ICSF: integrating inter-modal and cross-modal learning framework for self-supervised heterogeneous change detection.

ZHANG, E., ZONG, H., LI, X., FENG, M. and REN, J.

2025

ICSF: Integrating Inter-Modal and Cross-Modal Learning Framework for Self-Supervised Heterogeneous Change Detection

Erlei Zhang, *Member, IEEE*, He Zong, Xinyu Li, Mingchen Feng, and Jinchang Ren, *Senior Member, IEEE*

***Abstract*—Heterogeneous change detection (HCD) is a process to determine the change information by analyzing heterogeneous images of the same geographic location taken at different times, which plays an important role in remote sensing applications such as disaster response and environmental monitoring. However, the different imaging mechanisms result in different visual appearances in heterogeneous images, making it difficult to accurately detect changes through direct comparison. To address this problem, we propose a novel self-supervised dual-branch framework (ICSF) for HCD that incorporates inter-modal and cross-modal learning. First, in the inter-modal branch, we perform the contrastive learning on heterogeneous images within their respective modalities to learn the robust and discriminative features, rather than relying on the raw spectral or spatial information from these images. Second, in the cross-modal branch, we perform cross-modal reconstruction to ensure the obtained features exhibit consistent comparability, thereby facilitating the extraction of rich information of the real changes within the images. Next, the difference images (DIs) computed from both branches are further refined using a superpixel segmentation strategy to preserve the consistency of differences within the same ground object. Experimental results on five public datasets with different modality combinations and change events demonstrate the effectiveness of the proposed approach in comparison to ten state-of-the-art methods, achieving the best performance with an average overall accuracy of 95.88% and an average Kappa coefficient of 74.20%.**

***Index Terms*—Heterogeneous change detection (HCD), self-supervised learning, dual-branch, contrastive learning, structural relationship.**

I. Introduction

In recent years, with the advancement of remote sensing technology, the rapidly increasing remote sensing data with various modalities and resolutions provides a vast opportunity for Earth observation [1]. As a crucial task in the remote sensing community, remote sensing change detection (CD) is a process of identifying and analyzing changes that occur over time in images at the same geographical location [2]. This technology has been extensively applied in various fields, including damage assessment [3], disaster monitoring [4], and environmental monitoring [5].

With the advancement of deep learning, significant progress has been made in remote sensing applications [6], [7]. For example, Hong *et al.* [8] proposed a deep learning-based framework for multimodal remote sensing data classification, utilizing convolutional neural networks as the backbone. In [9], an effective coupling paradigm was proposed to address the performance bottleneck in hyperspectral anomaly detection, combining model-driven low-rank representation with data-driven deep learning techniques. For CD applications, the focus is primarily on homogeneous remote sensing images, where the pre- and post-event images are captured by the same sensors under identical parameters and are rigorously registered. This includes multispectral images (MSIs) [10], [11], [12], [13], synthetic aperture radar (SAR) images [14], [15], [16] and hyperspectral images (HSIs) [17] [18] [19]. It is worth noting that CD between corresponding regions is only meaningful if the two images are geographically aligned accurately [20], where effective image registration methods [21] [22] can be applied in the data pre-processing steps to ensure that the two images are well aligned. However, obtaining timely homogeneous images is challenging due to adverse weather and atmospheric conditions, especially in regions affected by natural disasters. Consequently, heterogeneous CD (HCD) using different types of remote sensing images has become a new focus and trend. Here, heterogeneous images can be categorized into two types: images captured from different sensors but employing the same image type (e.g., two MSIs with red, green and blue bands from the Pleiades and WorldView-2 in France dataset [23]) and images captured by different sensors in different image types (e.g., a pair of MSI and SAR images from the Landsat-8 and Radarsat-2 in Shuguang dataset [24]). In this paper, the combination of two common types of remote sensing images, MSI and SAR imagery, is focused for HCD, including (SAR, MSI) and (MSI, MSI). Here, (MSI, MSI) refers to MSIs captured from different sensors with potentially different spectral resolutions and spectral ranges but of the same image modality. Compared to the homogeneous CD, HCD offers better practicality and flexibility. The complementary nature of heterogeneous images with various imaging characteristics can potentially significantly increase the availability of remote sensing data.

This work was supported in part by the National Natural Science Foundation of China (No. 62376225), Chinese Universities Scientific Fund (No. 2452024392), the QinChuangyuan High-Level Innovation and Entrepreneurship Talent Program of Shaanxi (No. 2021QCYRC4-50). (Corresponding author: Jinchang Ren, jinchang.ren@ieee.org)

Erlei Zhang, He Zong, Xinyu Li, and Mingchen Feng are with the College of Information Engineering, Northwest A&F University, Yangling 712100, China.

Jinchang Ren is with the National Subsea Centre, Robert Gordon University, AB21 0BH Aberdeen, U.K.

By using any available pre- and post-event images, this can enable quick extraction of the change information and thus reduce the response time in case of emergency events, without being limited by the lack of homogeneous images.

HCD, where the pre-event and post-event images are captured from different sensors with different modalities, enables the utilization of any available images to efficiently extract changed objects. However, the changing visual appearances caused by the different imaging mechanisms in heterogeneous images render traditional methods [25], [26] inadequate for accurately detecting relevant changes. To address this challenge, HCD methods can be categorized into supervised, semi-supervised, and unsupervised, depending on whether the ground truth data is used in deriving the detection model. For supervised and semi-supervised methods, it is necessary to collect labeled data in advance; however, due to the complexity of the scenarios and the variations of objects, the collection of sufficient labeled data is costly, time-consuming, and labor-intensive [20]. In contrast, unsupervised HCD methods are more practical and challenging, which will be focused on in this paper.

The principle of unsupervised HCD is to transform incomparable heterogeneous images into a comparable domain space. Inspired by the classification strategy in [27] yet with certain refinements and extensions, we divide HCD methods into four categories based on the constructed domain space: i.e. similarity measure-based, classification-based, deep latent feature-based, and image translation-based. Among them, similarity measure-based methods utilize modality-independent structural relationships to distinguish whether changes occur in the corresponding regions. This type of method, such as [27]–[29], captures structural relationships by constructing non-local k-nearest neighbor (KNN) graphs. The strength of them lies in their simplicity and ease of implementation. However, there are two main challenges with them. On the one hand, they only use original spectral or low-level spatial information in heterogeneous images, which lacks robustness in complex detection conditions. For instance, it may struggle with diverse land-cover objects of varying sizes [30], or strong speckle noise in SAR images [15]. On the other hand, they detect changes solely by constructing similarity metrics based on structural relationships. However, the performance of the constructed similarity metrics depends on the complexity of the scenario. In some intricate cases, relying solely on these metrics may fail to distinguish between changed and unchanged regions, leading to the loss of certain change information and thus degrading the performance of CD.

To address the aforementioned challenges, we propose an integrating **I**nter-modal and **C**ross-modal **S**elf-supervised dual-branch learning **F**ramework (ICSF) for HCD. ICSF comprises inter-modal and cross-modal learning branches. Firstly, to address the limitations associated with using raw spectral and spatial information in image regions, inspired by the contrastive learning paradigm [R1], we establish Siamese networks in the inter-modal branch to extract valuable and robust features from these regions. This helps to construct more accurate structural relationships where similar regions are closer in the feature space. Secondly, a novel cross-modal branch is introduced to fully extract change information through cross-modal reconstruction for heterogeneous images. Furthermore, we employ a superpixel segmentation-based refinement strategy to enhance the quality of difference images (DIs) derived from both branches, better highlighting the degree of changes within the same ground object. The main contributions are summarized as follows.

(1) We present the first attempt to integrate inter- and cross-modal learning for unsupervised HCD. In the inter-modal branch, we first perform contrastive feature learning on heterogeneous images within their respective modalities. Efficient Siamese networks are established to learn robust and representative features while facilitating more accurate construction of KNN graphs.

(2) To better and more comprehensively extract change information, we establish a cross-modal learning branch. By performing cross-modal reconstruction, our cross-modal branch network enables their extracted features to be mapped into the same feature space respectively, where their features exhibit consistent comparability.

(3) An automated refinement strategy based on superpixel segmentation is proposed to highlight the degree of changes within the same ground object in the DIs.

(4) The impressive experimental results on five public datasets with different modality combinations and change events demonstrate the superiority and practicability of our proposed methods in comparison with ten unsupervised state-of-the-art (SOTA) HCD methods.

The rest of this paper is organized as follows. In Section II, we concisely introduce related works on unsupervised HCD methods and contrastive learning. Section III provides related background knowledge and details the proposed method. Sections IV and V present quantitative and qualitative experimental results and discussions, and finally, the conclusion of our work is given in Section VI.

II. Related Work

In this section, we provide a brief review of representative methods in each category of unsupervised HCD as mentioned in Section I, and introduce the relevant knowledge of contrastive learning.

A. Unsupervised HCD

1) Classification-based methods: They first classify pre- and post-event images to transform heterogeneous images into a common category domain, generating respective classification maps to detect changes. As Wan *et al.* [31] introduced a post-classification comparison method that integrated superpixel segmentation and classification for SAR and optical images CD. Building upon this approach, they further proposed a region-based multitemporal hierarchical Markov random field (RMH-MRF) model to enhance the performance of CD [32]. Han *et al.* [33] developed a hierarchical extreme learning machine to extract robust features from heterogeneous images, mitigating the impact of noise. Li *et al.* [34] presented a spatially self-paced convolutional network to efficiently select reliable samples and capture the relationships between heterogeneous images, thereby enhancing CD accuracy.

2) Deep latent feature-based methods: The core idea is to transform heterogeneous images into a high-dimensional feature space, where their features are continuous and can be directly compared. Leveraging the powerful feature extraction ability of deep learning, most feature transformation-based methods employ deep learning models to determine the latent feature space. In [24], an unsupervised symmetric convolutional coupling network (SCCN) was proposed, which transformed heterogeneous images into a shared feature space with consistent representations. Wu *et al.* [35] proposed a commonality autoencoder to discover common features of ground objects between heterogeneous images in the feature space. Liu *et al.* [36] presented a probabilistic model based on a bipartite convolution network, which learned to capture common distributions of heterogeneous images in an unsupervised manner. Xing *et al.* [37] proposed an iterative modality alignment approach in the feature space to reduce the influence of modality discrepancy and changed regions, thereby progressively improving detection accuracy. In [38], self-guided autoencoders (SGAEs) were initially established to generate an elementary change map as initial pseudo-labels. The pseudo-labels are then used iteratively to optimize the network, which helps extract the discriminative features in self-guided iterations.

3) Image translation-based methods: They aim to reduce modality differences by projecting the pre-event (or post-event) image from its modality to the modality of the post-event (or pre-event) image based on style transfer and adversarial learning. Niu *et al.* [39] utilized a conditional generative adversarial network (GAN) to convert optical images to images with similar statistical properties as SAR images. Li *et al.* [40] applied a cyclic GAN structure for modality translation, followed by training a CD network on the translated images to enhance performance. Luppino *et al.* [41] introduced an adversarial cyclic encoder network (ACE-Net) for modality translation, combining cycle consistency and adversarial learning. In [42], a code space alignment loss was introduced to mitigate the impact of change pixels and enhance image translation.

4) Similarity measure-based methods: These methods focus on the similarity measurement between heterogeneous images. By leveraging modality-independent structural relationships, they transform the heterogeneous images into specific metric space, enabling them to distinguish changed regions from unchanged ones. Luppino *et al.* [43] represented structural relationships by constructing local affinity matrices in different modalities and directly calculating the differences between affinity matrices. Mignotte *et al.* [44] started to explore the self-similarity property in heterogeneous images to construct structural relationships for HCD. They utilized fractal projection based on self-similarity to transform an image from its original modality to the modality of the given image, and the DI is then obtained by comparing the transferred image with the given image. In [28], [45], the structural relationships were represented by finding the KNN regions to construct nonlocal KNN graphs for image regions. The structure differences between heterogeneous images were then computed using graph mapping. Chen *et al.* [20] took advantage of two types of structural relationships in heterogeneous images to construct KNN graphs. Subsequently, they applied a graph convolutional autoencoder to extract robust features from these graphs. In [46], they treated each image as a graph signal defined on the corresponding constructed KNN graphs. Sun *et al.* [47] proposed an energy model based on image structural consistency, which can reduce the influence of image noise and varying imaging conditions.

It is worth noting that the self-similarity property can also enhance other methods like image regression [48]–[50], leading to improved CD accuracy.

B. Contrastive Learning

Contrastive learning, a technique for self-supervised representation learning, is widely applied in tasks without labeled data [51]–[53]. Its core idea is to construct positive and negative sample pairs from different views of the data, and then aggregate positive and separate negative sample pairs in the feature space [54]. This approach facilitates the acquisition of robust features and ensures that similar data is closer in the feature space, enhancing the discriminative capability of the features. The success of contrastive learning may depend on techniques like memory banks [55], momentum updates [56], projection heads [57], and stop gradient operations [58], [59]. BYOL [58] and Simsiam [59] directly remove negative samples, only requiring positive sample construction. Simsiam further simplifies BYOL by combining a Siamese network with a stop gradient operation, resulting in faster convergence. In this work, we employ Simsiam as our training strategy to extract rich features in heterogeneous images.

III. Methodology

In this section, we first introduce the preliminary knowledge about the problem statement regarding HCD and structural relationships, followed by a detailed depiction of our dual-branch self-supervised learning framework (ICSF). The overall diagram of ICSF is illustrated in Fig. 1, comprising the following five components: (1) data normalization and image patch generation; (2) inter-modal branch learning; (3) cross-modal branch learning; (4) change information computing; and (5) superpixel segmentation-based refinement.

A. Preliminaries

1) Problem Formulation: Let a pair of registered heterogeneous remote sensing images $X = \{x(h,w,c) \mid 1 \le h \le H, 1 \le w \le W, 1 \le c \le C_X\}$ with modality $\mathcal{X}$ and $Y = \{(y(h,w,c) \mid 1 \le h \le H, 1 \le w \le W, 1 \le c \le C_Y\}$ with modality $\mathcal{Y}$ be acquired over the same geographical area before and after a change event occurs, respectively. Here, H, W, and C_X (or C_Y) denote the height, weight, and number of channels of the image X (or Y). Both images are geometrically aligned and have the same size and spatial resolution. The purpose of HCD is to generate a binary change map $BM \in \mathbb{R}^{H \times W}$, where the changed pixels are labeled as "1" and those unchanged are labeled as "0". The CD step can be formulated as

$$BM = G(f_X(X) \ominus f_Y(Y)) \tag{1}$$

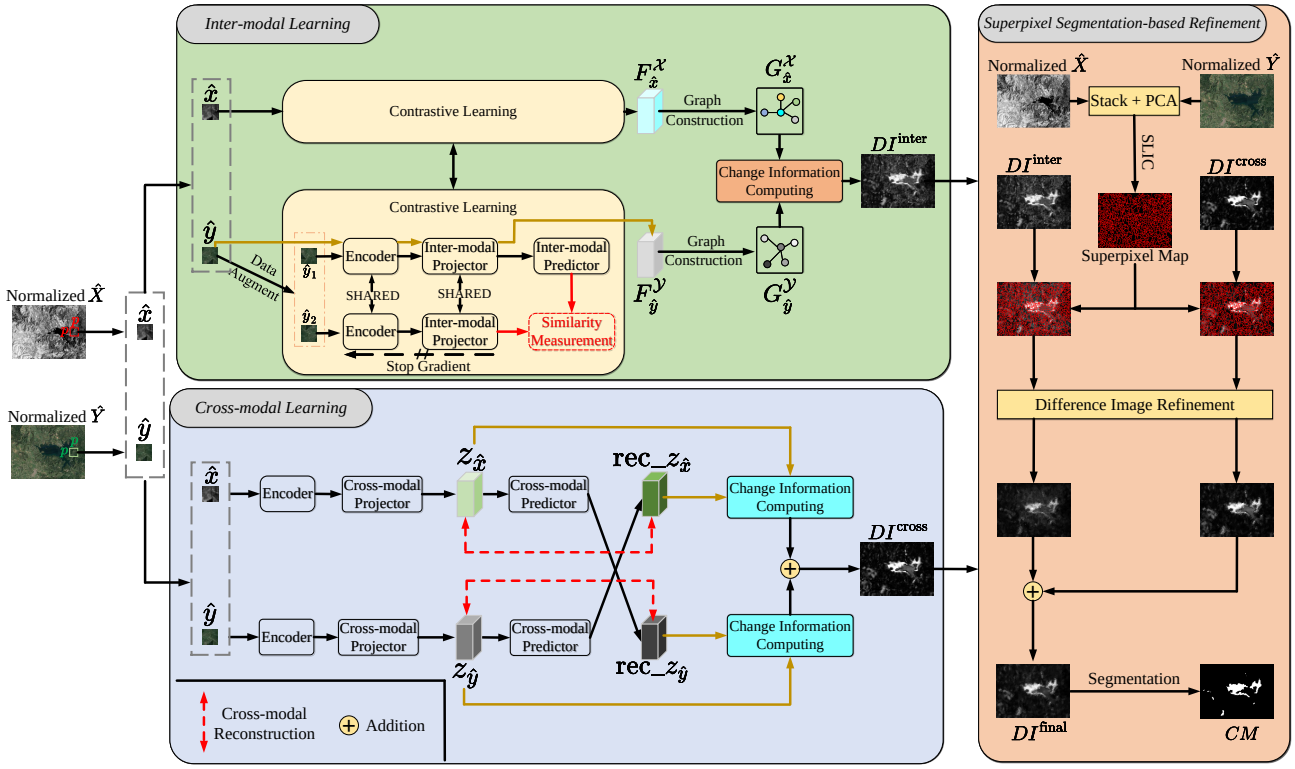

Fig. 1. The overall diagram of the proposed ICSF. First, from the inputs of $\hat{X}$ and $\hat{Y}$, image patches $\hat{x}$ and $\hat{y}$ are extracted, which then respectively undergo data augmentation to generate $\hat{x}_1$, $\hat{x}_2$, and $\hat{y}_1$, $\hat{y}_2$. In the inter-modal branch, $\hat{x}_1$ (or $\hat{y}_1$) and $\hat{x}_2$ (or $\hat{y}_2$) are taken as inputs to the Siamese network to learn robust features $F^{\mathcal{X}}_{\hat{x}}$ and $F^{\mathcal{Y}}_{\hat{y}}$. Additionally, these features capture the structural relationship through graph construction, resulting in $G^{\mathcal{X}}_{\hat{x}}$ and $G^{\mathcal{Y}}_{\hat{y}}$. In the cross-modal branch, $\hat{x}$ and $\hat{y}$ are encoded into the features $z_{\hat{x}}$ and $z_{\hat{y}}$, followed by cross-modal reconstruction to learn consistently comparable features. Finally, a superpixel segmentation-based refinement strategy is used to refine the computed DI^{inter} and DI^{cross}.

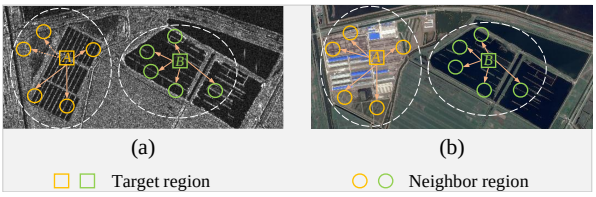

Fig. 2. Structural relationship in the heterogeneous images. (a) Pre-event image. (b) Post-event image. Here, the structure is defined by the similarity relationship between the target image regions and their neighbor regions. The solid lines indicate that the regions are similar to each other, and the dashed lines are opposite. Unchanged regions exhibit consistent structures, but changed regions do not.

where f_X (f_Y) represents the feature extraction operation, while $\ominus$ denotes the difference operator, aiming to generate a difference image (DI). In feature transformation-based methods, f_X (f_Y) is typically considered as feature extractors used to align X and Y into a common feature space [24], [35], [36]. In similarity measure-based methods, f_X (f_Y) can be used to represent modality-independent structural relationships [45], [27], [29]. Hereafter, a DI that shows the change degree can be analyzed by G to generate the final change map BM, which assigns a label to each pixel position and accurately identifies changes that occurred on the ground.

2) Structural Relationship: Due to the significant difference across different modalities, the same geographic targets present varying visual characteristics, rendering it impracticable to achieve accurate CD through direct comparison of heterogeneous images. Nonetheless, structural relationships can be leveraged for HCD.

As depicted in Fig. 2, for target region A (or B) in the pre-event image, K similar regions can be found to form a KNN graph. As the region B is unchanged after the event, the graph structure formed by B and its similar regions exhibits minimal changes in the post-event image. However, for A, the corresponding graph structure can no longer be maintained because the changes have occurred. We refer to this as the structural relationship, which is established by constructing a KNN graph for the image regions.

B. Data Normalization and Image Patch Generation

Given a pair of co-registered heterogeneous images, denoted as the pre-event image X with modality $\mathcal{X}$ and the post-event image Y with modality $\mathcal{Y}$, capturing the same geographical area at different times t_1 and t_2. The pixels in X and Y are denoted as $x(h,w,c)$ and $y(h,w,c)$, with c corresponding to the channel dimension.

Due to the different imaging mechanisms of heterogeneous images, the range of their pixel values is different. To address this problem, we first perform image normalization to scale their pixel values to the same range. It is helpful for the subsequent training of our proposed network. The heterogeneous images used in this work include near-infrared (NIR), MSI, and SAR images. Following [20], we consider all image types

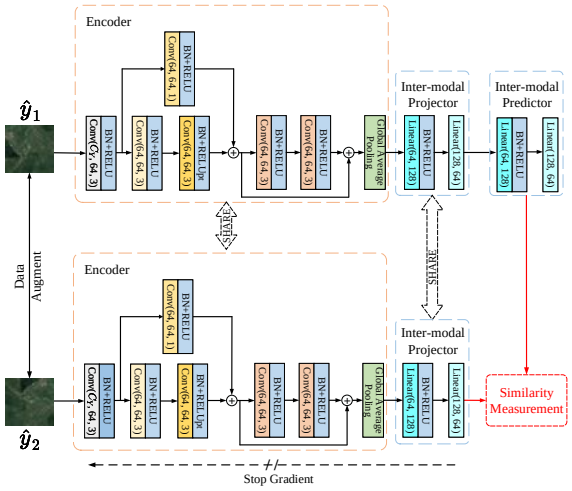

Fig. 3. Architecture of the proposed Self-Supervised Contrastive Learning Network (SCLN). $\hat{y_1}$ and $\hat{y_2}$ denote the image patches. The "Stop Gradient" arrow indicates that the features from the lower branch serve as the targets for the upper branch, and gradient backpropagation does not occur in the lower branch during the learning process.

except SAR as optical images for the same normalization. For optical images, we normalize their pixel values to the range of [0, 1] by

$$\hat{x}(h,w,c) = \frac{x(h,w,c) - \min_x}{\max_x - \min_x} \tag{2}$$

where $\max_x$ and $\min_x$ are the maximum and minimum values of the image pixels across all channels.

For SAR images, we first perform a logarithmic ratio to suppress the impact of speckle noise [14], followed by a similar normalization as in Eq. 2

$$\begin{cases} x_{\log}(h,w,c) = \log(1 + x(h,w,c)) \\ \hat{x}(h,w,c) = \frac{x_{\log}(h,w,c) - \min_{x_{\log}}}{\max_{x_{\log}} - \min_{x_{\log}}}. \end{cases} \tag{3}$$

The normalized heterogeneous images are denoted as $\hat{X}$ and $\hat{Y}$.

Next, we extract image patches from the normalized heterogeneous images $\hat{X}$ and $\hat{Y}$ using an overlapping sliding window. In this procedure, the image patch size is set to p, while the step size of the sliding window is set to $\lceil p/2 \rceil$. Corresponding patches $\hat{x}^i$ and $\hat{y}^i$ are extracted as training samples. Here, $i \in \{1, \ldots, |N|\}$, where N represents the number of patches.

C. Dual-branch Learning

Through the aforementioned steps, we have obtained the normalized image patches as our training samples. Here, assuming a pair of image patches $\hat{x}$ and $\hat{y}$ are extracted from $\hat{X}$ and $\hat{Y}$, respectively.

1) Inter-Modal Learning: In the inter-modal branch, we establish the self-supervised contrastive learning networks (SCLNs) to extract features from $\hat{X}$ and $\hat{Y}$, respectively. Here, a contrastive learning strategy called Simsiam [59] is applied for SCLN which presents an implicit contrastive learning way without the requirements to have negative samples during the network training. This approach allows us to extract

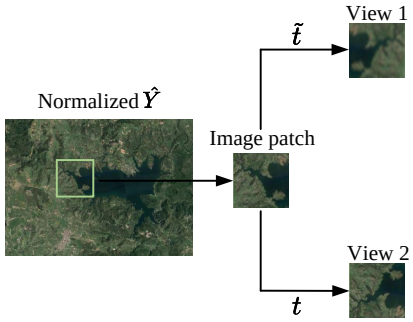

Fig. 4. Different views of an image patch obtained through data augmentation, where $t(\cdot)$ and $\tilde{t}(\cdot)$ represent data augmentation operations.

valuable features from image patches and ensures that similar patches are closer in feature space, enhancing the accuracy of the constructed KNN graphs. The network structure of the proposed SCLN is depicted in Fig. 3, which consists of a Siamese network with two branches. The upper branch is the online network that is composed of one encoder, one inter-modal projector, and one inter-modal predictor, while the lower branch is the target network that has the same network architecture without the inter-modal predictor module.

Following the mainstream contrastive learning paradigm, we use the random data augmentation method to obtain different views of the image patches. Specifically, we apply different data augmentation $t(\cdot)$ and $\tilde{t}(\cdot)$ for $\hat{y}$ to obtain two different views: $\hat{y_1} = t(\hat{y})$ and $\hat{y_2} = \tilde{t}(\hat{y})$, which are then regarded as a positive sample pair. We can perform the same operation on $\hat{x}$ to obtain $\hat{x_1}$ and $\hat{x_2}$. We follow the reference augmentations in [60], including random horizontal flipping, random vertical flipping, and random Gaussian blur. To illustrate the effect of data augmentation, visual examples of different views of an image patch are given in Fig. 4.

As shown in Fig. 3, $\hat{y_1}$ is input into the upper branch, obtaining feature vectors $p_{\hat{y_1}}$, and $\hat{y_2}$ is processed by the lower branch, generating $z_{\hat{y_2}}$. Symmetrically, we can obtain $p_{\hat{y_2}}$ and $z_{\hat{y_1}}$ by exchanging the input positions of $\hat{y_1}$ and $\hat{y_2}$. Here, we optimize the network associated with $\hat{Y}$ by minimizing the loss $\mathcal{L}_{\text{inter\_}\hat{y}}$, which is defined as

$$\mathcal{L}_{\text{inter\_}\hat{y}} = -\frac{1}{2}(\text{sim}(p_{\hat{y_1}}, z_{\hat{y_2}}) + \text{sim}(p_{\hat{y_2}}, z_{\hat{y_1}})) \tag{4}$$

where $\text{sim}(\cdot,\cdot)$ is the cosine similarity function, which is used to measure the feature similarity.

Similarly, we perform the same operation on $\hat{x}$ to calculate the loss function by

$$\mathcal{L}_{\text{inter\_}\hat{x}} = -\frac{1}{2}(\text{sim}(p_{\hat{x_1}}, z_{\hat{x_2}}) + \text{sim}(p_{\hat{x_2}}, z_{\hat{x_1}})) \tag{5}$$

Then, the loss function of the inter-modal branch can be written as

$$\mathcal{L}_{\text{inter}} = \frac{1}{2}(\mathcal{L}_{\text{inter\_}\hat{x}} + \mathcal{L}_{\text{inter\_}\hat{y}}) \tag{6}$$

It is worth noting that during the optimization process, the parameters of the target network are frozen, which is crucial to ensure stable training [59]. Once the parameters of the online network are updated, the parameters of the encoder and the projector are copied to the target network.

2) Cross-Modal Branch Learning: Some existing methods construct similarity metrics by establishing structural relationships within the image, which are modality-independent, and detect changes by computing these similarity metrics [20], [27]–[29]. However, relying solely on this approach may lead to the loss of change information, particularly for regions with subtle structural changes. To this end, we introduce a cross-modal branch in our learning framework, as shown in Fig. 1. In the cross-modal branch, we establish a cross-modal reconstruction network (CMRN). Similar to SCLN, it comprises an encoder, cross-modal projector, and cross-modal predictor modules. The network structure of the cross-modal projector and cross-modal predictor modules is the same as the inter-modal projector and inter-modal predictor modules. However, due to the different modalities of the input images, CMRN is not designed as a Siamese network with shared weights.

First, we can obtain the feature representations $z_{\hat{x}}$ and $z_{\hat{y}}$ for $\hat{x}$ and $\hat{y}$ by inputting them into the encoder and cross-modal projector modules of CMRN. To further explore the common features between the two inputs from different modalities, we propose cross-modal reconstruction to transform the feature representation $z_{\hat{x}}$ into the other and vice versa. Then, $z_{\hat{x}}$ and $z_{\hat{y}}$ are input into the cross-modal predictor module to predict the corresponding feature representations respectively, obtaining rec_$z_{\hat{y}}$ and rec_$z_{\hat{x}}$. Finally, the loss function of the cross-branch can be defined as

$$\mathcal{L}_{\text{cross}} = \frac{1}{2}(\|z_{\hat{x}} - \text{rec\_}z_{\hat{x}}\|_2^2 + \|z_{\hat{y}} - \text{rec\_}z_{\hat{y}}\|_2^2) \tag{7}$$

where $||\cdot||_2^2$ refers to the squared Euclidean distance. Specifically, in order to reduce the complexity of the proposed ICSF, we share the encoder parameters between CMRN and SSLN, as they process the same type of images.

The final loss function can be written as

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{inter}} + \mathcal{L}_{\text{cross}} \tag{8}$$

It can be seen that both the inter-modal and cross-modal branches contribute to the $\mathcal{L}_{\text{total}}$. The proposed ICSF can be trained by minimizing the loss function $\mathcal{L}_{\text{total}}$.

D. Change Information Computing

After training the proposed ICSF, all the image patches $\{\hat{x}^1,\ldots,\hat{x}^{|N|}\}$ and $\{\hat{y}^1,\ldots,\hat{y}^{|N|}\}$ from $\hat{X}$ and $\hat{Y}$ can be fed into the network to obtain their respective feature representations. Assuming we have a pair of image patches $\hat{x}^i$ and $\hat{y}^i$ sampled from all image patches. Here, the region denoted by Ω_i represents the same spatial location in two image patches: $\hat{x}^i$ and $\hat{y}^i$.

In the inter-modal branch, $\hat{x}^i$ and $\hat{y}^i$ are fed into the online network of corresponding SCLN, thus obtaining the feature representations $F_{\hat{x}^i}$ and $F_{\hat{y}^i}$. If a change event occurs within the region Ω_i, the relationships between Ω_i and its similar regions will no longer be consistent across bi-temporal heterogeneous images. Accordingly, we construct a KNN graph for $\hat{x}^i$ by finding its K most similar patches based on their feature similarity. The graph structure is then represented as follows:

$$\begin{cases} G_{\hat{x}^i}^{\mathcal{X}} = \{\mathcal{V}_{\hat{x}^i}^{\mathcal{X}}, \mathcal{E}_{\hat{x}^i}^{\mathcal{X}}, W_{\hat{x}^i}^{\mathcal{X}}\} \\ \mathcal{V}_{\hat{x}^i}^{\mathcal{X}} = \{F_{\hat{x}^k}^{\mathcal{X}}, k = 1,2,\ldots,K\} \\ \mathcal{E}_{\hat{x}^i}^{\mathcal{X}} = \{(F_{\hat{x}^i}^{\mathcal{X}}, F_{\hat{x}^k}^{\mathcal{X}}) \mid F_{\hat{x}^k}^{\mathcal{X}} \in \mathcal{V}_{\hat{x}^i}^{\mathcal{X}}\} \\ W_{\hat{x}^i}^{\mathcal{X}} = \{||F_{\hat{x}^i}^{\mathcal{X}} - F_{\hat{x}^k}^{\mathcal{X}}||_2^2 \mid (F_{\hat{x}^i}^{\mathcal{X}}, F_{\hat{x}^k}^{\mathcal{X}}) \in \mathcal{E}_{\hat{x}^i}^{\mathcal{X}}\} \end{cases} \tag{9}$$

where the kth patch that is most similar to $\hat{x}^i$ is represented by $\hat{x}^k$, and $||\cdot||_2^2$ refers to the squared Euclidean distance. The KNN graph constructed for $\hat{x}^i$ is denoted as $G_{\hat{x}^i}^{\mathcal{X}}$. $V_{\hat{x}^i}^{\mathcal{X}}$ represents the set of vertices, $E_{\hat{x}^i}^{\mathcal{X}}$ represents the edge between $F_{\hat{x}^i}^{\mathcal{X}}$ and $F_{\hat{x}^k}^{\mathcal{X}}$, and $W_{\hat{x}^i}^{\mathcal{X}}$ represents the weight assigned to the edge. Similarly, we can construct a graph $G_{\hat{y}^i}^{\mathcal{Y}} = \{\mathcal{V}_{\hat{y}^i}^{\mathcal{Y}}, \mathcal{E}_{\hat{y}^i}^{\mathcal{Y}}, W_{\hat{y}^i}^{\mathcal{Y}}\}$ for $\hat{y}^i$. Here, following the approach in [27], we set an adaptive $K = \lceil(\sqrt{N} + \frac{\sqrt{N}}{10})/2\rceil$, without the need for careful tuning of the K value.

Since the two constructed graph structures come from different modalities, we do not directly compare them to obtain the change information. Instead, we map each graph structure to the modality of the other graph, further reducing the impact of modality differences. We construct a mapping KNN graph $G_{\hat{x}^i}^{\mathcal{Y}} = \{\mathcal{V}_{\hat{x}^i}^{\mathcal{Y}}, \mathcal{E}_{\hat{x}^i}^{\mathcal{Y}}, W_{\hat{x}^i}^{\mathcal{Y}}\}$ in the post-event image Y. This graph is constructed using the spatial coordinates of K patches that are most similar to $\hat{x}^i$. Specifically, we calculate the differences of $(G_{\hat{x}^i}^{\mathcal{Y}}, G_{\hat{y}^i}^{\mathcal{Y}})$ to determine whether changes have occurred within the region Ω_i

$$d_{\Omega_i}^{\mathcal{X}} = \frac{1}{K}\sum_{k=1}^{K}(||F_{\hat{x}^i}^{\mathcal{Y}} - F_{\hat{x}^k}^{\mathcal{Y}}||_2^2 - ||F_{\hat{y}^i}^{\mathcal{Y}} - F_{\hat{y}^k}^{\mathcal{Y}}||_2^2) \tag{10}$$

where $||F_{\hat{x}^i}^{\mathcal{Y}} - F_{\hat{x}^k}^{\mathcal{Y}}||_2^2$ represents the weight of the kth edge in $G_{\hat{x}^i}^{\mathcal{Y}}$. Similar to $d_{\Omega_i}^{\mathcal{X}}$, we can obtain $d_{\Omega_i}^{\mathcal{Y}}$. Additionally, we can get inter-modal branch DI as $DI^{\text{inter}} \in \mathbb{R}^{H\times W}$ by assigning $d_{\Omega_i}^{\mathcal{X}}$ and $d_{\Omega_i}^{\mathcal{Y}}$ to the specific pixels according to Ω, which is defined as

$$DI^{\text{inter}}(h,w) = d_{\Omega_i}^{\mathcal{X}} + d_{\Omega_i}^{\mathcal{Y}} \tag{11}$$

where $(h,w) \in \Omega_i, i = 1,2,\ldots,|N|$.

In the cross-modal branch, we can obtain $z_{\hat{x}^i}$ and $p_{\hat{x}^i}$ by inputting $\hat{x}^i$ to the cross-modal projector and cross-modal predictor modules of CMRN. Similarly, we can acquire $z_{\hat{y}^i}$ and $p_{\hat{y}^i}$ from $\hat{y}^i$. Here, we calculate the distance of $(z_{\hat{x}^i}, p_{\hat{y}^i})$ and $(z_{\hat{y}^i}, p_{\hat{x}^i})$ to detect changes in region Ω_i as

$$d_{\Omega_i}^{\text{cross}} = (\|z_{\hat{x}^i} - p_{\hat{y}^i}\|_2^2 + \|z_{\hat{y}^i} - p_{\hat{x}^i}\|_2^2) \tag{12}$$

Take $(z_{\hat{x}^i}, p_{\hat{y}^i})$ as an example, the smaller distance between them indicates that they have more commonalities and a stronger correlation, and are less likely to have changed. Conversely, a larger distance shows they are less relevant, and thus more likely to be changed. The cross-modal branch DI^{cross} can be denoted as

$$DI^{\text{cross}}(h,w) = d_{\Omega_i}^{\text{cross}} \tag{13}$$

where $(h,w) \in \Omega_i, i = 1,2,\ldots,|N|$.

Algorithm 1 ICSF for Change Detection on Heterogeneous Images

Input: Pre-event image $X \in \mathbb{R}^{H\times W\times C_X}$, post-event image $Y \in \mathbb{R}^{H\times W\times C_Y}$, number of epochs $K \in \mathbb{N}^+$

Output: Difference image $DI^{\text{final}} \in \mathbb{R}^{H\times W}$ and change map $CM \in \{0,1\}^{H\times W}$

Normalize X and Y to obtain $\hat{X}$ and $\hat{Y}$ using Eq. (2) and Eq. (3)
Extract overlapping patches $\hat{x}^i \in \mathbb{R}^{p\times p\times C_X}$ and $\hat{y}^i \in \mathbb{R}^{p\times p\times C_Y}$ from $\hat{X}$ and $\hat{Y}$
Initialize network parameters ϑ randomly
for $k = 1$ to K **do**
    Compute inter-modal branch loss $\mathcal{L}_{\text{inter}}$ using Eq. (6)
    Compute cross-modal branch loss $\mathcal{L}_{\text{cross}}$ using Eq. (7)
    $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{inter}} + \mathcal{L}_{\text{cross}}$
    Update network parameters ϑ by minimizing $\mathcal{L}_{\text{total}}$
end for
Compute the difference image DI^{inter} and DI^{cross}.
Refine and Fuse DI^{inter} and DI^{cross} to get DI^{final} through Eq. (14) and Eq. (17).
Perform Otsu algorithm to get the change map CM:

$$CM = \text{Otsu}(DI^{\text{final}})$$

return DI^{final} and CM

E. Superpixel Segmentation-based Refinement

After generating DI^{inter} and DI^{cross}, they can be refined to enhance the detection results. Similar to other methods [28], [29], [45], our approach uses image patches as the fundamental unit. However, the distribution of land cover objects is often not square. An image patch may contain different types of objects, which are assigned the same change degree, leading to a decrease in the quality of DI. This will be discussed in Section V-B. To ensure that the same land cover objects display the same change degree and further obtain high-quality DIs, we propose the superpixel segmentation-based refinement strategy. Inspired by [10], we concatenate $\hat{X}$ and $\hat{Y}$ along the channel dimension and perform the principal component analysis (PCA), retaining only the first three principal components. We then apply the simple linear iterative clustering (SLIC) algorithm [61] for segmentation, which ensures consistent superpixel segmentation results for $\hat{X}$ and $\hat{Y}$. The cosegmentation superpixel result is defined as

$$\begin{cases} \Gamma = \{\Gamma_i \mid i = 1,2,\ldots,N_{\text{cs}}\} \\ \Gamma_i \cap \Gamma_j = \varnothing, \quad \text{if } i \neq j \\ \bigcup_{i=1}^{N_{\text{cs}}} \Gamma_i = \{(m,n) \mid m = 1,2,\ldots,H; n = 1,2,\ldots,W\} \end{cases} \tag{14}$$

where N_{cs} is the number of superpixels and H, W denote the height and width of $\hat{X}$ and $\hat{Y}$.

Denote the ith superpixel of DI^{inter} and DI^{cross} as $DI_i^{\text{inter}} = \{d(h,w) \mid (h,w) \in \Gamma_i\}$ and $DI_i^{\text{cross}} = \{d(h,w) \mid (h,w) \in \Gamma_i\}$, respectively. Within each superpixel, the land cover objects should exhibit the same degree of change. Therefore, we compute the mean change degree of all pixels within the superpixel as the superpixel's change degree:

$$DI_i^{\text{inter}}(h,w) = \text{mean}(DI_i^{\text{inter}}) \tag{15}$$

$$DI_i^{\text{cross}}(h,w) = \text{mean}(DI_i^{\text{cross}}) \tag{16}$$

where $\text{mean}(\cdot)$ represents the operation of calculating the mean value of a set, and $i = 1,2,\ldots,|N_{\text{cs}}|$.

Finally, the refined DI^{inter} and DI^{cross} can be obtained. We present a simple fusion strategy to fuse them

$$DI^{final} = (DI^{\text{inter}}/max(DI^{\text{inter}}) + DI^{\text{cross}}/max(DI^{\text{cross}}))/2 \tag{17}$$

where DI^{final} is the final DI.

During the DI analysis stage, the CD task can be treated as an image segmentation problem, which can be solved by using threshold segmentation [62], [63] or clustering [64], [65] methods. Here, we directly use a simple thresholding-based segmentation algorithm named Otsu [62] to the DI^{final}, thereby generating a change map CM that accurately reflects the observed changes on the land surface.

In summary, the entire CD procedure in ICSF is summarized in Algorithm 1.

IV. Results

In this section, five public heterogeneous datasets with different modality combinations and change events are first introduced, followed by the depictions of evaluation metrics and comparison methods. Next, the implementation details of our method are presented. Finally, comprehensive quantitative and qualitative experimental results of the proposed method are conducted.

A. Datasets

To evaluate the effectiveness of our proposed method, we conduct experiments on five public heterogeneous datasets. The detailed information of these five datasets is summarized in Table I.

The first dataset, Italy [23], consists of a near-infrared (NIR) image and a three-band MSI with a size of 300 × 412. These images were captured in 1995 and 1996 using Landsat-5 and Google Earth sensors, with a spatial resolution of 30m. The first row of Fig. 5(a)-(c) displays the images along with the corresponding ground truth, illustrating changes due to lake expansion.

The second dataset is the Tianhe dataset [66], which is illustrated in the second row of Fig. 5(a)-(c). It consists of a panchromatic (PAN) image and an MSI, which were captured from Landsat-7 in July 2002 and Google Earth in June 2013, respectively. Both images have a size of 666 × 615 pixels and have an approximate spatial resolution of 11m. The change event depicts the transformation of the Tianhe International Airport located in Wuhan, China.

The third dataset, known as the Shuguang dataset [24], includes a SAR image and an MSI collected in 2008 and 2012, respectively, covering a village in Shandong province, China. The third row of Fig. 5(a)-(c) shows the images and their ground truth, depicting changes related to building

TABLE I
INFORMATION OF THE FIVE HETEROGENEOUS DATASETS.

Datasets	Sensor	Size	Location	Dates	Event (& Spatial Resolution)
Italy	Landsat-5/Google Earth	$300 \times 412 \times 1(3)$	Sardinia, Italy	Sep. 1995/Jul. 1996	Lake expansion (30m)
Tianhe	Landsat-7/Google Earth	$615 \times 666 \times 1(3)$	Wuhan, China	Jul. 2002/Jun. 2013	Airport construction (≈11m)
Shuguang	Radarsat-2/Google Earth	$593 \times 921 \times 1(3)$	Shuguang Village, China	Jun. 2008/Sep. 2012	Building construction and river expansion (8m)
France	Pleiades/WorldView2	$2000 \times 2000 \times 3(3)$	Toulouse, France	May. 2012/Jul. 2013	Construction (0.52m)
Texas	Landsat-5/EO-1 ALI	$1534 \times 808 \times 7(10)$	Texas, USA	Sep. 2011/Oct. 2011	Wildfire (30m)

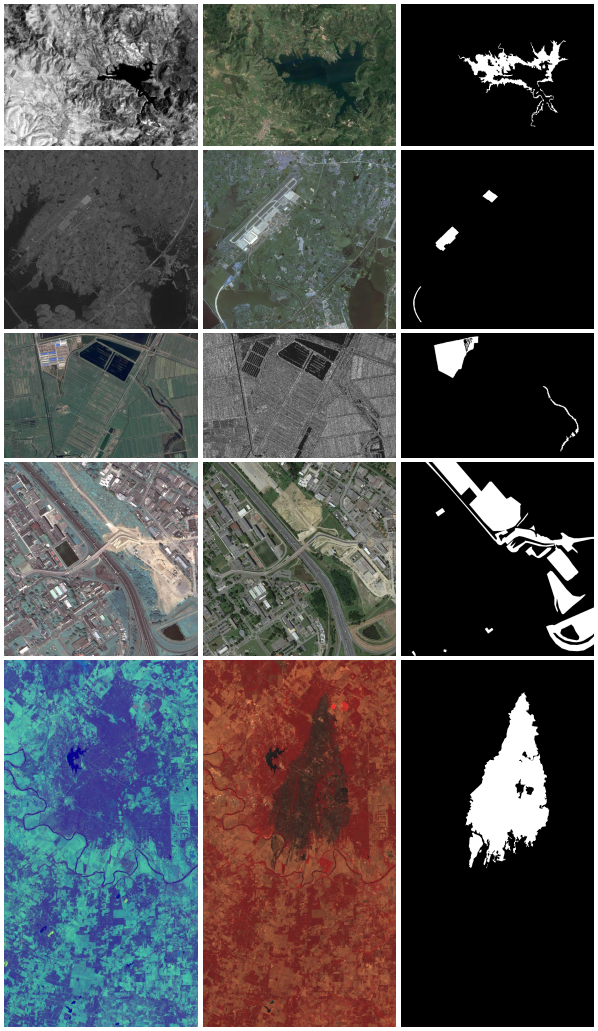

Fig. 5. Heterogeneous datasets. From top to bottom, they correspond to the Italy, Tianhe, Shuguang, France, and Texas datasets, respectively. (a) Pre-event image X captured at t_1; (b) Post-event image Y captured at t_2. (c) Ground truth;

construction and river expansion. Each image has a size of 593 × 921 pixels and a spatial resolution of 8m.

The fourth dataset is the France dataset [23], depicted in the fourth row of Fig. 5(a)-(c). These images were captured by Pleiades and WorldView 2, showcasing a road construction event in Toulouse, France, spanning from May 2012 to July 2013. Both images have a size of 2000 × 2000 pixels with a spatial resolution of 0.52m.

The final dataset [43] consists of two MSIs, depicted in the fifth row of Fig. 5(a)-(c). The pre-event image was captured by Landsat-5 in September 2011 with seven bands, while the post-event image was captured by EO-1 ALI in October 2011 with ten bands. Both images are 1534 × 808 in size with a spatial resolution of 30m. The change event captured is a wildfire that occurred in a forest area in Texas, USA.

B. Comparison Methods and Evaluation Metrics

To qualitatively assess the effectiveness and practicality of our method, we compare it with the SOTA approaches on five public heterogeneous datasets. Here, we choose ten unsupervised HCD methods (whose codes are publicly released) as our comparison algorithms, including Markov model based on a neighborhood adaptive class conditional likelihood (CCLMRF[1]) [67], X-Net[2] [41], ACE-Net[2] [41], nonlocal patch similarity graph-based method (NPSG[3]) [28], structure consistency based graph method (INLPG[4]) [45], iterative robust graph and Markovian cosegmentation method (IRG-MCS[5]) [27], structural relationship graph convolutional autoencoder (SRGCAE[6]) [20], vertex domain filtering (VDF-HCD[7]) [46], sparse constrained adaptive structure consistency based method (SCASC[8]) [48] and adaptive graph and structure cycle consistency-based method (AGSCC[9]) [49]. It is worth noting that the above methods are implemented with the recommended parameters described in their works. To quantitatively validate the performance of different CD methods, we employ two types of evaluation indices. Firstly, we utilize the receiver operating characteristic (ROC) curve and precision-recall (PR) curve to assess the quality of the DIs produced by different methods, with the areas under the ROC curve (AUC) and PR curve (AP) used as the quantitative criteria. The closer the ROC curve is to the upper left corner and the PR curve to the upper right corner, the higher the AUC and AP values, indicating better DI quality. Secondly, for the change maps generated by different methods, we employ six common evaluation metrics: false positives (FP), false negatives (FN), overall error (OE), overall accuracy (OA), F1 score (F1) as

[1]CCLMRF is available at https://www-labs.iro.umontreal.ca/~mignotte/
[2]X-Net, ACE-Net are available at https://github.com/llu025/Heterogeneous_CD
[3]NPSG is kindly available at https://github.com/yulisun/NPSG
[4]INLPG is kindly available at https://github.com/yulisun/INLPG
[5]IRG-MCS is kindly available at https://github.com/yulisun/IRG-McS
[6]SRGCAE is available at https://github.com/ChenHongruixuan/SRGCAE
[7]VDF-HCD is available at https://github.com/yulisun/HCD-GSP
[8]SCASC is available at https://github.com/yulisun/SCASC
[9]AGSCC is available at https://github.com/yulisun/AGSCC

TABLE II
THE VALUES OF AUC AND AP OF DIS GENERATED BY DIFFERENT METHODS ON THE FIVE DATASETS. THE BEST RESULTS AND THE SECOND-BEST RESULTS ARE IN RED AND BLUE, RESPECTIVELY.

Dataset		CCLMRF	X-Net	ACE-Net	NPSG	INLPG	IRG-MCS	SRGCAE	VDF-HCD	SCASC	AGSCC	ICSF
Italy	AUC	0.8808	0.9148	0.8596	0.9373	0.9485	0.8898	0.7863	0.8990	0.8915	0.9053	0.9652
	AP	0.3157	0.6446	0.4526	0.6240	0.7040	0.6472	0.2214	0.6089	0.4387	0.5319	0.8073
Tianhe	AUC	0.9866	0.9825	0.9650	0.9980	0.9938	0.9856	0.9896	0.9920	0.9900	0.9727	0.9957
	AP	0.6310	0.4063	0.2212	0.8550	0.7817	0.4110	0.6138	0.7887	0.5386	0.7169	0.8417
Shuguang	AUC	0.9430	0.9711	0.9600	0.9836	0.9826	0.9766	0.8828	0.9716	0.9642	0.9603	0.9877
	AP	0.6356	0.7465	0.6459	0.7011	0.7927	0.7747	0.7514	0.8168	0.6605	0.7993	0.8573
France	AUC	0.5745	0.8543	0.7593	0.6596	0.8019	0.8347	0.8185	0.8134	0.8139	0.8199	0.9287
	AP	0.2187	0.5571	0.4257	0.3437	0.5031	0.4346	0.5656	0.4680	0.5092	0.5167	0.6892
Texas	AUC	0.8276	0.9495	0.9804	0.9434	0.9710	0.9369	0.9564	0.9498	0.9602	0.9665	0.9927
	AP	0.4363	0.7177	0.8319	0.5058	0.7291	0.5621	0.8261	0.6829	0.6484	0.8305	0.9450
Average	AUC	0.8425	0.9344	0.9047	0.9044	0.9396	0.9247	0.8867	0.9252	0.9240	0.9249	0.9740
	AP	0.4475	0.6144	0.5155	0.6059	0.7021	0.5659	0.5957	0.6731	0.5591	0.6791	0.8281

well as the Kappa coefficient (KC).

$$\mathrm{OE} = \mathrm{FP} + \mathrm{FN} \tag{18}$$

$$\mathrm{OA} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \tag{19}$$

$$\mathrm{F1} = \frac{2\mathrm{TP}}{2\mathrm{TP} + \mathrm{FP} + \mathrm{FN}} \tag{20}$$

$$\mathrm{KC} = \frac{\mathrm{Po} - \mathrm{Pe}}{1 - \mathrm{Pe}}, (\mathrm{Po} = \mathrm{OA}, \tag{21}$$

$$\mathrm{Pe} = \frac{(\mathrm{TP} + \mathrm{FP})(\mathrm{TP} + \mathrm{FN}) + (\mathrm{FN} + \mathrm{TN})(\mathrm{FP} + \mathrm{TN})}{(\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN})^2}). \tag{22}$$

where TP and TN represent the values of true positives and true negatives, respectively. The lower the values of FP, FN, and OE, the higher the values of OA, F1, and KC, and the better the performance of one method.

C. Implementation Details

The proposed method is implemented with the Pytorch library. We employ the SGD optimizer with a momentum of $9e^{-1}$, weight decay of $1e^{-5}$, and a learning rate of $5e^{-2}$ to optimize the network. In the training process, the number of epochs is 100, with a batch size of 2048. For all datasets, the image patch size p is uniformly set to 9, and the superpixel number $N_{\mathrm{cs}} \approx 5000$ in the superpixel segmentation-based refinement. Here, we divide the image patches into the training and test sets in an 8:2 ratio. Additionally, the experiments are conducted on a personal computer with an Intel Core i9-13000K CPU running at 3.00 GHz, NVIDIA GeForce RTX 3090 GPU, 16.00 GB RAM, and Ubuntu 22.04 LTS 64-bit OS.

D. Experimental Results and Analysis

1) Difference Images: The DIs generated by the proposed ICSF method and ten comparison methods over the five different datasets are illustrated in Fig. 6, along with their corresponding ROC and PR curves shown in Fig. 7. Additionally, the AUC values of ROC curves and the AP values of PR curves for different methods are listed in Table II.

As shown in Fig. 6, the DIs generated by most methods can reflect some change information. However, due to the complexity of scenarios and the variety of ground objects, the quality of the generated DIs is uneven. For some simple scenarios (e.g., the Sardinia dataset), most methods can highlight the regions that are most likely to have changed. For some complex scenarios, like the France dataset, the quality of DIs generated by similarity measure-based methods (e.g., NPSG [28], INLPG [45], IRG-MCS [27]) only utilize the low-level information is poor, as they fail to accurately highlight the true changes. In contrast, our method utilizes contrastive learning to extract robust features for change information computing, which can better identify the changed regions and suppress the unchanged regions. On the Texas dataset, it can seen that our method can better capture change information with brighter colors compared to other methods. much changed information is not captured by SRGCAE and VDF-HCD with darker colors, which only compute constructed similarity metrics. It demonstrates that the proposed dual-branch learning can better capture more diverse change information. This is further supported by the results in Fig. 7 and Table II, which show that our proposed ICSF has achieved nearly the best performance compared to other comparison methods. For example, the ROC curves reach close to the top-left corner, with AUC values up to 0.9652, 0.9957, 0.9877, 0.9287, and 0.9927 on the Italy, Tianhe, Shuguang, France, and Texas datasets, respectively. Additionally, ICSF achieves average AUC and AP values of 0.9740 and 0.8281, respectively, outperforming all other methods. Notably, when compared to self-supervised HCD methods that utilize self-supervised networks, such as X-Net, ACE-Net, and SRGCAE, our method can also demonstrate significant superiority, with average AUC values surpassing those of X-Net (3.96%), ACE-Net (6.93%), and SRGCAE (8.73%). Overall, the experimental results for DIs demonstrate that our method can generate high-quality DI, which can be used to obtain accurate change maps reflecting the ground changes.

2) Change Maps: We present the change maps generated by different methods on five datasets in Fig. 8, where TP, TN, FP, and FN are marked in white, black, red, and green, respectively. These change maps are obtained through image segmentation of the above difference images. Additionally, a detailed comparison of evaluation metrics is provided in Table III.

From Fig. 8, it can be seen that the comparison methods still lack robustness. While they can accurately detect changes

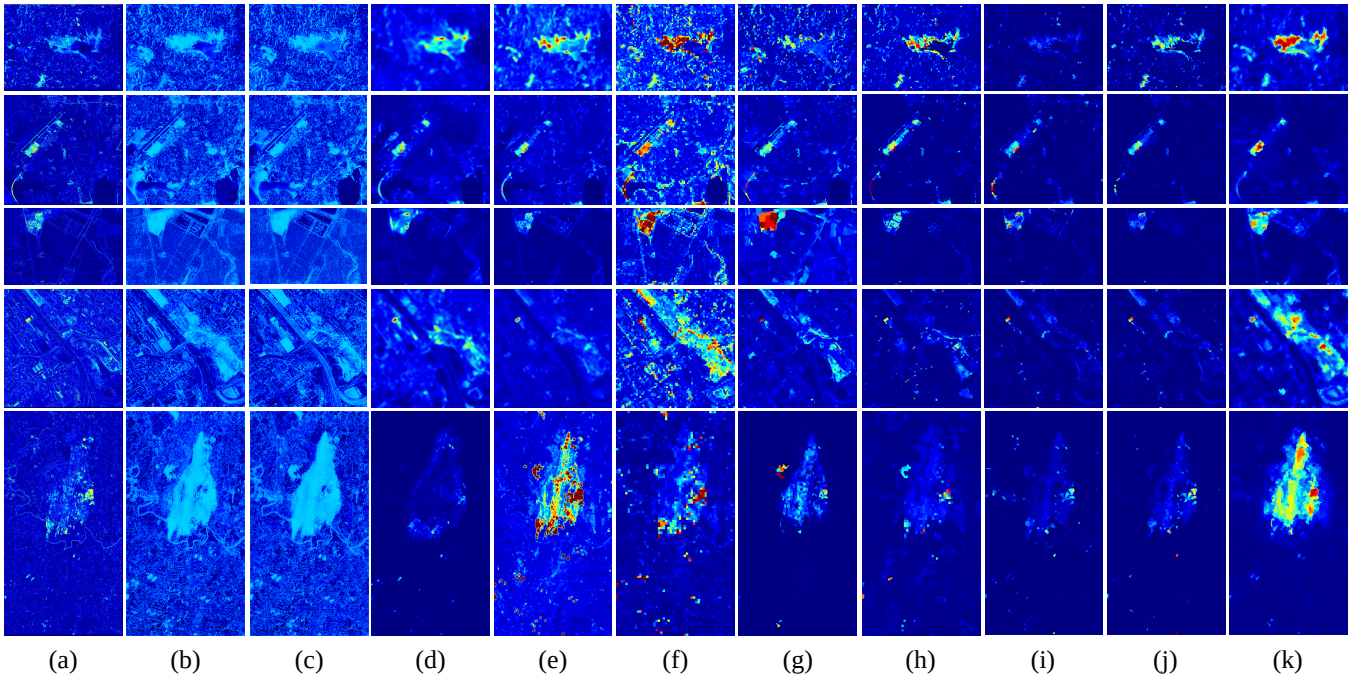

Fig. 6. Visual comparison of the DIs generated by different methods on the five datasets. (a) CCLMRF; (b) X-Net; (c) ACE-Net; (d) NPSG; (e) INLPG; (f) IRG-MCS; (g) SRGCAE; (h) VDF-HCD; (i) SCASC; (j) AGSCC; (k) ICSF. From top to down, the DIs from Italy, Tianhe, Shuguang, France, and Texas datasets are displayed in "jet" colormap. (Brighter regions are more likely to be changed, whereas darker regions are more likely to remain unchanged.)

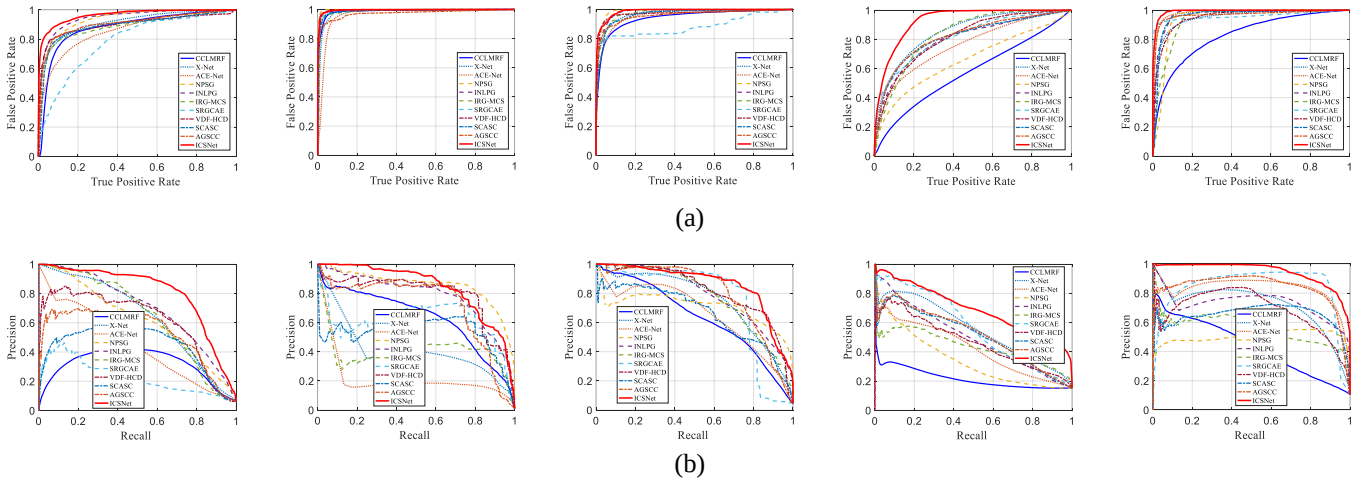

Fig. 7. ROC (a) and PR (b) curves of DIs generated by different methods. From left to right are the results on datasets Italy, Tianhe, Shuguang, France, and Texas.

in some datasets, they fail to deliver satisfactory results on all datasets. CCLMRF achieves the lowest FN on the Tianhe and Shuguang datasets, with 8 and 137 respectively. However, it presents a significant number of FP, resulting in low KC scores of 0.2122 and 0.4041. X-Net and ACE-net exhibit numerous false detections on the Tianhe and France datasets, leading to high OE. In particular, the similarity measure-based methods (including INLPG, IRG-MCS, SRGCAE) perform poorly on the Texas dataset, with many missed detections and consequently very small KC values. These methods struggle to accurately distinguish changed regions from unchanged ones, especially when the proportion of change regions is large. For instance, IRG-MCS and INLPG only achieve KC values of 0.1478 and 0.5684 on the Texas dataset, respectively, which are much lower than the proposed method.

Compared to the aforementioned methods, our proposed method achieves the best CD performance on all five datasets. The change maps generated by our method, as shown in Fig. 8(k), exhibit robustness with minimal FP (marked in red) and FN (marked in green). Moreover, the main changed regions are accurately identified, effectively illustrating that our method can accurately distinguish between unchanged and changed regions. As reported in Table III, the proposed method achieves the highest F1 and KC. For instance, the proposed ICSF exhibits significant improvements in KC compared to the second-best method on all datasets: 1.33%

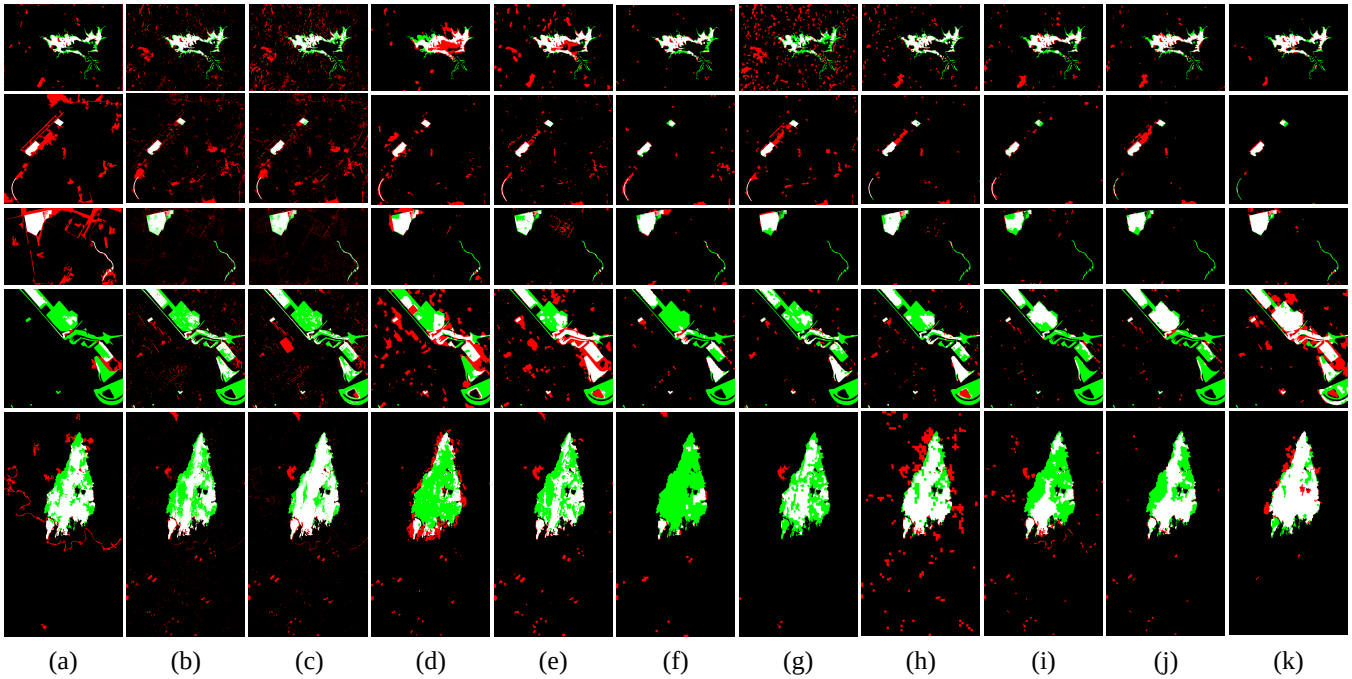

Fig. 8. Visual comparison of the change maps obtained by different methods on the five datasets, where the first to fifth rows are the Italy, Tianhe, Shuguang, France, and Texas datasets, respectively. (a) CCLMRF; (b) X-Net; (c) ACE-Net; (d) NPSG; (e) INLPG; (f) IRG-MCS; (g) SRGCAE; (h) VDF-HCD; (i) SCASC; (j) AGSCC; (k) ICSF. (Changed and unchanged regions are in white and black colors. Red indicates FP, green indicates FN)

TABLE III
PERFORMANCE COMPARISON ON CHANGE MAPS OBTAINED BY DIFFERENT METHODS ON THE FIVE DATASETS. THE BEST RESULTS AND THE SECOND-BEST RESULTS ARE IN RED AND BLUE, RESPECTIVELY.

Dataset		CCLMRF	X-Net	ACE-Net	NPSG	INLPG	IRG-MCS	SRGCAE	VDF-HCD	SCASC	AGSCC	ICSF
Italy	FP	2703	5325	8804	5702	7403	1389	11435	1268	3177	2840	1424
	FN	2716	2398	3040	2293	1494	2454	4388	2250	2630	2231	2003
	OE	5419	7723	11844	7995	8897	3843	15823	3518	5807	5071	3427
	OA	0.9562	0.9375	0.9042	0.9353	0.9280	0.9689	0.8720	0.9715	0.9530	0.9590	0.9723
	F1	0.6444	0.5752	0.4364	0.5716	0.5796	0.7291	0.2904	0.7535	0.6324	0.6803	0.7664
	KC	0.6210	0.5424	0.3883	0.5378	0.5435	0.7127	0.2277	0.7384	0.6074	0.6584	0.7517
Tianhe	FP	31757	25593	28589	11442	10448	6999	13850	5708	3666	7880	1155
	FN	8	250	412	16	147	809	182	370	880	617	1109
	OE	31765	25843	29010	11458	10595	7808	14032	6078	4546	8497	2264
	OA	0.9224	0.9369	0.9292	0.9720	0.9741	0.9809	0.9657	0.9851	0.9889	0.9792	0.9945
	F1	0.2282	0.2563	0.2283	0.4500	0.4624	0.4994	0.3919	0.5878	0.6272	0.4903	0.7605
	KC	0.2122	0.2414	0.2126	0.4402	0.4530	0.4914	0.3807	0.5814	0.6220	0.4819	0.7578
Shuguang	FP	61735	17355	23209	12068	6635	6897	1913	2529	3549	1077	5462
	FN	137	5196	5687	4904	8226	5487	7919	7065	11029	8435	4427
	OE	61872	22551	28896	16972	14861	12384	9832	9594	14578	9512	9889
	OA	0.8867	0.9587	0.9471	0.9689	0.9728	0.9773	0.9820	0.9824	0.9733	0.9826	0.9819
	F1	0.4466	0.6384	0.5733	0.7041	0.6943	0.7600	0.7775	0.7899	0.6587	0.7780	0.8070
	KC	0.4041	0.6173	0.5471	0.6880	0.6800	0.7481	0.7683	0.7808	0.6453	0.7692	0.7975
France	FP	52905	134003	167376	421538	362959	56391	60169	80051	39318	47564	326298
	FN	506870	393800	427776	371615	289265	431542	402158	374852	399325	367586	174836
	OE	559775	527803	595152	793153	652224	487933	462327	454903	438643	415150	501134
	OA	0.8601	0.8680	0.8512	0.8017	0.8369	0.8781	0.8844	0.8863	0.8903	0.8962	0.8747
	F1	0.2622	0.4461	0.3750	0.3718	0.4930	0.4174	0.4690	0.5044	0.4856	0.5349	0.6326
	KC	0.2144	0.3775	0.2977	0.2543	0.3962	0.3642	0.4152	0.4475	0.4362	0.4848	0.5582
Texas	FP	25377	19296	20585	36416	18239	7392	4519	59436	19289	5959	16583
	FN	43785	58693	26174	98944	66938	118744	71650	33795	81831	55244	19659
	OE	69162	77989	46759	135360	85177	126136	76169	93231	101120	61203	36242
	OA	0.9442	0.9371	0.9623	0.8908	0.9313	0.8982	0.9385	0.9248	0.9184	0.9506	0.9708
	F1	0.7181	0.6524	0.8189	0.3273	0.6039	0.1723	0.6126	0.6778	0.4974	0.7146	0.8610
	KC	0.6873	0.6189	0.7978	0.2740	0.5684	0.1478	0.5834	0.6356	0.4576	0.6891	0.8446
Average	OA	0.9139	0.9276	0.8901	0.9137	0.9286	0.9407	0.9285	0.9500	0.9448	0.9535	0.9588
	F1	0.4599	0.5137	0.4864	0.4850	0.5666	0.5156	0.5083	0.6627	0.5803	0.6396	0.7655
	KC	0.4278	0.4795	0.4487	0.4387	0.5282	0.4928	0.4751	0.6367	0.5537	0.6167	0.7420

(Italy),13.58% (Tianhe),1.67% (Shuguang), 7.34% (France), and 4.68% (Texas). Moreover, the average F1 and KC values of ICSF on the five datasets are 0.7655 and 0.7420, respectively, which are 10.28% and 10.53% higher than the second-

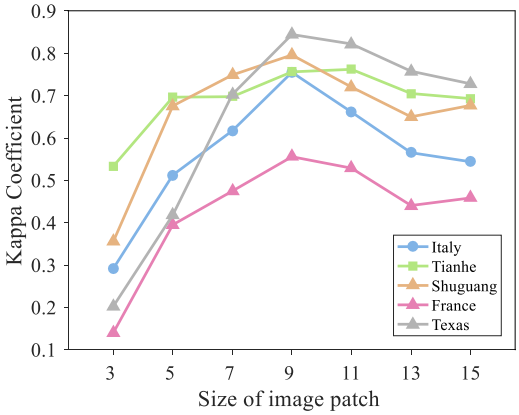

Fig. 9. Relationship between the image patch size p and Kappa Coefficient (KC) values of the proposed ICSF on different datasets.

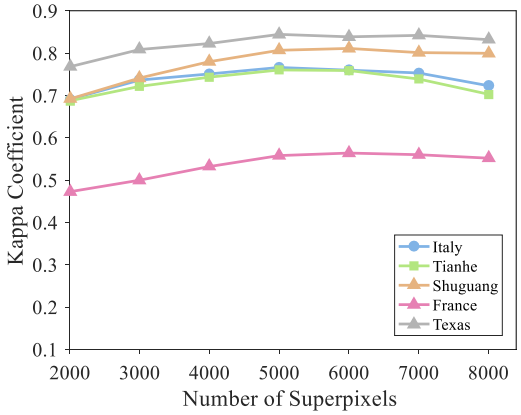

Fig. 10. Relationship between the number of superpixels N_{cs} and Kappa coefficient (KC) values of the proposed ICSF on different datasets.

best method. Overall, the experimental results on different datasets demonstrate that ICSF can achieve the best CD performance and outperform other methods in both quantitative and qualitative evaluations, further verifying its effectiveness and practicality.

V. Discussion

In this section, we conduct extensive experiments to comprehensively validate and analyze the effectiveness and practicality of the proposed method. First, we test and analyze the effect of different p and N_{cs} on the CD performance of the proposed method. Then, to validate our contributions in this work, we conduct multiple ablation studies to analyze and verify the roles of different components in our proposed method, including the refinement strategy based on superpixel segmentation, the application of contrastive learning, and dual-branch learning. Additionally, we report the computational time of our proposed method and ten comparison methods on the three datasets with different scales, further validating the practicality of the proposed method. Finally, we perform a stability analysis to verify the robustness of our method.

A. Parameter Analysis

In our framework, the size of the input image patch p and the number of superpixels N_{cs} play a key role in the proposed method, affecting the final CD performance. Here, the parameter setting is analyzed as follows. On the one hand, we set p to 3, 5, 7, 9, 11, 13, and 15 to analyze the influence of p. The relationship between p and KC over the five datasets is depicted in Fig. 9, from which we can see that the values of KC initially rise and then drop as p increases. When p is set too small, little spatial information is used that is not enough to support the network to learn useful feature representations, thus leading to more false detections. When p is set too large, much irrelevant spatial information is introduced, resulting in the inundation of important information. Both cases result in poor CD performance. From Fig. 9, it can be seen that when $p = 9$, the overall performance is relatively good. Therefore, we adopt $p = 9$ for our proposed method on all five datasets based on the above analysis.

On the other hand, keeping other settings constant, we explore the influence of N_{cs} through experiments. Here, N_{cs} is gradually adjusted between 2000 and 8000, with an increment step of 1000, to select an appropriate value. Fig. 10 illustrates the variation of KC values with the increasing N_{cs}. When N_{cs} is smaller, we have large superpixels that encompass multiple land cover classes and exhibit blurred image boundaries. This may cause different land cover types to be assigned the same change degree, thereby degrading the CD performance. In contrast, a large N_{cs} may lead to over-segmentation and increase the running time, thus affecting the effectiveness and efficiency of the proposed refinement strategy. In this paper, we simply set $N_{cs} = 5000$ for all datasets as a compromise choice, which can be adjusted according to practical circumstances and requirements.

B. Analysis of Superpixel Segmentation-based Refinement Strategy

As mentioned in Section III-E, we refine DI^{inter} and DI^{cross} to obtain high-quality DIs. Here, we first explore the effectiveness of this strategy through a comparative ablation study across five datasets, employing AUC and AP metrics to better accurately reflect the quality of DI. For simplicity, we analyze the fused difference image, which encompasses all the difference information from both DI^{inter} and DI^{cross}, rather than the individual DI^{inter} or DI^{cross}, as depicted in Table IV. The results indicate that our proposed refinement strategy has demonstrated improvements in terms of AUC and AP metrics across all five datasets, thereby enhancing the quality of the DIs and substantiating the effectiveness of our method. Fig. 11 visualizes the DIs before and after employing the proposed refinement strategy on the Texas dataset (AUC increased by 0.76%, AP increased by 4.05%). It is evident that the application of the proposed strategy to the original DI enables a more precise representation of ground object information in terms of superpixels. By unifying the change degrees of the same objects, irrelevant changes are effectively suppressed, thereby enhancing DI quality.

TABLE IV
ABLATION STUDY OF THE SUPERPIXEL SEGMENTATION-BASED REFINEMENT STRATEGY USING THE FUSED DI

Refinement	Dataset									
	Italy		Tianhe		Shuguang		France		Texas	
	AUC	AP	AUC	AP	AUC	AP	AUC	AP	AUC	AP
✘	0.9609	0.7802	0.9945	0.8294	0.9831	0.8263	0.9147	0.6608	0.9851	0.9045
✔	0.9652	0.8073	0.9957	0.8417	0.9877	0.8573	0.9287	0.6892	0.9927	0.9450

TABLE V
THE EFFECTIVENESS OF CONTRASTIVE LEARNING

Contrastive Learning Strategies	Dataset									
	Italy		Tianhe		Shuguang		France		Texas	
	F1	KC	F1	KC	F1	KC	F1	KC	F1	KC
✘	0.5901	0.5556	0.6774	0.6727	0.6929	0.6741	0.4395	0.3292	0.5687	0.5271
✔ (+BYOL)	0.7408	0.7284	0.7272	0.7236	0.7398	0.7221	0.6110	0.5186	0.6809	0.6402
✔ (Ours)	0.7294	0.7132	0.7388	0.7345	0.7429	0.7288	0.6068	0.5154	0.7248	0.6848

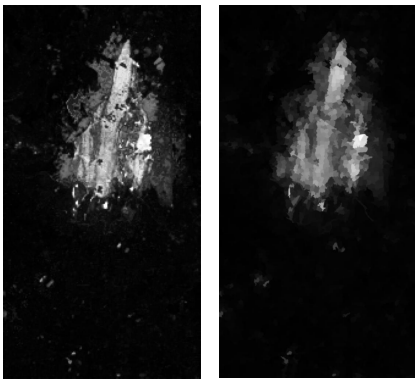

Fig. 11. The DIs acquired before and after implementing the superpixel segmentation-based refinement strategy on the Texas dataset. (a) before implementation; (b) after implementation.

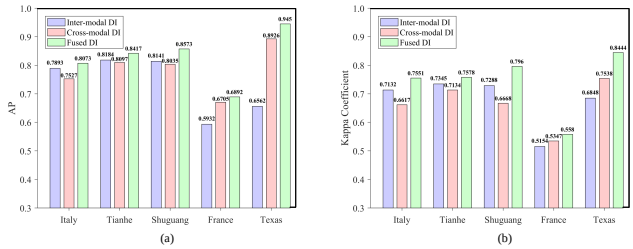

Fig. 12. The CD performance obtained from different DIs across five datasets. (a) AP metric of DIs; (b) KC metric of DIs.

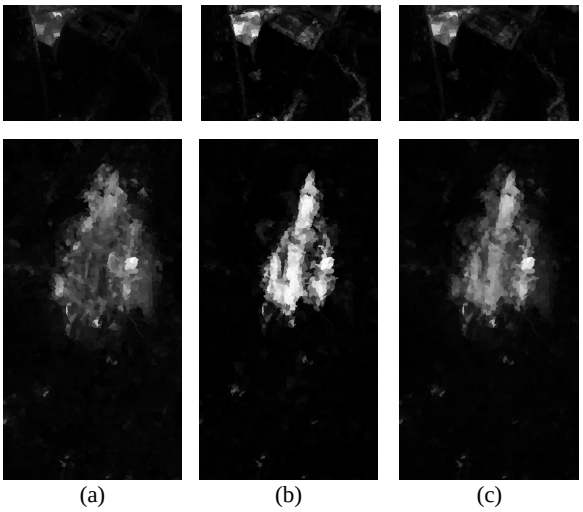

Fig. 13. The DIs obtained by computing change information from different branches on the two datasets. (a) Inter-modal branch DI; (b) Cross-modal branch DI; (c) Fused DI.

C. Analysis of Contrastive Learning in the Inter-modal Branch

The utilization of contrastive learning in the inter-modal branch is effective in our proposed framework, facilitating robust and discriminative feature extraction. Here, we conduct the ablation study of contrastive learning and the selected Simsiam strategy to validate their effectiveness, as shown in Table V. Here, we focus on the inter-modal branch. We use the original spectral and spatial information from the image patches to replace the features learned by contrastive learning, as depicted in the first row of Table V. In the second row of Table V, we employ another self-supervised contrastive learning strategy, named BYOL [58], to replace Simsiam in the inter-modal branch. It is evident that integrating the contrastive learning paradigm, whether employing BYOL or the selected Simsiam strategy, significantly improves CD performance, validating the efficacy of contrastive learning in our proposed framework. Furthermore, compared to BYOL, the selected SimSiam strategy can also demonstrate comparable performance. Specifically, on the Texas dataset, it achieves a 4.46% improvement in the KC score, further highlighting its rationale and effectiveness. In summary, utilizing features extracted through contrastive learning outperforms low-level information utilization, highlighting contrastive learning's effectiveness in extracting more representative and robust features and facilitating the accurate construction of KNN graphs.

D. Analysis of the Dual-branch DIs Fusion

In the proposed method, we compute change information to obtain DI^{inter} and DI^{cross} from the inter-modal and cross-modal branches, respectively. Subsequently, we fuse them to obtain the final DI. Here, the Otsu algorithm is employed

TABLE VI
COMPUTATIONAL TIME (IN SECONDS) OF DIFFERENT METHODS ON THREE DATASETS WITH DIFFERENT SCALES

Datasets	Image Size	CCLMRF	X-Net	ACE-Net	NPSG	INLPG	IRG-MCS	SRGCAE	VDF-HCD	SCASC	AGSCC	ICSF
Italy	300 × 412	92.08	497.10	317.09	79.75	23.18	8.36	288.93	64.14	6.5	84.12	43.76
Shuguang	593 × 921	212.84	1211.26	1086.48	237.64	210.74	25.72	332.72	82.88	10.21	93.51	108.05
Texas	1534 × 808	430.19	3731.53	2575.95	660.97	578.78	46.41	640.12	126.50	21.24	97.69	260.86

TABLE VII
COMPUTATIONAL TIME (IN SECONDS) ON EACH PART OF THE PROPOSED ICSF

Dataset	ICSF				
	t_{pre}	t_{train}	t_{compute}	t_{refine}	t_{total}
Italy	0.12	39.49	3.12	1.03	43.76
Shuguang	0.96	84.28	19.49	3.32	108.05
Texas	1.44	182.26	69.70	7.46	260.86

for generating the respective change maps from DI^{inter} and DI^{cross}. In Fig.12, we evaluate the performance of DI^{inter}, DI^{cross} and their fused DI based on the AP and KC metrics across the five datasets. The results indicate that the inter-modal DI outperforms the cross-modal DI in the Italy, Tianhe, and Shuguang datasets. Conversely, the cross-modal DI exhibits superior performance in the France and Texas datasets. Furthermore, the fused DI achieves the highest KC metric across all five datasets, demonstrating that the inclusion of the cross-modal branch learning enables the proposed method to capture more diverse change information, achieving the complementarity of change information and better distinguishing between changed and unchanged regions. For example, on the Texas dataset, the cross-modal branch (KC in 0.7538) outperforms the inter-modal branch (KC in 0.6848), further validating its effectiveness and addressing the limitations of relying solely on computing constructed similarity metrics to detect changes. Additionally, as shown in Fig.13, we provide a visual comparison of the DIs on the Shuguang and Texas datasets, which shows that changed areas are highlighted while unchanged areas are mostly suppressed after fusion, leading to an improvement in CD performance.

E. Running Time Cost

When enhancing the performance of HCD methods, it is important to consider their computational cost to ensure better practical applicability. Without loss of generality, we select three datasets of different scales (Italy, Shuguang, and Texas) to validate the practicality of our proposed ICSF method. The computational time of our method and ten comparison methods are listed in Table VI. Among these methods, CCLMRF is implemented with C++; X-Net, ACE-Net, SRGCAE, and the proposed ICSF are implemented with Python, and other methods are implemented with MATLAB. The computational time of each part of ICSF is also calculated in Table VII, where t_{pre}, t_{train}, t_{compute}, and t_{refine} represents the computational time spent in data preprocessing, dual-branch learning, change information computing and superpixel segmentation-based refinement. From Table VI, it can be seen that ICSF is still a bit time-consuming compared to IRG-MCS and SCASC. Considering the good performance achieved by ICSF, the use of advanced devices and efficient programming languages such as C can further reduce the running time, thus highlighting its effectiveness and practicality.

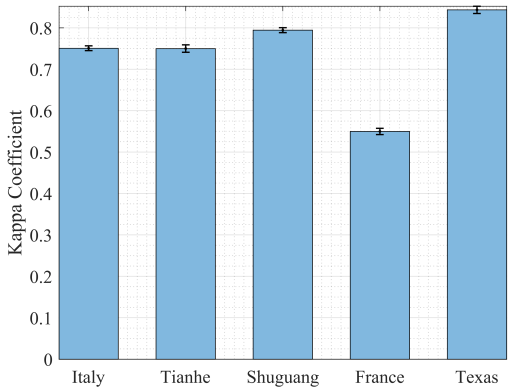

Fig. 14. Error bar plot of Kappa coefficient (KC) values from 20 independent runs of ICSF across five datasets.

F. Stability Analysis

As described in Section IV-C, all image patches are randomly divided into the training and testing sets in a ratio of 8:2. However, the sample selection of the training and test sets may affect the CD performance of the proposed ICSF. Hence, we conduct a stability analysis of ICSF on the sample selection to evaluate the stability of ICSF. Specifically, we execute ICSF for 20 independent runs, each using the random sample selection to obtain the corresponding KC value. The average KC values and their standard deviations for the Italy, Tianhe, Shuguang, France, and Texas datasets are 0.7506 ($\pm$0.0058), 0.7499 ($\pm$0.0089), 0.7941 ($\pm$0.0060), 0.5499 ($\pm$0.0076), and 0.8434 ($\pm$0.0088), respectively. Additionally, the error bar plot of the KC values across different datasets is presented in Fig. 1. As can be seen, the performance fluctuation of the proposed method is very small, which demonstrates that ICSF can achieve stable detection results from different datasets.

VI. CONCLUSION

In this article, to more thoroughly detect changes in ground objects within heterogeneous images, we propose an integrating inter-modal and cross-modal self-supervised learning framework for HCD. In the inter-modal branch, we establish efficient Siamese networks to learn representative and discriminative features from heterogeneous images, instead of using the spatial information and spectral features from the heterogeneous images themselves. To capture more diverse change

information, we introduce a cross-modal branch involving conducting cross-modal reconstruction on heterogeneous images, thereby ensuring that the heterogeneous images are mapped to a shared feature space. Furthermore, a refinement strategy based on superpixel segmentation is applied to enhance the quality of the obtained DIs from both branches. The remarkable experimental results on five heterogeneous datasets have validated the superiority and practicality of the proposed method over ten existing unsupervised SOTA methods.

As elaborated in Section III-E, we apply PCA to construct a false RGB image using the extracted first three principal components. This process may result in a certain loss of information, potentially causing inaccuracies in following superpixel segmentation. Second, the proposed framework is capable of identifying whether the region has changed; however, it lacks the ability to accurately distinguish the specific types of changes. Therefore, future research will involve using more advanced image segmentation methods and implementing fine-grained semantic CD to address these limitations.

REFERENCES

[1] P. Ma, M. Macdonald, S. Rouse, and J. Ren, "Automatic geolocation and measuring of offshore energy infrastructure with multimodal satellite data," *IEEE J. Ocean. Eng.*, vol. 49, no. 1, pp. 66–79, 2024.

[2] R. Radke, S. Andra, O. Al-Kofahi, and B. Roysam, "Image change detection algorithms: a systematic survey," *IEEE Trans. Image Process.*, vol. 14, no. 3, pp. 294–307, 2005.

[3] H. Chen, E. Nemni, S. Vallecorsa, X. Li, C. Wu, and L. Bromley, "Dual-tasks siamese transformer framework for building damage assessment," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, 2022, pp. 1600–1603.

[4] Y. Qing, D. Ming, Q. Wen, Q. Weng, L. Xu, Y. Chen, Y. Zhang, and B. Zeng, "Operational earthquake-induced building damage assessment using cnn-based direct remote sensing change detection on superpixel level," *Int. J. Appl. Earth Obs. Geoinf.*, vol. 112, p. 102899, 2022.

[5] K. Kalinaki, O. A. Malik, and D. T. Ching Lai, "Fcd-attresu-net: An improved forest change detection in sentinel-2 satellite images using attention residual u-net," *Int. J. Appl. Earth Obs. Geoinf.*, vol. 122, p. 103453, 2023.

[6] X. Wu, D. Hong, and J. Chanussot, "Uiu-net: U-net in u-net for infrared small object detection," *IEEE Trans. Image Process.*, vol. 32, pp. 364–376, 2023.

[7] D. Hong, J. Yao, C. Li, D. Meng, N. Yokoya, and J. Chanussot, "Decoupled-and-coupled networks: Self-supervised hyperspectral image super-resolution with subpixel fusion," *IEEE Trans. Geosci. Remote Sensing*, vol. 61, pp. 1–12, 2023.

[8] X. Wu, D. Hong, and J. Chanussot, "Convolutional neural networks for multimodal remote sensing data classification," *IEEE Trans. Geosci. Remote Sensing*, vol. 60, pp. 1–10, 2022.

[9] C. Li, B. Zhang, D. Hong, X. Jia, A. Plaza, and J. Chanussot, "Learning disentangled priors for hyperspectral anomaly detection: A coupling model-driven and data-driven paradigm," *IEEE Trans. Neural Netw. Learn. Syst.*, pp. 1–14, 2024.

[10] T. Zhan, M. Gong, X. Jiang, and E. Zhang, "S3net: Superpixel-guided self-supervised learning network for multitemporal image change detection," *IEEE Geosci. Remote Sens. Lett.*, vol. 20, pp. 1–5, 2023.

[11] B. Du, L. Ru, C. Wu, and L. Zhang, "Unsupervised deep slow feature analysis for change detection in multi-temporal remote sensing images," *IEEE Trans. Geosci. Remote Sensing*, vol. 57, no. 12, pp. 9976–9992, 2019.

[12] H. Du, Y. Zhuang, S. Dong, C. Li, H. Chen, B. Zhao, and L. Chen, "Bilateral semantic fusion siamese network for change detection from multitemporal optical remote sensing imagery," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.

[13] C. Wu, H. Chen, B. Du, and L. Zhang, "Unsupervised change detection in multitemporal vhr images based on deep kernel pca convolutional mapping network," *IEEE T. Cybern.*, vol. 52, no. 11, pp. 12 084–12 098, 2022.

[14] J. Ma, D. Li, X. Tang, Y. Yang, X. Zhang, and L. Jiao, "Unsupervised sar image change detection based on feature fusion of information transfer," *IEEE Geosci. Remote Sens. Lett.*, vol. 20, pp. 1–5, 2023.

[15] S. Fang, C. Qi, S. Yang, Z. Li, W. Wang, and Y. Wang, "Unsupervised sar change detection using two-stage pseudo labels refining framework," *IEEE Geosci. Remote Sens. Lett.*, vol. 21, pp. 1–5, 2024.

[16] B. Cui, Y. Peng, Y. Zhang, H. Yin, H. Fang, S. Guo, and P. Du, "Enhanced edge information and prototype constrained clustering for sar change detection," *IEEE Trans. Geosci. Remote Sensing*, vol. 62, pp. 1–16, 2024.

[17] J. Zhao, S. Xiao, W. Dong, J. Qu, and Y. Li, "Dictionary learning-guided deep interpretable network for hyperspectral change detection," *IEEE Geosci. Remote Sens. Lett.*, vol. 20, pp. 1–5, 2023.

[18] Y. Li, J. Ren, Y. Yan, Q. Liu, P. Ma, A. Petrovski, and H. Sun, "Cbanet: An end-to-end cross-band 2-d attention network for hyperspectral change detection in remote sensing," *IEEE Trans. Geosci. Remote Sensing*, vol. 61, pp. 1–11, 2023.

[19] J. Ding, X. Li, J. Li, and S. Chen, "Multiple spatial–spectral features aggregated neural network for hyperspectral change detection," *IEEE Geosci. Remote Sens. Lett.*, vol. 21, pp. 1–5, 2024.

[20] H. Chen, N. Yokoya, C. Wu, and B. Du, "Unsupervised multimodal change detection based on structural relationship graph representation learning," *IEEE Trans. Geosci. Remote Sensing*, vol. 60, pp. 1–18, 2022.

[21] D. Xiang, X. Pan, H. Ding, J. Cheng, and X. Sun, "Two-stage registration of sar images with large distortion based on superpixel segmentation," *IEEE Trans. Geosci. Remote Sensing*, vol. 62, pp. 1–15, 2024.

[22] D. Xiang, H. Ding, X. Sun, J. Cheng, C. Hu, and Y. Su, "Polsar image registration combining siamese multiscale attention network and joint filter," *IEEE Trans. Geosci. Remote Sensing*, vol. 62, pp. 1–14, 2024.

[23] R. Touati, M. Mignotte, and M. Dahmane, "Multimodal change detection in remote sensing images using an unsupervised pixel pairwise-based markov random field model," *IEEE Trans. Image Process.*, vol. 29, pp. 757–767, 2020.

[24] J. Liu, M. Gong, K. Qin, and P. Zhang, "A deep convolutional coupling network for change detection based on heterogeneous optical and radar images," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 3, pp. 545–559, 2018.

[25] Y. Zhan, K. Fu, M. Yan, X. Sun, H. Wang, and X. Qiu, "Change detection based on deep siamese convolutional network for optical aerial images," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 10, pp. 1845–1849, 2017.

[26] R. Caye Daudt, B. Le Saux, and A. Boulch, "Fully convolutional siamese networks for change detection," in *Proc. IEEE Int. Conf. Inf. Process. (ICIP)*, 2018, pp. 4063–4067.

[27] Y. Sun, L. Lei, D. Guan, and G. Kuang, "Iterative robust graph for unsupervised change detection of heterogeneous remote sensing images," *IEEE Trans. Image Process.*, vol. 30, pp. 6277–6291, 2021.

[28] Y. Sun, L. Lei, X. Li, H. Sun, and G. Kuang, "Nonlocal patch similarity based heterogeneous remote sensing change detection," *Pattern Recognit.*, vol. 109, p. 107598, 2021.

[29] T. Han, Y. Tang, B. Zou, and H. Feng, "Unsupervised multimodal change detection based on adaptive optimization of structured graph," *Int. J. Appl. Earth Obs. Geoinf.*, vol. 126, p. 103630, 2024.

[30] L. Zhang, L. Zhang, and B. Du, "Deep learning for remote sensing data: A technical tutorial on the state of the art," *IEEE Geosci. Remote Sens. Mag.*, vol. 4, no. 2, pp. 22–40, 2016.

[31] L. Wan, Y. Xiang, and H. You, "A post-classification comparison method for sar and optical images change detection," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 7, pp. 1026–1030, 2019.

[32] L. Wan, Y. Xiang, and H. You , "An object-based hierarchical compound classification method for change detection in heterogeneous optical and sar images," *IEEE Trans. Geosci. Remote Sensing*, vol. 57, no. 12, pp. 9941–9959, 2019.

[33] T. Han, Y. Tang, X. Yang, Z. Lin, B. Zou, and H. Feng, "Change detection for heterogeneous remote sensing images with improved training of hierarchical extreme learning machine (helm)," *Remote Sens.*, vol. 13, no. 23, 2021.

[34] H. Li, M. Gong, M. Zhang, and Y. Wu, "Spatially self-paced convolutional networks for change detection in heterogeneous images," *IEEE J. Sel. Top. Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 4966–4979, 2021.

[35] Y. Wu, J. Li, Y. Yuan, A. K. Qin, Q.-G. Miao, and M.-G. Gong, "Commonality autoencoder: Learning common features for change detection from heterogeneous images," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 9, pp. 4257–4270, 2022.

[36] J. Liu, W. Zhang, F. Liu, and L. Xiao, "A probabilistic model based on bipartite convolutional neural network for unsupervised change detection," *IEEE Trans. Geosci. Remote Sensing*, vol. 60, pp. 1–14, 2022.
[37] Y. Xing, Q. Zhang, L. Ran, X. Zhang, H. Yin, and Y. Zhang, "Progressive modality-alignment for unsupervised heterogeneous change detection," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 61, pp. 1–12, 2023.
[38] J. Shi, T. Wu, A. Kai Qin, Y. Lei, and G. Jeon, "Self-guided autoencoders for unsupervised change detection in heterogeneous remote sensing images," *IEEE Transactions on Artificial Intelligence*, vol. 5, no. 6, pp. 2458–2471, 2024.
[39] X. Niu, M. Gong, T. Zhan, and Y. Yang, "A conditional adversarial network for change detection in heterogeneous images," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 1, pp. 45–49, 2019.
[40] X. Li, Z. Du, Y. Huang, and Z. Tan, "A deep translation (gan) based change detection network for optical and sar remote sensing images," *ISPRS-J. Photogramm. Remote Sens.*, vol. 179, pp. 14–34, 2021.
[41] L. T. Luppino, M. Kampffmeyer, F. M. Bianchi, G. Moser, S. B. Serpico, R. Jenssen, and S. N. Anfinsen, "Deep image translation with an affinity-based change prior for unsupervised multimodal change detection," *IEEE Trans. Geosci. Remote Sensing*, vol. 60, pp. 1–22, 2022.
[42] L. T. Luppino, M. A. Hansen, M. Kampffmeyer, F. M. Bianchi, G. Moser, R. Jenssen, and S. N. Anfinsen, "Code-aligned autoencoders for unsupervised change detection in multimodal remote sensing images," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 1, pp. 60–72, 2024.
[43] L. T. Luppino, F. M. Bianchi, G. Moser, and S. N. Anfinsen, "Unsupervised image regression for heterogeneous change detection," *IEEE Trans. Geosci. Remote Sensing*, vol. 57, no. 12, pp. 9960–9975, 2019.
[44] M. Mignotte, "A fractal projection and markovian segmentation-based approach for multimodal change detection," *IEEE Trans. Geosci. Remote Sensing*, vol. 58, no. 11, pp. 8046–8058, 2020.
[45] Y. Sun, L. Lei, X. Li, X. Tan, and G. Kuang, "Structure consistency-based graph for unsupervised change detection with homogeneous and heterogeneous remote sensing images," *IEEE Trans. Geosci. Remote Sensing*, vol. 60, pp. 1–21, 2022.
[46] Y. Sun, L. Lei, D. Guan, G. Kuang, and L. Liu, "Graph signal processing for heterogeneous change detection," *IEEE Trans. Geosci. Remote Sensing*, vol. 60, pp. 1–23, 2022.
[47] G. K. Yuli SUN, Lin LEI, "A structure consistency-based energy model for heterogeneous optical and sar images change detection," *SCIENTIA SINICA Informationis*, vol. 53, no. 10, pp. 2016–, 2023.
[48] Y. Sun, L. Lei, D. Guan, M. Li, and G. Kuang, "Sparse-constrained adaptive structure consistency-based unsupervised image regression for heterogeneous remote-sensing change detection," *IEEE Trans. Geosci. Remote Sensing*, vol. 60, pp. 1–14, 2022.
[49] Y. Sun, L. Lei, D. Guan, J. Wu, G. Kuang, and L. Liu, "Image regression with structure cycle consistency for heterogeneous change detection," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 2, pp. 1613–1627, 2024.
[50] Y. Sun, L. Lei, Z. Li, and G. Kuang, "Similarity and dissimilarity relationships based graphs for multimodal change detection," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 208, pp. 70–88, 2024.
[51] H. Li, Y. Li, G. Zhang, R. Liu, H. Huang, Q. Zhu, and C. Tao, "Global and local contrastive self-supervised learning for semantic segmentation of hr remote sensing images," *IEEE Trans. Geosci. Remote Sensing*, vol. 60, pp. 1–14, 2022.
[52] H. Jung, Y. Oh, S. Jeong, C. Lee, and T. Jeon, "Contrastive self-supervised learning with smoothed representation for remote sensing," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.
[53] Z. Chen, C. Zhang, B. Zhang, and Y. He, "Triplet contrastive learning framework with adversarial hard-negative sample generation for multimodal remote sensing images," *IEEE Trans. Geosci. Remote Sensing*, vol. 62, pp. 1–15, 2024.
[54] X. Liu, F. Zhang, Z. Hou, L. Mian, Z. Wang, J. Zhang, and J. Tang, "Self-supervised learning: Generative or contrastive," *IEEE Trans. Knowl. Data Eng.*, vol. 35, no. 1, pp. 857–876, 2023.
[55] Z. Wu, Y. Xiong, S. X. Yu, and D. Lin, "Unsupervised feature learning via non-parametric instance discrimination," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2018, pp. 3733–3742.
[56] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2020, pp. 9726–9735.
[57] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Int. Conf. Mach. Learn. (ICML)*. PMLR, 2020, pp. 1597–1607.
[58] J. Grill, F. Strub, F. Altché, C. Tallec, P. H. Richemond, E. Buchatskaya, C. Doersch, B. Á. Pires, Z. D. Guo, M. G. Azar, B. Piot, K. Kavukcuoglu, R. Munos, and M. Valko, "Bootstrap your own latent: A new approach to self-supervised learning," *CoRR*, vol. abs/2006.07733, 2020.
[59] X. Chen and K. He, "Exploring simple siamese representation learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2021, pp. 15 745–15 753.
[60] F. Jiang, M. Gong, H. Zheng, T. Liu, M. Zhang, and J. Liu, "Self-supervised global–local contrastive learning for fine-grained change detection in vhr images," *IEEE Trans. Geosci. Remote Sensing*, vol. 61, pp. 1–13, 2023.
[61] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, "Slic superpixels compared to state-of-the-art superpixel methods," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 11, pp. 2274–2282, 2012.
[62] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Trans. Syst. Man Cybern.*, vol. 9, no. 1, pp. 62–66, 1979.
[63] G. Moser and S. Serpico, "Generalized minimum-error thresholding for unsupervised change detection from sar amplitude imagery," *IEEE Trans. Geosci. Remote Sensing*, vol. 44, no. 10, pp. 2972–2982, 2006.
[64] A. M. Ikotun, A. E. Ezugwu, L. Abualigah, B. Abuhaija, and J. Heming, "K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data," *Inf. Sci.*, vol. 622, pp. 178–210, 2023.
[65] J. C. Bezdek, R. Ehrlich, and W. Full, "Fcm: The fuzzy c-means clustering algorithm," *Comput. Geosci.*, vol. 10, no. 2, pp. 191–203, 1984.
[66] J. Liu, W. Zhang, F. Liu, and L. Xiao, "A probabilistic model based on bipartite convolutional neural network for unsupervised change detection," *IEEE Trans. Geosci. Remote Sensing*, vol. 60, pp. 1–14, 2022.
[67] M. Mignotte, "Mrf models based on a neighborhood adaptive class conditional likelihood for multimodal change detection," *AI, Comput. Sci. Robot. Technol.*, Mar 2022.

Erlei Zhang is currently an Associate Professor in the School of Information Engineering, Northwest A&F University. He received the Ph.D. degrees in electronic engineering from Xidian University, Xi'an, China, in 2015. From 2018 to 2020, he was a postdoctoral fellow at the UT Southwestern Medical Center, USA and Northwest University, China.

His current research interests include pattern recognition, machine learning, and remote sensing image analysis and understanding. He has published over 40 papers in top tier academic journals and conferences, and has over 10 patents.

He Zong received the B.S.degree from Northwest A&F University, Xianyang, China, in 2023. He is currently pursuing the M.S. degree in computer science and technology with the College of Northwest A&F University.

His research interests include computer vision, remote sensing image processing, and multitemporal image analysis.

Xinyu Li received a Bachelor's degree from Northwest A&F University in Yangling, China, in 2022. She is currently pursuing a Master's degree in Computer Science and Technology at the College of Information Engineering, Northwest A&F University.

Her research interests include machine learning and image processing.

Mingchen Feng received the M.S. degree in software engineering from the School of Software and Microelectronics, Northwestern Polytechnical University, Xi'an, China, in 2015. He was a Visiting Researcher with the Department of Electronic and Electrical Engineering, University of Strathclyde, Glasgow, U.K., in 2018.

His research interests include big data analytics, data mining, machine learning, deep learning, and data visualization.

Jinchang Ren (Senior Member, IEEE) received the B.Eng., M.Eng., and D.Eng. degrees from Northwestern Polytechnical University, Xi'an, China, in 1992, 1997, and 2000, respectively, and the Ph.D. degree from the University of Bradford, Bradford, U.K., in 2019.

He is currently a Professor of computing with Robert Gordon University, Aberdeen, U.K. He has authored or coauthored more than 300 peer-reviewed journal articles or conference papers. His research interests include hyperspectral imaging, image processing, computer vision, big data analytics, and machine learning.