sensing image change detection with transformers. IEEE Trans Geosci Remote Sens, 2021, 60: 1–14
58 Ding J G, Li X R, Xiang S, et al. Multilevel features fused and change information enhanced neural network for hyperspectral image change detection. IEEE Trans Geosci Remote Sens, 2023, 62: 1–13