Bi-DiffCD: Bidirectional Diffusion Guided Collaborative Change Detection for Arbitrary-Modal Remote Sensing Images

Jingyu Zhao , **Jiahui Qu** * and **Wenqian Dong**
State Key Laboratory of Integrated Service Network, Xidian University, Xi'an, China
jingyuzhao@stu.xidian.edu.cn, {jhqu, wqdong}@xidian.edu.cn

Abstract

Change detection aims to identify land cover changes by analyzing multitemporal images that cover the same area. However, It may be difficult to effectively obtain high-quality multitemporal images with the same modality in real dynamic scenarios. The rapid development of remote sensing technology enables collaborative observation of multimodal images, but it is challenging for uni-modal image-specific methods to overcome modal discrepancy and achieve complementary advantage detection. To this end, we propose a bidirectional diffusion guided collaborative change detection model (Bi-DiffCD) for arbitrary-modal images, which eliminates the modal discrepancy between arbitrary-modal images through the bidirectional diffusion and makes full use of the multilevel complementary advantage features to improve the detection accuracy. Specifically, a conditional diffusion-based bidirectional modal alignment module (CDBMA) is designed to step-wise align the modal attribute bidirectionally while preserving the multimodal complementary features. Furthermore, a multilevel complementary feature collaborative change detection module (MLCCD) is proposed to collaborate the multilevel enhanced complementary change information from transformed images and potential features for change detection. Experiments have been conducted on three widely used and one self-made multimodal datasets to demonstrate the effectiveness of the proposed method with different combinations of modalities. Code is available at https://github.com/Jiahuiqu/Bi-DiffCD.

1 Introduction

With the rapid development of remote sensing imaging technology, the types and quantities of remote sensing images are increasing, which provides strong support for Earth observation [Qu *et al.*, 2025b; Yang *et al.*, 2024]. Change detection, as an attractive topic within the Earth observation domain, can identify the changes between the multitemporal images covering the same area [Chen *et al.*, 2023]. It has been widely used in a lot of fields, such as disaster monitoring [Zhang and Xia, 2022], environmental monitoring [Chang *et al.*, 2021], and urban planning [Basavaraju *et al.*, 2022].

However, due to the limitations of atmospheric conditions and satellites revisit period, it is difficult to obtain the multitemporal images with same modality covering the same area timely and efficiently [Longbotham *et al.*, 2012]. This limitation significantly hampers the application of change detection. The rapid development of remote sensing technology makes multi-satellite collaborative observation possible [Liu *et al.*, 2019]. Therefore, especially for disaster emergency response, multimodal change detection has received increasing attention. It aims to use multitemporal images with different modalities (e.g., multispectral images and synthetic aperture radar (SAR) images) to provide change information of ground objects [Sun *et al.*, 2021b]. However, since the imaging mechanisms are different, multimodal images depict the same object in varying ways, leading to significant modal discrepancy [Sun *et al.*, 2024a]. Thus, it is difficult to accurately detect changes between multimodal images by directly applying methods designed for uni-modal images.

In recent years, many multimodal change detection approaches have been developed [Qu *et al.*, 2025a]. These methods can be generally divided into three categories, i.e., classification-based methods [Lv *et al.*, 2022], feature transformation-based methods [Li *et al.*, 2021a], and image regression-based methods [Liu *et al.*, 2022]. Classification-based methods can provide the category information, but they excessively rely on the performance of the classifier, which easily leads to cumulative errors. The feature transformation-based methods aim to transform the multimodal images into a common space where the change information can be captured by measuring the similarity of the features [Sun *et al.*, 2021a]. However, it is difficult to capture the common feature space when remote sensing images cover a large area with complex types of ground objects. This imposes limitation on the accuracy of change detection. Recently, the image regression-based methods have gradually attracted the attention of researchers. They can eliminate modal discrepancy by converting images from one modality to another, so that the unique features of multimodal images are preserved [Chen *et al.*, 2022]. However, these deep learning methods typi-

*corresponding author

cally use Generative Adversarial Networks (GANs) for modal transformation. They fail to utilize the features in potential space, and the model training process is unstable, which is prone to mode collapse. Moreover, most existing methods lack experimental validation on high-dimensional data, like hyperspectral image (HSI).

In this paper, we propose a bidirectional diffusion guided collaborative change detection model (Bi-DiffCD) for arbitrary-modal remote sensing images. Specifically, we first transform the "uncomparable" multimodal images into the high fidelity "comparable" uni-modal images through the bidirectional diffusion to eliminate the modal discrepancy between arbitrary-modal images. Then, the complementary advantage difference information can be obtained from the two pair uni-modal images and the process of potential feature space during the modal alignment, which is collaborated to improve the detection accuracy. Furthermore, we have created an RGB-HSI multimodal dataset to evaluate the effectiveness of the proposed method on high dimension data. The main contributions are summarized as follows,

- We propose a bidirectional diffusion guided collaborative change detection model (Bi-DiffCD) for arbitrary-modal remote sensing images, which utilizes the bidirectional diffusion mechanism to perform modal alignment and effectively leverages the multilevel complementary information to improve the detection accuracy.
- A conditional diffusion-based bidirectional modal alignment module (CDBMA) is proposed, which gradually generates two pairs of uni-modal images with high fidelity semantic and modal information through the U-Attention-based conditional diffusion mechanism, so as to eliminate the modal discrepancy.
- We propose a multilevel complementary feature collaborative change detection module (MLCCD), in which a bitemporal anti-attention-based difference feature extractor is designed to mine the difference information of potential feature level and two image levels for multimodal collaborative change detection.
- We create a real RGB-HSI multimodal image change detection dataset, i.e., Liyukou dataset, to further demonstrate the superiority of the proposed method.

2 Related Work

2.1 Classification-Based Methods

The classification-based change detection methods first classify the multimodal images separately, and then compare the classification results of the corresponding areas between the multimodal images to derive changed information [Mubea and Menz, 2012; Hedhli *et al.*, 2014]. Wan et al. [Wan *et al.*, 2019b] proposed a post-classification comparison method for SAR and optical images change detection, which first captures homogeneous objects through multitemporal segmentation, and then utilizes compound classification at the object level to obtain the changed information. Afterwards, they improved the previous work by integrating cooperative multitemporal segmentation and hierarchical compound classification [Wan *et al.*, 2019a]. Han et al. [Han *et al.*, 2021] proposed multimodal change detection method based on post-classification, which utilizes the improved training of hierarchical extreme learning machine to obtain meaningful feature representations and achieve high learning efficiency. However, these methods are rely on the accuracy of classification results and are sensitive to the misclassification.

2.2 Feature Transformation-Based Methods

The feature transformation-based methods transform the multimodal images into a common feature space and capture change information by measuring the similarity of features in this shared space [Volpi *et al.*, 2015; Hedhli *et al.*, 2014]. Some of these methods manually construct the modality-invariant similarity measurement between the multimodal images [Touati and Mignotte, 2018; Mercier *et al.*, 2008; Prendes *et al.*, 2015; Volpi *et al.*, 2015]. Others learn the common feature space through deep learning technology [Wu *et al.*, 2022]. Liu et al. [Liu *et al.*, 2018a] proposed a deep convolutional coupling network, which leverages symmetric network with a convolutional layer and coupling layers to transform the input heterogeneous images into the same feature space. Jiang et al. [Jiang *et al.*, 2022] proposed a transfer learning-based change detection network for heterogeneous images, which uses a transfer learning strategy to freeze low-level features and learn deep features of heterogeneous images, and extracts homogeneous features for change detection. The accuracy of feature transformation-based methods heavily depends on the quality of extracted features. When dealing with complex scenes or being seriously disturbed by image noise, the difficulty of feature extraction increases.

2.3 Image Regression-Based Methods

Regression-based methods map the domain of one image to another, which transform the task into uni-modal change detection [Li *et al.*, 2021b; Luppino *et al.*, 2024; Liu *et al.*, 2018b]. Luppino et al.[Luppino *et al.*, 2019] proposed a framework based on the comparison of affinity matrices and image regression, which quantifies the similarity of affinity matrices to identify the unchanged pixels for transformation process learning. Furthermore, they designed an adversarial cyclic encoder network (ACE-Net) to improve the detection accuracy [Luppino *et al.*, 2022]. Niu et al. [Niu *et al.*, 2019] proposed a conditional adversarial network (cGAN) to realize modal transformation, and used an approximation network to reduce the pixel difference. Gong et al. [Gong *et al.*, 2019] proposed a coupling translation network for change detection, which used coupled generative adversarial networks (GANs) associated with a coupling variational autoencoder (VAE) to translate the heterogeneous images into homogeneous. Although these methods have made some progress, they fail to make full use of the features in potential space during the modal transformation, and the model training process is unstable, which is prone to mode collapse.

3 Methods

3.1 Overview

Given a pair of arbitrary-modal remote sensing images acquired at different times, we denote the image patch with

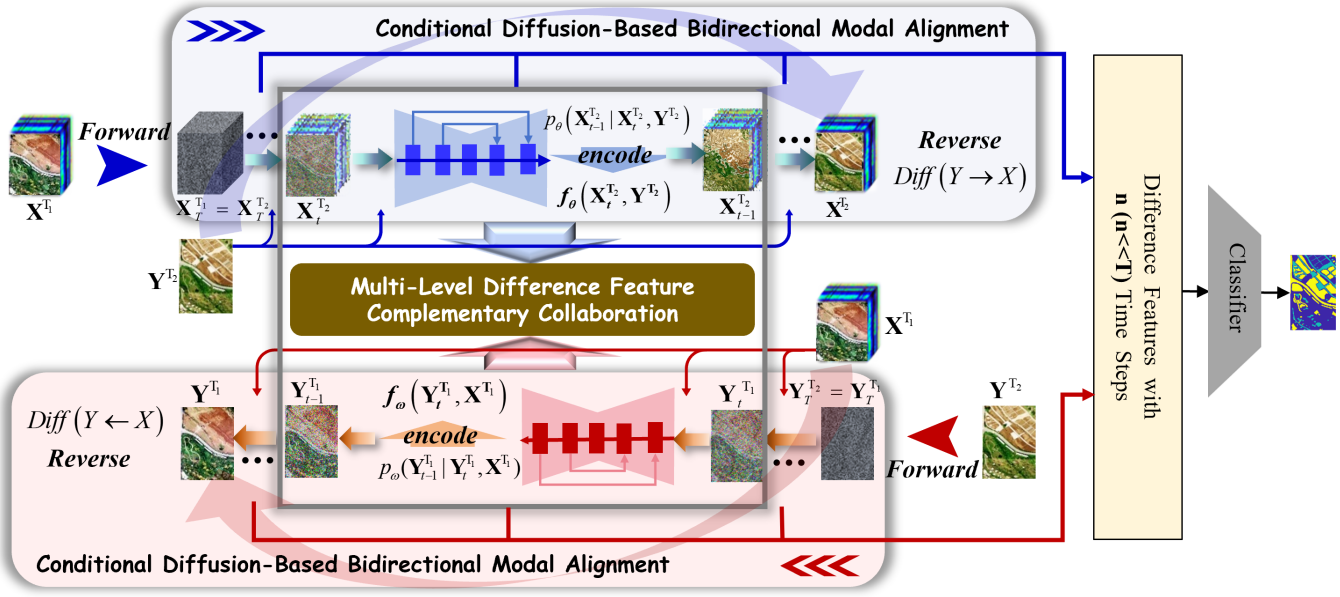

Figure 1: The illustration of the bidirectional diffusion guided collaborative change detection model (Bi-DiffCD) for arbitrary-modal remote sensing images. $\mathbf{X}^{\mathrm{T}_1}$ and $\mathbf{Y}^{\mathrm{T}_2}$ represent two images with different modalities taken at different times ($\mathbf{X} \neq \mathbf{Y}$).

modality X obtained at time T_1 as $\mathbf{X}^{\mathrm{T}_1} \in \mathbb{R}^{H_x \times W_x \times C_x}$, and the image patch with modality Y obtained at time T_2 as $\mathbf{Y}^{\mathrm{T}_2} \in \mathbb{R}^{H_y \times W_y \times C_y}$, where $\mathbf{X} \neq \mathbf{Y}$, $\{H_x, H_y\}$ is the height, $\{W_x, W_y\}$ represents the width, and $\{C_x, C_y\}$ is the number of channels. Due to different imaging mechanisms, the same ground object is represented differently in multimodal images, leading to a significant modal discrepancy between multimodal images. It is difficult to overcome the modal discrepancy and achieve high detection accuracy with the model designed specifically for uni-modal images.

To this end, we propose a bidirectional diffusion-guided collaborative change detection model (Bi-DiffCD) for arbitrary-modal remote sensing images, which performs bidirectional modal alignment through conditional diffusion mechanism and makes full use of the multilevel complementary advantage features to achieve the high accuracy. To be specific, as shown in Figure 1, The modal discrepancy can be gradually mitigated through the conditional diffusion mechanism while preserving the distinctive benefits of multimodal images. In addition, the difference information obtained from the potential feature space and the two image domains can be extracted for multilevel complementary feature collaborative change detection. The proposed method can achieve the remote sensing change detection in arbitrary modality, thereby expanding the application scope of the change detection.

3.2 Conditional Diffusion-Based Bidirectional Modal Alignment

The CDBMA is pretrained with the unchanged patch pairs, which contains two process (forward diffusion and reverse denoising).

Forward Diffusion Process for Adding Noise to MultiModal Images

The forward diffusion process aims to add stochastic noise to the clean multimodal images step by step through a Markov chain to generate the noisy images that subject to Gaussian distribution [Ho *et al.*, 2020]. Supposing that the clean multimodal patch pairs $\mathbf{X}^{\mathrm{T}_1} = \mathbf{X}_0^{\mathrm{T}_1}$ and $\mathbf{Y}^{\mathrm{T}_2} = \mathbf{Y}_0^{\mathrm{T}_2}$ subject to the distribution of $q(\mathbf{X}_0^{\mathrm{T}_1})$ and $q(\mathbf{Y}_0^{\mathrm{T}_2})$, i.e., $\mathbf{X}^{\mathrm{T}_1} \sim q(\mathbf{X}_0^{\mathrm{T}_1})$ and $\mathbf{Y}^{\mathrm{T}_2} \sim q(\mathbf{Y}_0^{\mathrm{T}_2})$. After T time steps, the forward diffusion process can transform the data distribution of multimodal images into standard Gaussian distribution, i.e., $\mathbf{X}_T^{\mathrm{T}_1} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and $\mathbf{Y}_T^{\mathrm{T}_2} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$.

Reverse Denoising Process for Bidirectional Modal Alignment

The reverse denoising process aims to reduce the modal discrepancy by gradually removing the noise added in the forward diffusion process. In this process, the bidirectional diffusion mechanism guide to generate the image $\mathbf{X}^{\mathrm{T}_2}$ with the semantic information of $\mathbf{Y}^{\mathrm{T}_1}$ and the modal attribute information of $\mathbf{X}^{\mathrm{T}_1}$, and generate the image $\mathbf{Y}^{\mathrm{T}_1}$ with the semantic information of $\mathbf{X}^{\mathrm{T}_1}$ and the modal attribute information of $\mathbf{Y}^{\mathrm{T}_2}$, viz., the pair of "uncomparable" multimodal images are transformed into two pairs of "comparable" uni-modal images $\left(\left\{\left(\mathbf{X}^{\mathrm{T}_1}, \mathbf{Y}^{\mathrm{T}_2}\right)\right\} \rightarrow \left\{\left(\mathbf{X}^{\mathrm{T}_1}, \mathbf{X}^{\mathrm{T}_2}\right), \left(\mathbf{Y}^{\mathrm{T}_1}, \mathbf{Y}^{\mathrm{T}_2}\right)\right\}\right)$. The process can be expressed as follows,

$$\begin{array}{l} \mathbf{Y}^{\mathrm{T}_2} \xrightarrow{Diff(Y \rightarrow X)} \mathbf{X}^{\mathrm{T}_2} \\ \mathbf{Y}^{\mathrm{T}_1} \xleftarrow[Diff(Y \leftarrow X)]{} \mathbf{X}^{\mathrm{T}_1} \end{array} \tag{1}$$

where $Diff(Y \rightarrow X)$ represents the transformation process from $\mathbf{Y}^{\mathrm{T}_2}$ to $\mathbf{X}^{\mathrm{T}_2}$, in which the clean image $\mathbf{Y}^{\mathrm{T}_2}$ and the noisy image $\mathbf{X}_T^{\mathrm{T}_1}$ are concatenated as the input, and

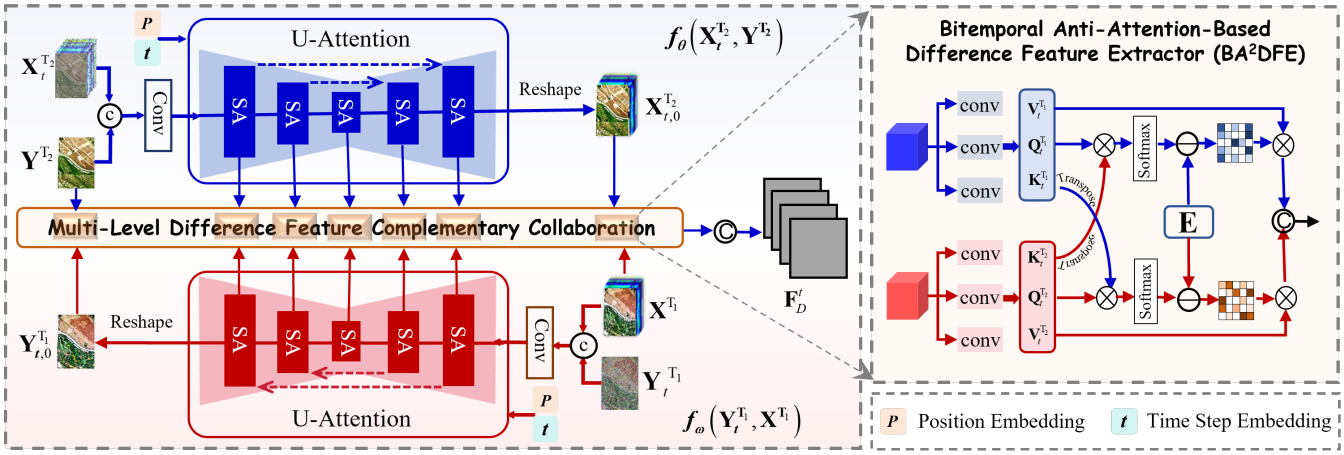

Figure 2: The illustration of Bi-DiffCD in the t-th time step.

$Diff\,(Y \leftarrow X)$ is the transformation process from $\mathbf{X}^{\mathrm{T}_1}$ to $\mathbf{Y}^{\mathrm{T}_1}$, in which the clean image $\mathbf{X}^{\mathrm{T}_1}$ and the noisy image $\mathbf{Y}_T^{\mathrm{T}_2}$ are concatenated as the input.

The posterior probability distribution of the reverse process can be encoded into the deep network, that is,

$$p_\theta\left(\mathbf{X}_{t-1}^{\mathrm{T}_2}|\mathbf{X}_t^{\mathrm{T}_2},\mathbf{Y}^{\mathrm{T}_2}\right)\Rightarrow f_\theta\left(\mathbf{X}_t^{\mathrm{T}_2},\mathbf{Y}^{\mathrm{T}_2}\right) \tag{2}$$

$$p_\omega\left(\mathbf{Y}_{t-1}^{\mathrm{T}_1}|\mathbf{Y}_t^{\mathrm{T}_1},\mathbf{X}^{\mathrm{T}_1}\right)\Rightarrow f_\omega\left(\mathbf{Y}_t^{\mathrm{T}_1},\mathbf{X}^{\mathrm{T}_1}\right) \tag{3}$$

where "$\Rightarrow$" represents the encoding operation, $f_\theta\left(\mathbf{X}_t^{\mathrm{T}_2},\mathbf{Y}^{\mathrm{T}_2}\right)$ represents the $Diff(Y\rightarrow X)$ branch, and $f_\omega\left(\mathbf{Y}_t^{\mathrm{T}_1},\mathbf{X}^{\mathrm{T}_1}\right)$ represents the $Diff(Y\leftarrow X)$ branch. Each time step of the reverse denoising process can provide the prediction of $\mathbf{X}_{t,0}^{\mathrm{T}_2}$ and $\mathbf{Y}_{t,0}^{\mathrm{T}_1}$.

Specifically, to obtain the images with high fidelity semantic information and modal attribute information, we design a U-Attention block to encode the posterior conditional distribution of modal transformation process. The U-Attention block is consisted of K-layer self-attention (SA) [Vaswani *et al.*, 2023] units with skip connection, in which the SA unit can fully couple the semantic feature of conditional image and the modal attribute feature of the noisy image, and skip connection between the shallow and deep layers can prevent information loss during the transition process. The modal transition process in each time step can be expressed as,

$$\begin{aligned}\mathbf{X}_{t,0}^{\mathrm{T}_2}&=f_\theta\left(\mathbf{X}_t^{\mathrm{T}_2},\mathbf{Y}^{\mathrm{T}_2}\right)\\&=F_{\mathrm{MT},\theta}^t\left[f_{SA,\theta}^{K,t}\left(\cdots f_{SA,\theta}^{k,t}\left(\cdots f_{SA,\theta}^{1,t}\left(\mathbf{X}_t^{\mathrm{T}_2};\mathbf{Y}^{\mathrm{T}_2}\right)\right)\right)\right]\end{aligned} \tag{4}$$

$$\begin{aligned}\mathbf{Y}_{t,0}^{\mathrm{T}_1}&=f_\omega\left(\mathbf{Y}_t^{\mathrm{T}_1},\mathbf{X}^{\mathrm{T}_1}\right)\\&=F_{\mathrm{MT},\omega}^t\left[f_{SA,\omega}^{K,t}\left(\cdots f_{SA,\omega}^{k,t}\left(\cdots f_{SA,\omega}^{1,t}\left(\mathbf{Y}_t^{\mathrm{T}_1};\mathbf{X}^{\mathrm{T}_1}\right)\right)\right)\right]\end{aligned} \tag{5}$$

where $f_{SA,\theta}^{k,t}$ and $f_{SA,\omega}^{k,t}$ represent the k-th SA unit in the t-th time step, and $F_{\mathrm{MT},\theta}^t$ and $F_{\mathrm{MT},\omega}^t$ denote the time step embedding and the skip connection.

The transformed images can be further optimized through continuous iteration. The input for the next time step can be calculated from the variance and mean of the posterior distribution as follows,

$$\begin{aligned}\mathbf{X}_{t-1}^{\mathrm{T}_2}&=\frac{\sqrt{\bar{\alpha}_{t-1,X}}\left(1-\alpha_{t,X}\right)}{1-\bar{\alpha}_{t,X}}f_\theta\left(\mathbf{X}_t^{\mathrm{T}_2},\mathbf{Y}^{\mathrm{T}_2}\right)\\&+\frac{\sqrt{\alpha_{t,X}}\left(1-\bar{\alpha}_{t-1,X}\right)}{1-\bar{\alpha}_{t,X}}\mathbf{X}_t^{\mathrm{T}_2}+\sqrt{\frac{\left(1-\bar{\alpha}_{t-1,X}\right)\left(1-\alpha_{t,X}\right)}{1-\bar{\alpha}_{t,X}}}\varepsilon\end{aligned} \tag{6}$$

$$\begin{aligned}\mathbf{Y}_{t-1}^{\mathrm{T}_1}&=\frac{\sqrt{\bar{\alpha}_{t-1,Y}}\left(1-\alpha_{t,Y}\right)}{1-\bar{\alpha}_t}f_\omega\left(\mathbf{Y}_t^{\mathrm{T}_1},\mathbf{X}^{\mathrm{T}_1}\right)\\&+\frac{\sqrt{\alpha_{t,Y}}\left(1-\bar{\alpha}_{t-1,Y}\right)}{1-\bar{\alpha}_{t,Y}}\mathbf{Y}_t^{\mathrm{T}_1}+\sqrt{\frac{\left(1-\bar{\alpha}_{t-1,Y}\right)\left(1-\alpha_{t,Y}\right)}{1-\bar{\alpha}_{t,Y}}}\varepsilon\end{aligned} \tag{7}$$

where $\varepsilon\in N\left(\mathbf{0},\mathbf{I}\right)$, $\overline{\alpha}_{t,X}=\prod_{i=0}^{t}\alpha_{i,X}$, $\overline{\alpha}_{t,Y}=\prod_{i=0}^{t}\alpha_{i,Y}$, $\{\alpha_{i,X},\alpha_{i,Y}\}$ is the hyperparameter.

3.3 Multilevel Difference Feature Complementary Collaborative Change Detection

Bidirectional modal transformation processes progressively achieve the modal alignment by step-wise denoising operation. The features in potential space during the transformation process can usually provide more abstract high-dimensional information. The two image domains content abundant shape structure and texture information, which is conducive to providing details of changes. To maximize the change detection accuracy, we design a multilevel complementary feature collaborative change detection module (MLCCD) to collaboratively utilize the discriminative features in both image domains and potential feature spaces. Thus, the model can achieve high-accuracy change detection by fully exploiting the features with complementary advantages at multiple levels. Specifically, to effectively capture the discriminative difference information, we design a bitemporal anti-attention-based difference feature extractor (BA^{2}DFE) as shown in Figure 2, which highlights the changed information and suppress the unchanged information by re-weighting the similarity of bitemporal inputs.

Take the feature extraction process of bitemporal images with modal X in the t-th time step serves as an illustrative

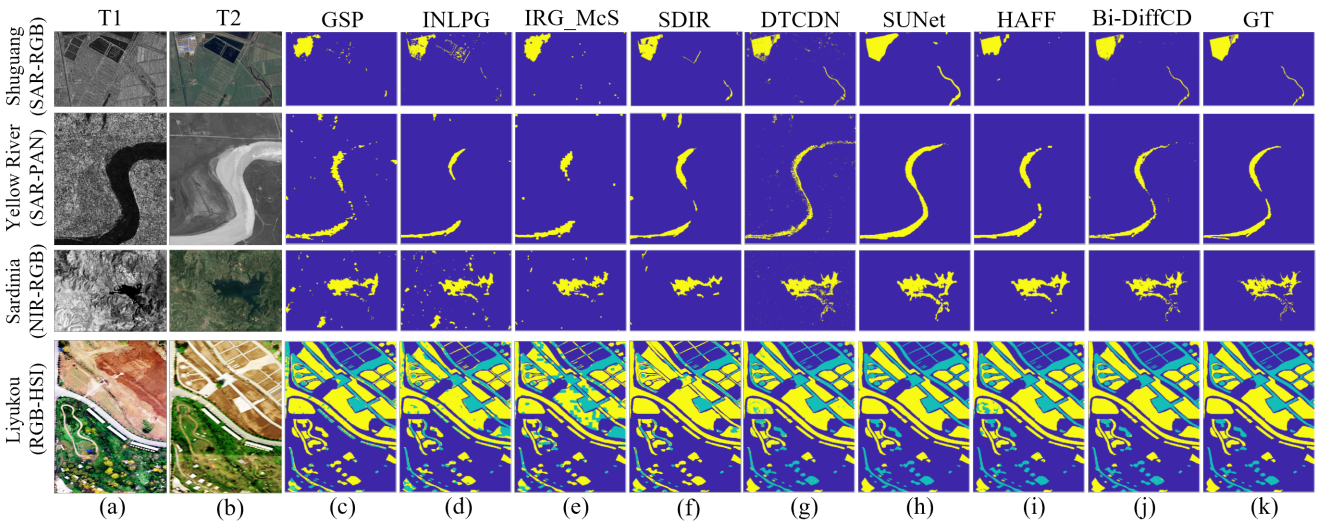

Figure 3: Visualization results of different methods on four datasets. (a) The image captured at time T_1. (b) The image captured at time T_2. (c) GSP. (d) INLPG. (e) IRG-McS. (f) SDIR (g) DTCDN. (h) SUNet. (i)HAFF. (j) Bi-DiffCD. (k) Ground Truth (GT).

example. We assign lower weights to regions exhibiting high similarity and higher weights to regions displaying low similarity. The weight matrices can be obtained by follows,

$$\begin{cases} \mathbf{W}_{t,X}^{\mathrm{T}_1} = \mathbf{E} - f_\sigma\left(\mathbf{Q}_{t,X}^{\mathrm{T}_1}\left(\mathbf{K}_{t,X}^{\mathrm{T}_2}\right)^{\mathrm{T}}\right) \\ \mathbf{W}_{t,X}^{\mathrm{T}_2} = \mathbf{E} - f_\sigma\left(\mathbf{Q}_{t,X}^{\mathrm{T}_2}\left(\mathbf{K}_{t,X}^{\mathrm{T}_1}\right)^{\mathrm{T}}\right) \end{cases} \tag{8}$$

where $\mathbf{W}_{t,X}^{\mathrm{T}_1}$ and $\mathbf{W}_{t,X}^{\mathrm{T}_2}$ are the weight matrices of bitemporal images, $\mathbf{Q}_{t,X}^{\mathrm{T}_1}$, $\mathbf{K}_{t,X}^{\mathrm{T}_1}$, and $\mathbf{V}_{t,X}^{\mathrm{T}_1}$ are the feature matrices obtained by projecting $\mathbf{X}_t^{\mathrm{T}_1}$ into different spaces through convolutional operation, $\mathbf{Q}_{t,X}^{\mathrm{T}_2}$, $\mathbf{K}_{t,X}^{\mathrm{T}_2}$, and $\mathbf{V}_{t,X}^{\mathrm{T}_2}$ are the feature matrices obtained by projecting $\mathbf{X}_t^{\mathrm{T}_2}$ into different spaces through convolutional operation, $\mathbf{E}$ is the identity matrix, and $f_\sigma(\cdot)$ is the activation function. The difference feature extraction process can be formulated as follows,

$$\mathbf{F}_{D_X}^t = \left[\frac{\mathbf{W}_{t,X}^{\mathrm{T}_1}}{\sqrt{B}}\mathbf{V}_{t,X}^{\mathrm{T}_1}; \frac{\mathbf{W}_{t,X}^{\mathrm{T}_2}}{\sqrt{B}}\mathbf{V}_{t,X}^{\mathrm{T}_2}\right] \tag{9}$$

where $\mathbf{F}_{D_X}^t$ is the difference feature obtained from the images with modal X in the t-th time step, $[\cdot;\cdot]$ represents the concatenation operation, and B is the scale factor.

Similarly, BA^{2}DFE is used to extract the difference features from output features of the K-layer SA units. The difference features obtained from the potential feature space in the modal alignment process are denoted as $\left\{\mathbf{F}_{D_{SA}}^{t,1}, \cdots, \mathbf{F}_{D_{SA}}^{t,k}, \cdots \mathbf{F}_{D_{SA}}^{t,K}\right\}$. In addition, the difference feature captured from the bitemporal images with modal Y can be denoted as $\mathbf{F}_{D_Y}^t$. The difference feature from the t-th time step can be obtained by the concatenate operation,

$$\mathbf{F}_{D_{SA}}^t = [\mathbf{F}_{D_{SA}}^{t,1}; \cdots; \mathbf{F}_{D_{SA}}^{t,k}; \cdots; \mathbf{F}_{D_{SA}}^{t,K}] \tag{10}$$

$$\mathbf{F}_D^t = [\mathbf{F}_{D_X}^t; \mathbf{F}_{D_{SA}}^t; \mathbf{F}_{D_Y}^t] \tag{11}$$

where $\mathbf{F}_{D_{SA}}^t$ is the feature during the bidirectional modal alignment process, $\mathbf{F}_D^t$ is the multilevel difference feature. Further, we select the difference features with n ($n \ll T$) time steps to construct the enhanced difference feature $\mathbf{F}_D$. The results can be obtained from $\mathbf{F}_D$ through the classifier,

$$\widetilde{\mathbf{Z}} = f_C(\mathbf{F}_D) \tag{12}$$

where $\widetilde{\mathbf{Z}}$ is the change detection result.

3.4 Training Objective

Given a training set $\left\{\left(\mathbf{X}_n^{\mathrm{T}_1}, \mathbf{Y}_n^{\mathrm{T}_2}, \mathbf{Z}_n\right) | n \in [1, N]\right\}$, where $\left\{\mathbf{X}_n^{\mathrm{T}_1}, \mathbf{Y}_n^{\mathrm{T}_2}\right\}$ is the n-th training samples, $\mathbf{Z}_n$ is the corresponded ground truth, and N is the number of training samples. To ensure the high fidelity of semantic information and modal attributes during the modal alignment process, the consistency constraint strategy is used for model training,

$$L_1 = \frac{1}{N}\sum_{n=1}^{N}\left(\left\|\mathbf{X}_n^{\mathrm{T}_2} - \mathbf{X}_n^{\mathrm{T}_1}\right\|_1 + \left\|\mathbf{Y}_n^{\mathrm{T}_1} - \mathbf{Y}_n^{\mathrm{T}_2}\right\|_1\right) \tag{13}$$

Additionally, the cross-entropy loss measures the proximity between prediction results and ground truth values,

$$L_2 = -\frac{1}{N}\sum_{n=1}^{N}\left(\mathbf{Z}_n \log \widetilde{\mathbf{Z}}_n + (1 - \mathbf{Z}_n)\log\left(1 - \widetilde{\mathbf{Z}}_n\right)\right) \tag{14}$$

where $\widetilde{\mathbf{Z}}_n$ is the prediction of n-th sample pair. Therefore, the consistency constraint and cross-entropy loss are used to optimize the model for optimal performance. The total loss function L can be expressed as follows,

$$L = \{L_1, L_2\} \tag{15}$$

When the loss function value converges to the minimum value and remains relatively unchanged during model training, the model achieves optimization.

Dataset	Liyukou (RGB-HSI)			Sardinia (NIR-RGB)			Yellow River (SAR-PAN)			Shuguang (SAR-RGB)		
Index	OA	KC	F1	OA	KC	F1	OA	KC	F1	OA	KC	F1
GSP	0.8436	0.6294	0.7398	0.9605	0.6822	0.7033	0.9770	0.7061	0.7180	0.9763	0.7489	0.7611
INLPG	0.8126	0.5669	0.7036	0.9605	0.6821	0.7003	0.9763	0.6878	0.7002	0.9656	0.6523	0.6703
IRG-McS	0.7499	0.4321	0.6181	0.9720	0.7522	0.7670	0.9736	0.6197	0.6333	0.9801	0.8066	0.8171
SDIR	0.8085	0.5084	0.6225	0.9679	0.6910	0.7075	0.9804	0.7378	0.7479	0.9803	0.8014	0.8116
HAFF	0.9702	0.9329	0.9553	0.9823	0.8557	0.8652	0.9901	0.8666	0.8718	0.9811	0.7983	0.8080
SUNet	0.9860	0.9686	0.9790	0.9748	0.8187	0.8320	0.9811	0.7926	0.8022	0.9817	0.8526	0.8621
DTCDN	0.9717	0.9359	0.9571	0.9768	0.8233	0.8357	0.9763	0.6878	0.7003	0.9823	0.8395	0.8489
Bi-DiffCD	**0.9935**	**0.9853**	**0.9901**	**0.9877**	**0.8991**	**0.9056**	**0.9918**	**0.8884**	**0.8927**	**0.9926**	**0.9300**	**0.9339**

Table 1: Quantitative results obtained by different methods on the four datasets.

4 Experiment

4.1 Datasets

Liyukou Dataset (RGB-HSI): Liyukou dataset is a self-made dataset, which describes the changes in Liyukou Village, Xi'an, Shaanxi Province, China. The dataset includes a HSI taken by the Headwall VNIR hyperspectral camera on April 6, 2023 and a RGB image captured by the FILR DUO Rro R sensor on November 18, 2022. The RGB image contains 729×351 pixels. The HSI contains 243×117 pixels and has 270 bands covering a spectral coverage of $0.4 \sim 1.0 \mu m$. 194 bands were selected for change detection. In this paper, the HSI is upsampled to the same size as the RGB image before being fed into the model.

Sardinia Dataset (NIR-RGB): Sardinia dataset describes the changes due to lake overflows in the region of Sardinia, Italy. The dataset consists of a near-infrared (NIR) image taken by Landsat-5 in September 1995 and a RGB image captured on Google Earth in July 1996. Each image has 3 bands, and contains 405×297 pixels.

Yellow River Dataset (SAR-PAN): Yellow River dataset describes the changes along the Yellow River, China. The dataset contains a SAR image and a Panchromatic (PAN) image. The SAR image was captured by Radarsat-2 in June 2008, and the PAN image was obtained from Landsat-7 in September 2010. Each image has 3 bands, and contains 324×270 pixels.

Shuguang Dataset (SAR-RGB): Shuguang dataset describes the changes caused by the construction of Shuguang Village in Dongying, China. The dataset contains a SAR image and a RGB image. The SAR image was taken by the Radarsat-2 satellite in June 2008, and the RGB image was obtained from Google Earth in September 2012. Each image has 3 bands, and contains 810×540 pixels.

4.2 Experimental Settings and Competing Methods

All experiments were conducted with the PyTorch framework on two NVIDIA GeForce RTX 3090 GPU. Adam optimizer was chosen to train the model. The batch size was set to 64. The CDBMA module has been pre-trained for 1000 epochs with a learning rate of 0.0001. The time step T was set to 2000, and the uniform increment range of hyperparameter sequence $\{\alpha_1, \alpha_2, \ldots, \alpha_n\}$ was $0.98 \sim 1$. The MLCCD module has been pre-trained for 2000 epochs with a learning rate of 0.0001. The difference features obtained at the time step $t = 1, 1000, 2000$ are used for change detection.

Dataset	Number	OA	KC	F1
Liyukou (RGB-HSI)	1	0.9858	0.9676	0.9781
	3	0.9892	0.9755	0.9835
	5	**0.9935**	**0.9853**	**0.9901**
	7	0.9914	0.9804	0.9868
Sardinia (NIR-RGB)	1	0.9819	0.8510	0.8607
	3	0.9859	0.8861	0.8936
	5	**0.9877**	**0.8991**	**0.9056**
	7	0.9874	0.8986	0.9053
Yellow River (SAR-PAN)	1	0.9865	**0.9330**	0.8399
	3	0.9906	0.8755	0.8804
	5	**0.9918**	0.8884	**0.8927**
	7	0.9911	0.8739	0.8785
Shuguang (SAR-RGB)	1	0.9889	0.8994	0.9053
	3	0.9901	0.9034	0.9086
	5	**0.9926**	**0.9300**	**0.9339**
	7	0.9918	0.9257	0.9301

Table 2: Experimental results of Bi-DiffCD with different number of self-attention (SA) units in U-Attention block.

To verify the superiority of the proposed method, seven representative methods are selected for comparison experiments, including GSP [Sun *et al.*, 2022a], INLPG [Sun *et al.*, 2022b], IRG-McS [Sun *et al.*, 2021a], SDIR [Sun *et al.*, 2024b], HAFF [Lv *et al.*, 2023], SUNeT [Shao *et al.*, 2021], and DTCDN [Li *et al.*, 2021b]. Three widely used indexes are used for quantitative evaluation of the experiments, including overall accuracy (OA), Kappa coefficient (KC) and F1-score (F1). The higher the values of OA, KC, and F1, the better the detection performance of the method.

4.3 Quantitative and Qualitative Comparison

The visualization results in Figure 3 show that for the Liyukou dataset (RGB-HSI), yellow pixels represent unchanged areas, green pixels indicate changed areas, and blue pixels represent unmarked areas. For the other three datasets, blue pixels indicate unchanged areas, and yellow pixels denote changed areas. The first three methods show significant missing detections. The changed areas, especially in river detection, cannot be accurately identified. The deep learning-based methods can achieve the great performance in multi-

Dataset	Model	OA	KC	F1
Liyukou	w/ U-Net	0.9892	0.9756	0.9836
(RGB-HSI)	Bi-DiffCD	**0.9935**	**0.9853**	**0.9901**
Sardinia	w/ U-Net	0.9873	0.8905	0.8973
(NIR-RGB)	Bi-DiffCD	**0.9877**	**0.8991**	**0.9056**
Yellow River	w/ U-Net	0.9911	0.8797	0.8843
(SAR-PAN)	Bi-DiffCD	**0.9918**	**0.8884**	**0.8927**
Shuguang	w/ U-Net	0.9899	0.8712	0.8764
(SAR-RGB)	Bi-DiffCD	**0.9926**	**0.9300**	**0.9339**

Table 3: Experimental results of Bi-DiffCD and its variant with U-Net (w/ U-Net) on four datasets.

Dataset	Model	OA	KC	F1
Liyukou	w/o $\mathbf{F}^t_{D_{SA}}$	0.9879	0.9726	0.9815
(RGB-HSI)	Bi-DiffCD	**0.9935**	**0.9853**	**0.9901**
Sardinia	w/o $\mathbf{F}^t_{D_{SA}}$	0.9819	0.8432	0.8528
(NIR-RGB)	Bi-DiffCD	**0.9877**	**0.8991**	**0.9056**
Yellow River	w/o $\mathbf{F}^t_{D_{SA}}$	0.9885	0.8424	0.8483
(SAR-PAN)	Bi-DiffCD	**0.9918**	**0.8884**	**0.8927**
Shuguang	w/o $\mathbf{F}^t_{D_{SA}}$	0.9883	0.8933	0.8994
(SAR-RGB)	Bi-DiffCD	**0.9926**	**0.9300**	**0.9339**

Table 4: Experiments results of Bi-DiffCD and its variant that without $\mathbf{F}^t_{D_{SA}}$ (w/o $\mathbf{F}^t_{D_{SA}}$) on four datasets.

modal change detection tasks. Among all methods, SUNet and DTCDN can detect the changes of river in the lower right corner of Shuguang dataset. However, there are more noise in the change map from DTCDN. Bi-DiffCD demonstrates optimal performance across all datasets.

To further evaluate the performance of different methods, the values of OA, KC, and F1 for detection results are calculated as shown in Table 1. The best performance is labeled in bold. GSP and INLPG achieve similar results. The detection performance of deep learning-based methods is usually better than that of traditional methods. Bi-DiffCD exhibits the best performance across four datasets, and achieves a good balance between accuracy and recall. Especially on Liyukou dataset, the proposed method achieves high values of 0.9935, 0.9853, and 0.9901 for OA, KC, and F1 respectively. It is particularly effective in detecting subtle changes.

4.4 Ablation Study

Number of SA Units in U-Attention Block

Insufficient units can weaken the capability of feature extraction, while too many units can increase training complexity and risk overfitting. To find an appropriate number of SA units, we designed variants of Bi-DiffCD with different number of SA units. Results in Table 2 show that the model performs best with 5 SA units in the U-Attention block.

Effectiveness of U-Attention Block

To verify the effectiveness of U-Attention block, we designed a variant of Bi-DiffCD (w/ U-Net), in which the U-Attention block is replaced with the widely used denoising framework, U-Net. Experimental results are shown in Table 3. Since the U-Attention block can fully couple semantic information and modal attributes to obtain high-quality modal transformation results, Bi-DiffCD can achieve the better detection results. This can demonstrate the effectiveness of U-Attention block.

Dataset	Model	OA	KC	F1
Liyukou	w/ Conv	0.9850	0.9658	0.9768
(RGB-HSI)	Bi-DiffCD	**0.9935**	**0.9853**	**0.9901**
Sardinia	w/ Conv	0.9827	0.8545	0.8637
(NIR-RGB)	Bi-DiffCD	**0.9877**	**0.8991**	**0.9056**
Yellow River	w/ Conv	0.9889	0.8488	0.8546
(SAR-PAN)	Bi-DiffCD	**0.9918**	**0.8884**	**0.8927**
Shuguang	w/ Conv	0.9919	0.9244	0.9286
(SAR-RGB)	Bi-DiffCD	**0.9926**	**0.9300**	**0.9339**

Table 5: Experimental results of Bi-DiffCD and its variant replacing BA^{2}DFE with convolutional layers (w/ Conv) on four datasets.

Effectiveness of Difference Features in Potential Feature Space

To verify the effectiveness of difference features extracted from potential feature space, we designed a variant of the proposed model that did not use $\mathbf{F}^t_{D_{SA}}$ for change detection (w/o $\mathbf{F}^t_{D_{SA}}$), and conducted experiments on four datasets. The experimental results are shown in Table 4. The proposed method can achieve better performance with the difference information of the potential feature domain, which proved that difference features extracted from potential feature space is useful for change detection.

Effectiveness of BA^{2}DFE

To verify the effectiveness of BA^{2}DFE, we replaced the BA^{2}DFE with convolution layers that have an equal number of parameters to construct a variant of Bi-DiffCD (w/ Conv). Compared with the variant, Bi-DiffCD designs a bitemporal anti-attention mechanism that assigns larger weights to dissimilar features to obtain enhanced difference features. The experimental results are shown in Table 5, the detection performance of the proposed method is better than the variant, which demonstrates the effectiveness of BA^{2}DFE.

5 Conclusion

We propose a change detection model (Bi-DiffCD) for arbitrary-modal remote sensing images. The modal alignment can be first achieved through the bidirectional conditional diffusion process. Then, the multilevel difference information from the potential feature level and two image levels in the modal alignment process can be fully exploited for change detection. Extensive experiments are conducted on four datasets with different modality combination to demonstrate the effectiveness of Bi-DiffCD. It is proved that Bi-DiffCD is not limited to a specific data source, and can effectively expand the application scope of change detection.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant 62201423 and 62471359, the Key Research and Development Program of Shaanxi under Grant 2025SF-YBXM-513, the Young Talent Fund of Association for Science and Technology in Shaanxi under Grant 20230117 and 20250133, the Young Talent Fund of Xi'an Association for Science and Technology under Grant 959202313052, in part by the Fundamental Research Funds for the Central Universities under Grant QTZX25084.

References

[Basavaraju *et al.*, 2022] K. S. Basavaraju, N. Sravya, Shyam Lal, J. Nalini, Chintala Sudhakar Reddy, and Fabio Dell'Acqua. Ucdnet: A deep learning model for urban change detection from bi-temporal multispectral sentinel-2 satellite images. *IEEE Trans. Geosci. Remote Sens.*, 60:1–10, 2022.

[Chang *et al.*, 2021] Minghui Chang, Xiangchao Meng, Weiwei Sun, Gang Yang, and Jiangtao Peng. Collaborative coupled hyperspectral unmixing based subpixel change detection for analyzing coastal wetlands. *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, 14:8208–8224, 2021.

[Chen *et al.*, 2022] Hao Chen, Fachuan He, and Jinming Liu. Heterogeneous images change detection based on iterative joint global–local translation. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, 15:9680–9698, 2022.

[Chen *et al.*, 2023] Chao-Peng Chen, Jun-Wei Hsieh, Ping-Yang Chen, Yi-Kuan Hsieh, and Bor-Shiun Wang. Sarasnet: scale and relation aware siamese network for change detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 14187–14195, 2023.

[Gong *et al.*, 2019] Maoguo Gong, Xudong Niu, Tao Zhan, and Mingyang Zhang. A coupling translation network for change detection in heterogeneous images. *Int. J. Remote Sens.*, 40(9):3647–3672, 2019.

[Han *et al.*, 2021] Te Han, Yuqi Tang, Xin Yang, Zefeng Lin, Bin Zou, and Huihui Feng. Change detection for heterogeneous remote sensing images with improved training of hierarchical extreme learning machine (helm). *Remote Sens.*, 13(23), 2021.

[Hedhli *et al.*, 2014] Ihsen Hedhli, Gabriele Moser, Josiane Zerubia, and Sebastiano B. Serpico. New cascade model for hierarchical joint classification of multitemporal, multiresolution and multisensor remote sensing data. In *2014 IEEE International Conference on Image Processing (ICIP)*, pages 5247–5251, 2014.

[Ho *et al.*, 2020] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models, 2020.

[Jiang *et al.*, 2022] Xiao Jiang, Gang Li, Xiao-Ping Zhang, and You He. A semisupervised siamese network for efficient change detection in heterogeneous remote sensing images. *IEEE Trans. Geosci. Remote Sens.*, 60:1–18, 2022.

[Li *et al.*, 2021a] Hao Li, Maoguo Gong, Mingyang Zhang, and Yue Wu. Spatially self-paced convolutional networks for change detection in heterogeneous images. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, 14:4966–4979, 2021.

[Li *et al.*, 2021b] Xinghua Li, Zhengshun Du, Yanyuan Huang, and Zhenyu Tan. A deep translation (gan) based change detection network for optical and sar remote sensing images. *ISPRS J. Photogramm. Remote Sens.*, 179:14–34, 2021.

[Liu *et al.*, 2018a] Jia Liu, Maoguo Gong, Kai Qin, and Puzhao Zhang. A deep convolutional coupling network for change detection based on heterogeneous optical and radar images. *IEEE Trans. Neural Networks Learn. Syst.*, 29(3):545–559, 2018.

[Liu *et al.*, 2018b] Zhunga Liu, Gang Li, Gregoire Mercier, You He, and Quan Pan. Change detection in heterogenous remote sensing images via homogeneous pixel transformation. *IEEE Trans. Image Process.*, 27(4):1822–1834, 2018.

[Liu *et al.*, 2019] Sicong Liu, Daniele Marinelli, Lorenzo Bruzzone, and Francesca Bovolo. A review of change detection in multitemporal hyperspectral images: Current techniques, applications, and challenges. *IEEE Geosci. Remote Sens. Mag.*, 7(2):140–158, 2019.

[Liu *et al.*, 2022] Zhun-Ga Liu, Zuo-Wei Zhang, Quan Pan, and Liang-Bo Ning. Unsupervised change detection from heterogeneous data based on image translation. *IEEE Trans. Geosci. Remote Sens.*, 60:1–13, 2022.

[Longbotham *et al.*, 2012] Nathan Longbotham, Fabio Pacifici, Taylor Glenn, Alina Zare, Michele Volpi, Devis Tuia, Emmanuel Christophe, Julien Michel, Jordi Inglada, Jocelyn Chanussot, and Qian Du. Multi-modal change detection, application to the detection of flooded areas: Outcome of the 2009–2010 data fusion contest. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.*, 5(1):331–342, 2012.

[Luppino *et al.*, 2019] Luigi Tommaso Luppino, Filippo Maria Bianchi, Gabriele Moser, and Stian Normann Anfinsen. Unsupervised image regression for heterogeneous change detection. *IEEE Trans. Geosci. Remote Sens.*, 57(12):9960–9975, 2019.

[Luppino *et al.*, 2022] Luigi Tommaso Luppino, Michael Kampffmeyer, Filippo Maria Bianchi, Gabriele Moser, Sebastiano Bruno Serpico, Robert Jenssen, and Stian Normann Anfinsen. Deep image translation with an affinity-based change prior for unsupervised multimodal change detection. *IEEE Trans. Geosci. Remote Sens.*, 60:1–22, 2022.

[Luppino *et al.*, 2024] Luigi Tommaso Luppino, Mads Adrian Hansen, Michael Kampffmeyer, Filippo Maria Bianchi, Gabriele Moser, Robert Jenssen, and Stian Normann Anfinsen. Code-aligned autoencoders for unsupervised change detection in multimodal remote sensing images. *IEEE Trans. Neural Networks Learn. Syst.*, 35(1):60–72, 2024.

[Lv *et al.*, 2022] ZhiYong Lv, HaiTao Huang, Xinghua Li, MingHua Zhao, Jón Atli Benediktsson, WeiWei Sun, and

Nicola Falco. Land cover change detection with heterogeneous remote sensing images: Review, progress, and perspective. *Proc. IEEE*, 110(12):1976–1991, 2022.

[Lv *et al.*, 2023] Zhiyong Lv, Jie Liu, Weiwei Sun, Tao Lei, Jón Atli Benediktsson, and Xiuping Jia. Hierarchical attention feature fusion-based network for land cover change detection with homogeneous and heterogeneous remote sensing images. *IEEE Trans. Geosci. Remote Sens.*, 61:1–15, 2023.

[Mercier *et al.*, 2008] GrEgoire Mercier, Gabriele Moser, and Sebastiano B. Serpico. Conditional copulas for change detection in heterogeneous remote sensing images. *IEEE Trans. Geosci. Remote Sens.*, 46(5):1428–1441, 2008.

[Mubea and Menz, 2012] Kenneth Mubea and Gunter Menz. Monitoring land-use change in nakuru (kenya) using multi-sensor satellite data. *Advances in Remote Sensing*, 01, 12 2012.

[Niu *et al.*, 2019] Xudong Niu, Maoguo Gong, Tao Zhan, and Yuelei Yang. A conditional adversarial network for change detection in heterogeneous images. *IEEE Geosci. Remote Sens. Lett.*, 16(1):45–49, 2019.

[Prendes *et al.*, 2015] Jorge Prendes, Marie Chabert, Frédéric Pascal, Alain Giros, and Jean-Yves Tourneret. A new multivariate statistical model for change detection in images acquired by homogeneous and heterogeneous sensors. *IEEE Trans. Image Process.*, 24(3):799–812, 2015.

[Qu *et al.*, 2025a] Jiahui Qu, Wenqian Dong, Qian Du, Yufei Yang, Yunshuang Xu, and Yunsong Li. Cyclic consistency constrained multiview graph matching network for unsupervised heterogeneous change detection. *IEEE Trans. Geosci. Remote Sens.*, 63:1–15, 2025.

[Qu *et al.*, 2025b] Jiahui Qu, Xiaoyang Wu, Wenqian Dong, Jizhou Cui, and Yunsong Li. IR&ArF: Toward deep interpretable arbitrary resolution fusion of unregistered hyperspectral and multispectral images. *IEEE Trans. Image Process.*, 34:1934–1949, 2025.

[Shao *et al.*, 2021] Ruizhe Shao, Chun Du, Hao Chen, and Jun Li. Sunet: Change detection for heterogeneous remote sensing images from satellite and uav using a dual-channel fully convolution network. *Remote Sens.*, 13(18):3750, 2021.

[Sun *et al.*, 2021a] Yuli Sun, Lin Lei, Dongdong Guan, and Gangyao Kuang. Iterative robust graph for unsupervised change detection of heterogeneous remote sensing images. *IEEE Trans. Image Process.*, 30:6277–6291, 2021.

[Sun *et al.*, 2021b] Yuli Sun, Lin Lei, Xiao Li, Hao Sun, and Gangyao Kuang. Nonlocal patch similarity based heterogeneous remote sensing change detection. *Pattern Recognit.*, 109:107598, 2021.

[Sun *et al.*, 2022a] Yuli Sun, Lin Lei, Dongdong Guan, Gangyao Kuang, and Li Liu. Graph signal processing for heterogeneous change detection. *IEEE Trans. Geosci. Remote Sens.*, 60:1–23, 2022.

[Sun *et al.*, 2022b] Yuli Sun, Lin Lei, Xiao Li, Xiang Tan, and Gangyao Kuang. Structure consistency-based graph for unsupervised change detection with homogeneous and heterogeneous remote sensing images. *IEEE Trans. Geosci. Remote Sens.*, 60:1–21, 2022.

[Sun *et al.*, 2024a] Yuli Sun, Lin Lei, Dongdong Guan, Junzheng Wu, Gangyao Kuang, and Li Liu. Image regression with structure cycle consistency for heterogeneous change detection. *IEEE Trans. Neural Networks Learn. Syst.*, 35(2):1613–1627, 2024.

[Sun *et al.*, 2024b] Yuli Sun, Lin Lei, Zhang Li, and Gangyao Kuang. Similarity and dissimilarity relationships based graphs for multimodal change detection. *ISPRS J. Photogramm. Remote Sens.*, 208:70–88, 2024.

[Touati and Mignotte, 2018] Redha Touati and Max Mignotte. An energy-based model encoding nonlocal pairwise pixel interactions for multisensor change detection. *IEEE Trans. Geosci. Remote Sens.*, 56(2):1046–1058, 2018.

[Vaswani *et al.*, 2023] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023.

[Volpi *et al.*, 2015] Michele Volpi, Gustau Camps-Valls, and Devis Tuia. Spectral alignment of multi-temporal cross-sensor images with automated kernel canonical correlation analysis. *ISPRS J. Photogramm. Remote Sens.*, 107:50–63, 2015.

[Wan *et al.*, 2019a] Ling Wan, Yuming Xiang, and Hongjian You. An object-based hierarchical compound classification method for change detection in heterogeneous optical and sar images. *IEEE Trans. Geosci. Remote Sens.*, 57(12):9941–9959, 2019.

[Wan *et al.*, 2019b] Ling Wan, Yuming Xiang, and Hongjian You. A post-classification comparison method for sar and optical images change detection. *IEEE Geosci. Remote Sens. Lett.*, 16(7):1026–1030, 2019.

[Wu *et al.*, 2022] Yue Wu, Jiaheng Li, Yongzhe Yuan, A. K. Qin, Qi-Guang Miao, and Mao-Guo Gong. Commonality autoencoder: Learning common features for change detection from heterogeneous images. *IEEE Trans. Neural Networks Learn. Syst.*, 33(9):4257–4270, 2022.

[Yang *et al.*, 2024] Yueguang Yang, Jiahui Qu, Wenqian Dong, Tongzhen Zhang, Song Xiao, and Yunsong Li. Tmcfn: Text-supervised multidimensional contrastive fusion network for hyperspectral and lidar classification. *IEEE Trans. Geosci. Remote Sens.*, 62:1–15, 2024.

[Zhang and Xia, 2022] Lianchong Zhang and Junshi Xia. Flood detection using multiple chinese satellite datasets during 2020 china summer floods. *Remote Sens.*, 14(1), 2022.