Unsupervised Multimodal Change Detection Based on Structural Relationship Graph Representation Learning

Hongruixuan Chen, Student Member, IEEE, Naoto Yokoya, Member, IEEE, Chen Wu, Member, IEEE, and Bo Du, Senior Member, IEEE

Abstract—Unsupervised multimodal change detection is a practical and challenging topic that can play an important role in time-sensitive emergency applications. To address the challenge that multimodal remote sensing images cannot be directly compared due to their modal heterogeneity, we take advantage of two types of modality-independent structural relationships in multimodal images. In particular, we present a structural relationship graph representation learning framework for measuring the similarity of the two structural relationships. First, structural graphs are generated from preprocessed multimodal image pairs by means of an object-based image analysis approach. Then, a structural relationship graph convolutional autoencoder (SR-GCAE) is proposed to learn robust and representative features from graphs. Two loss functions aiming at reconstructing vertex information and edge information are presented to make the learned representations applicable for structural relationship similarity measurement. Subsequently, the similarity levels of two structural relationships are calculated from learned graph representations, and two difference images are generated based on the similarity levels. After obtaining the difference images, an adaptive fusion strategy is presented to fuse the two difference images. Finally, a morphological filtering-based postprocessing approach is employed to refine the detection results. Experimental results on six datasets with different modal combinations demonstrate the effectiveness of the proposed method.

Index Terms—Change detection, graph convolutional autoencoder, graph representation learning, multimodal remote sensing images, structural relationship.

I. INTRODUCTION

CHANGE detection is the process of identifying the changes of objects or phenomena in the same geographic area by analyzing remote sensing images acquired at different times [1]. It has been widely used in many real-world applications, such as urban studies, ecosystem monitoring, resource management, armed conflict monitoring, and damage assessment [2], [3], [4], [5], [6], [7].

Unimodal change detection or homogeneous change detection, where prechange and postchange images are collected by the same kind of sensors with the same sensor parameters, has been widely studied over the past decades. Many mature and effective paradigms, including traditional and deep learning-based models, have been proposed for different kinds of remote sensing images, including multispectral images [8], [9], [10], [11], hyperspectral images [12], [13], [14], [15], [16], very-high-resolution optical images [17], [18], [19], [20], [21], [22], and synthetic aperture radar (SAR) images [23], [24], [25], [26], [27], [28]. Recently, with the rapid development of earth observation technology, more and more remote sensing images representing land-cover information can be obtained from different sensors at the same time. This development provides data support for the research of multimodal change detection. Compared to unimodal change detection, the prechange and postchange images in multimodal change detection are obtained by different sensors. This means that multimodal change detection can alleviate the constraints of atmospheric conditions and revisit period cycles of satellites, thereby providing land-cover change information timely. Thus, multimodal change detection has great practical significance, especially for immediate evaluation and emergency disasters.

However, research on multimodal change detection is not as well established as that on unimodal change detection, despite the need for more advanced methods due to their practical importance. Under the scope of multimodal change detection, the images could be captured by different sensors (e.g., an optical image at $T_1$ and an SAR image at $T_2$ or a multispectral image captured by Landsat-5 at $T_1$ and a multispectral image captured by EO-1 at $T_2$) or recorded with different sensor parameters (e.g., a near-infrared band image at $T_1$ and an RGB bands image at $T_2$ or the images at $T_1$ and $T_2$ acquired by Radarsat-2, but with different looks) [29], [30]. As a result, the prechange and postchange images have different modalities, making it difficult for those intuitive paradigms designed for unimodal images to obtain accurate detection results from multimodal images. Generally,
multimodal change detection can be divided into supervised and unsupervised ones depending on whether or not label information is provided for the detection model. However, annotating labels for multimodal data is very labor-intensive and requires extensive expert knowledge as a guide [31], [32]. Thus, unsupervised methods are more popular in practice and undoubtedly more challenging. In this article, we focus on unsupervised multimodal change detection.

The fundamental idea of unsupervised multimodal change detection is to find a domain, where incomparable multimodal images become comparable. According to the transformation way and domain type, the existing methods can be generally divided into four categories: classification methods [33], [34], [35], [36], [37], modality translation methods [38], [39], [40], [41], [42], [43], feature learning-based methods [44], [45], [46], [47], [48], [49], [50], and similarity measure-based methods [30], [51], [52], [53], [54], [55], [56], [57]. Among them, similarity measure-based methods define a modality-independent metric that can be used to distinguish the changed and unchanged areas. The main advantages of this type of methods are that they are intuitive and easy to implement in practice. In this article, we follow the research line of the similarity measure-based method. However, the similarity measures proposed in these methods only utilize the original spectral features or the low-level spatial information in remote sensing images [55], [56], [58]. They are not robust enough when detection conditions are complex [59], [60], such as the diversity of land-cover objects in different research sites, the difference in imaging conditions, or the strong speckle noise in SAR images.

Therefore, we propose an unsupervised structural relationship graph representation learning framework for multimodal change detection (SRGRL-CD). The proposed framework not only employs self-similarity (called nonlocal structural relationship in this article) but also explores the local structural relationship observed from multimodal remote sensing images. Since the structural information of images can be expressed in the form of graph data, we naturally introduce graph representation learning and design a structural relationship graph convolutional autoencoder (SR-GCAE). The proposed network can learn representative and robust graph features from information contained in the vertices and edges of graphs through two reconstruction optimization objectives. Benefiting from this, the degree of local and nonlocal structural similarity can be calculated from the features learned by SR-GCAE instead of calculating it from low-level spectral features. In addition, a simple but effective adaptive fusion strategy is presented to fuse the difference images obtained by calculating the similarity levels of two structural relationships.

In particular, the main contribution of our work can be summarized as follows.

1) We present the first attempt at designing a graph representation learning framework for unsupervised multimodal change detection. With two kinds of reconstruction objectives as the loss function, the proposed network can learn robust high-level graph representations to measure the similarity levels of local and nonlocal structural relationships.

2) An adaptive fusion strategy based on the discrimination of change intensity in difference images is proposed to better highlight changed pixels and suppress unchanged pixels.

3) The proposed method outperforms the state-of-the-art (SOTA) methods on six change detection datasets with different modality combinations, showing its superiority. The source code for this work is publicly available for contributing to this field.

The remainder of this article is organized as follows. Section II reviews the representative work in unsupervised multimodal change detection. In Section III, the preliminaries of two structural relationships and graph convolutional networks are introduced. Section IV elaborates on the multimodal change detection framework based on the proposed SR-GCAE. To evaluate the proposed method, the experiments on six multimodal datasets are carried out in Section V. Finally, Section VI draws the conclusion of our work in this article.

II. RELATED WORK

In this section, we briefly review some representative methods in each category of unsupervised multimodal change detection in Section I.

A. Classification Methods

They transform the multimodal images into a common category domain by classifying the prechange and postchange images separately, and then compare the classification maps in the category domain to detect the changes. Wan et al. [33] presented a postclassification comparison method based on the superpixel segmentation for SAR and optical image change detection. Based on this work, they further proposed a region-based multitemporal hierarchical Markov random field to improve classification accuracy [34]. Liu et al. [35] proposed to use multidimensional evidential reasoning to find a common joint-class space. In [36], a hierarchical extreme learning machine was introduced to classify multimodal images. Li et al. [37] designed a spatially self-paced convolutional network to enhance classification performance. As can be seen, this type of method relies heavily on classification accuracy. However, unsupervised classification methods still have difficulty in obtaining very accurate classification results.

B. Modality Translation Methods

They reduce the modality difference by transforming one image from its modality to the modality of the other image. The paradigms of style transfer in the field of computer vision are introduced and developed. In [38], a conditional generative adversarial network (GAN) was applied to translate the optical image to the style of the SAR image. In [39], an original concentric circular invariant convolution model was proposed for modality transformation. Jiang et al. [40] proposed a deep homogeneous feature fusion model based on image style transfer. The optical images were transferred to the style of the SAR images, and change detection was performed. In [41] and [42], the ideas of adversarial learning, image-to-image translation, and cycle consistency were introduced for
designing networks and transforming image modality. In [43], to better match the distributions of real images, the CutMix transformation [61] was applied for training the discriminators of GANs. Nevertheless, the running time and computational overhead of style transfer learning approaches and GANs are quite large.

C. Feature Learning-Based Methods

They design an appropriate model, mainly deep learning models, to find a high-dimensional feature space where features of multimodal images can be directly compared. Zhang et al. [44] proposed to utilize the denoising autoencoder to learn high-level representations from image patches for multiresolution images. In [45], an unsupervised symmetric convolutional coupling network (SCCN) was proposed to transform multimodal images into a feature space where their feature representations become more consistent. In [46], the deep belief network (DBN) was introduced to learn features. Zhan et al. [47] proposed an iterative feature mapping network to learn multiple land-cover change types from multimodal images. Zhan et al. [48] further used the logarithmic transformation to transform SAR images so that they have similar statistical distributions as the optical images. Touati et al. [49] utilized a stacked sparse autoencoder to build an anomaly detection model to find the latent common feature space. Recently, Wu et al. [50] proposed a commonality autoencoder for heterogeneous change detection based on commonality exploration. These methods assume that unchanged areas occupy a large proportion of multimodal image pairs. Under this circumstance, they can directly shrink the distance between paired features to learn the common feature space. However, if the changed areas occupy a large proportion, the learning process will be affected, thereby degrading the accuracy of the results.

D. Similarity Measure-Based Methods

They define a modality-independent metric that can be used to distinguish the changed and unchanged areas. In [51], the distance of the sorted histogram was generated within the image, and the dissimilarity between the multimodal images was estimated by this measure. Liu et al. [52] presented a homogeneous pixel transformation method to detect changes between panchromatic and multispectral images. For each pixel in the prechange/postchange image, they estimated its mapping pixel in the postchange/prechange image using the K-nearest neighbor (KNN) algorithm. Touati et al. [58] proposed to utilize a one-order spatial–temporal gradient as the modality-independent metric. Luppino et al. [53] proposed an affinity matrix distance to calculate the change possibility of each pixel, which can be directly used to generate the change map. Also, subsequent methods, such as unsupervised image regression [53] and deep image translation-based methods [41], [42], can be employed to get more accurate detection results. In [54], the self-similarity property of images was introduced for multimodal change detection. The prechange image was transformed into the domain of the postchange image by fractal projection according to self-similarity. Then, the difference image was obtained by comparing the transformed image with the postchange image. Also, exploring self-similarity, Sun et al. [55] presented a patch similarity graph matrix (PSGM) and multiplied multimodal images with the PSGM for image regression. They also applied self-similarity to construct graphs, called NPSG [56], representing the structures for each image, and built the similarity relationships between heterogeneous images. In [57], they designed an iterative robust graph framework to reduce the computational overhead of NPSG and achieve better detection performance. In order to overcome the problem of ignoring useful priors in most existing methods, a graph fusion framework driven by graph signal smoothness representation based on a change prior [62] was proposed in [63] for homogeneous and heterogeneous change detection.

III. PRELIMINARIES

A. Structural Relationship in Multimodal Data

Due to the large difference in modality, it is difficult to accurately detect changes by comparing multimodal remote sensing images directly in the original spectral domain. Therefore, unsupervised multimodal change detection is essentially about finding a domain that makes multitemporal images with modal heterogeneity comparable. In this work, we exploit the local and nonlocal structural relationships in multimodal remote sensing images.

Fig. 1 shows a pair of multimodal remote sensing images and two local areas $A$ and $B$ within them. The prechange image is an SAR image, and the postchange image is an optical image. Although prechange and postchange images show a large modality difference, the relationship between inner pixels of area $A$ should be similar in both prechange and postchange images, as its corresponding land-cover objects are the same. On the other hand, the relationship between a large number of inner pixels of area $B$ cannot retain consistency between prechange and postchange images as a result of the occurring changes in land-cover objects. We refer to this relationship as the local structural relationship.

The second relationship is based on the self-similarity property of images, which has been widely used in image denoising [64], [65], [66]. According to the self-similarity property, for each small area in the image, some similar areas can be founded within the same image. In Fig. 2, for two areas $C$ and $D$ in the prechange image, we can find
some similar areas for them. Since area $C$ is unchanged, the relationship between area $C$ and its similar areas can preserve nearly consistent in the postchange image. However, for area $D$, this relationship cannot be retained because the change event has occurred. We call this relationship as the nonlocal structural relationship.

Since these two relationships are measured within the image, they can be seen as modality-independent. This means that we can use the similarity degree of these two relationships as an indirect measure of the changes. The next questions, then, are how to represent these two relationships in an appropriate model and how to use them to detect changes.

### B. Graph Convolutional Networks

In our method, the graph model is employed to represent structural information contained in multimodal remote sensing images. Then, we present an SR-GCAE to learn the information represented by graphs for multimodal change detection. Thus, we briefly introduce the basic concepts of graph and graph convolutional networks as the preliminaries of our method.

As shown in Fig. 3, a graph is a non-Euclidean data structure consisting of vertices (or nodes) and edges, which can be represented as $G = (V, E)$, where $V$ is the set of vertices with the number $|V| = N$ and $E$ is the set of edges. Let $v_i \in V$ denote a vertex and $e_{ij} = (v_i, v_j) \in E$ denote an edge. Then, the adjacency matrix $A$ can be defined as an $N \times N$ matrix with $A_{ij} = 1$ if $e_{ij} \in E$ and $A_{ij} = 0$ if $e_{ij} \notin E$. Based on $A$, the graph Laplacian matrix $L$ can be defined

$$L = D - A$$

where $D$ is a diagonal matrix of $A$, namely, $D_{ii} = \sum_{j=1}^{N} A_{ij}$.

We can further calculate the normalized Laplacian matrix $\tilde{L}$ for enhancing the generalization ability of the graph

$$\tilde{L} = I - D^{-\frac{1}{2}}AD^{-\frac{1}{2}}$$

where $I$ is the identity matrix.

As an emerging network architecture, graph convolutional networks can effectively handle graph structure data by modeling relations between vertices and have been initially applied in homogeneous change detection [28] and supervised multimodal change detection [32]. Now, there are many variants of graph convolutional networks [67], [68], [69], [70], [71]. Our work follows the standard graph convolutional network proposed in [67], which is essentially a first-order approximation of localized spectral convolution on graphs.

Given a signal $s \in \mathbb{R}^N$ (a scalar for each vertex) and a filter $g_\theta = \text{diag}(\theta)$ parameterized by $\theta \in \mathbb{R}^N$, the spectral convolution of $s$ and $g_\theta$ can be performed by decomposing $s$ on the Fourier domain and then multiplying each frequency by $g_\theta$ as

$$g_\theta * s = U g_\theta U^\top s$$

where $U$ is the matrix of eigenvectors of the normalized Laplacian matrix $\tilde{L} = I - D^{-\frac{1}{2}}AD^{-\frac{1}{2}} = U \Lambda U^\top$. $\Lambda$ is the diagonal matrix of eigenvalues of $\tilde{L}$. $U^\top s$ denotes the graph Fourier transform of $s$. $g_\theta$ can be understood as a function of the eigenvalues of $\tilde{L}$, i.e., $g_\theta(\Lambda)$. However, evaluating (3) requires explicitly calculating the Laplacian eigenvectors, which is not computationally feasible for large-scale graph data. To solve this problem, a feasible way is to approximate the filter $g_\theta$ by the Chebyshev polynomials up to the $K$th order [72]

$$g_\theta * s \approx \sum_{k=0}^{K} \theta_k T_k(\tilde{L}) s$$

where $T_k$ is the Chebyshev polynomials.

Going back to the graph convolutional network involved in our work, it limits $K = 1$ and approximates the largest eigenvalue of $\tilde{L}$ as $\lambda_{\max} \approx 2$ [67]. By doing so, (4) can be further simplified to

$$g_\theta * s \approx \theta \left( I + D^{-\frac{1}{2}} A D^{-\frac{1}{2}} \right) s.$$  

Based on (5), there is the following propagation rule of the graph convolutional layer [67]:

$$H^{(l+1)} = \sigma \left( \tilde{D}^{-\frac{1}{2}} \tilde{A} \tilde{D}^{-\frac{1}{2}} H^{(l)} W^{(l)} \right).$$

Here, $\tilde{A} = A + I$ and $\tilde{D}_{ii} = \sum_{j=1}^{N} \tilde{A}_{ij}$ are the renormalized terms of $A$ and $D$, $H^{(l)}$ and $H^{(l+1)}$ are the input and output of the $l$th graph convolutional layer, $W^{(l)}$ denotes the layer trainable weights, and $\sigma(.)$ is the activation function to introduce nonlinearity for the learned features. For more details about the graph convolutional network involved in this article, please refer to [67].
In this section, we present an unsupervised multimodal change detection framework based on the aforementioned two structural relationships and graph convolutional networks, as shown in Fig. 4. First, multimodal images are preprocessed by geometric alignment and image normalization. Then, we construct graphs to represent the crucial structural information in multimodal images. After that, we present the SR-GCAE and employ it on the constructed graphs to learn structural information. Our framework has two SR-GCAEs. The first SR-GCAE learns the edge information of graphs, which is then used to reflect the similarity degree of the local structural relationship between two images. The second network learns the vertex information of graphs, which is then used for nonlocal structural similarity measurement. Using the deep graph representations learned by two SR-GCAEs, we can measure the change level and generate corresponding difference images. After that, a fusion strategy is proposed to adaptively fuse the two difference images. Finally, a simple postprocessing step based on morphological filtering is used to refine the change map.

**A. Data Preprocessing**

Given a pair of multimodal remote sensing images acquired at times $T_1$ and $T_2$, we denote the prechange image with modality $\mathcal{X}$ as $X \in \mathbb{R}^{H_x \times W_x \times C_x}$ and the postchange image with modality $\mathcal{Y}$ as $Y \in \mathbb{R}^{H_y \times W_y \times C_y}$, where $H_x$, $W_x$, $C_x$ and $H_y$, $W_y$, $C_y$ are the height, the width, and the number of channels of prechange image and postchange image, respectively. The pixels in both images are denoted as $x(h,w,c)$ and $y(h,w,c)$, respectively.

The first step of preprocessing is image alignment. It is the process of aligning two or more remote sensing images of the same scene acquired at different times [73]. Given a pair of multimodal remote sensing images, change detection between the corresponding regions is only meaningful if the two images are geometrically aligned. In our framework, there are four main steps in image alignment: collecting matched point pairs, building a transformation model, transforming the image, and resampling. As the spatial resolutions of multimodal remote sensing images are generally different, the resampling step is to resample images with a relatively higher resolution to the reference image with a relatively lower resolution. With the four steps described above, a given multimodal image pair is geometrically aligned. We denote the coaligned images as $\tilde{X} \in \mathbb{R}^{H \times W \times C_x}$ and $\tilde{Y} \in \mathbb{R}^{H \times W \times C_y}$.

The second step of preprocessing is image normalization. For unimodal change detection, the major purpose of normalization is to eliminate the radiometric difference between multitemporal images caused by different imaging conditions [74]. For multimodal change detection, image normalization is also meaningful because it can assign the value of various dimensions of the input multimodal images to be in a similar range. It is beneficial for the performance of the subsequent steps. Optical and SAR images are the two most common types of images used for change detection tasks. The multimodal images used in the experiments of this work belong to these two types. Therefore, we will only present here the normalization methods for these two types of images. For optical images, we directly normalize their pixel values to the range $[0, 1]$

$$\tilde{x}(h, w, c) = \frac{x(h, w, c) - \min_c}{\max_c - \min_c} \quad (7)$$

where $\max_c$ and $\min_c$ are the maximum and minimum pixel values of the image in the $c$th band.

For SAR images, we first take logarithmic operations to suppress the speckle noise and then normalize the logarithmic transformed image

$$\begin{cases} \tilde{x}^{\log}(h, w, c) = \log(1 + \tilde{x}(h, w, c)) \\ \tilde{x}(h, w, c) = \frac{\tilde{x}^{\log}(h, w, c) - \min_c}{\max_c - \min_c}. \end{cases} \quad (8)$$

The normalized two images are denoted as $\tilde{X} \in \mathbb{R}^{H \times W \times C_x}$ and $\tilde{Y} \in \mathbb{R}^{H \times W \times C_y}$.

**IV. METHODOLOGY**

Fig. 4. Overview of unsupervised multimodal change detection based on structural relationship graph representation learning.
B. Structural Graph Construction

The second step is to construct structural graph data to represent the structural information of multimodal remote sensing images. To construct the graph data, the basic analysis unit in our method is not a pixel but a small part of the image. Image patches are an optional and common way to represent the structural information of an image [24], [56], [75]. However, generating image patches by sliding windows is a time-consuming process, and the distribution of land-cover objects is often not square. With these in mind, we apply image segmentation algorithms to obtain image objects (also called superpixels) as the basic units. They can better reflect the image structures than square image patches and are generated in a more efficient way.

In particular, the fractal net evolution approach (FNEA) [76] is employed to get the image objects. Compared to some segmentation methods proposed for RGB images, such as SLIC [77], FNEA has no band number restrictions on the segmentation methods proposed for RGB images, such as is employed to get the image objects. Compared to some image structures than square image patches and are generated superpixels) as the basic units. They can better reflect the segmentation algorithms to obtain image objects (also called

\[
\begin{align*}
&\text{where } f = w_{\text{channel}} h_{\text{channel}} + w_{\text{spatial}} h_{\text{spatial}} < T \quad (9) \\
&\text{where } f \text{ is the merging criteria based on the channel heterogeneity } h_{\text{channel}} \text{ and spatial heterogeneity } h_{\text{spatial}}, \text{ as defined in [76]; } w_{\text{channel}} \text{ and } w_{\text{spatial}} \text{ are the weight of corresponding heterogeneity; and } T \text{ is a preset merging threshold.}
\end{align*}
\]

Due to the different modality and land-cover objects distribution, executing FNEA on \( \tilde{X} \) and \( \tilde{Y} \) separately leads to different segmentation results. In this article, we adopt the cosegmentation strategy. The multimodal image pair is stacked in the channel dimension as a single image to get a unified segmentation result by FNEA. We denote the cosegmentation map as

\[
\begin{align*}
&\Omega = \{\Omega(i) \mid i = 1, 2, \ldots, N_c\} \\
&\Omega(i) \cup \Omega(j) = \emptyset, \quad \text{if } i \neq j \\
&\bigcup_{i=1}^{N_c} \Omega(i) = \{(h, w) \mid h = 1, \ldots, H; w = 1, \ldots, W\}
\end{align*}
\]

where \( N_c \) is the number of objects.

The ith object of \( \tilde{X} \) and \( \tilde{Y} \) can be expressed as \( O^X_i = \{x(h, w, c) \mid h, w \in \Omega(i), c = 1, 2, \ldots, C_X\} \) and \( O^Y_i = \{y(h, w, c) \mid h, w \in \Omega(i), c = 1, 2, \ldots, C_Y\} \), respectively. Based on the generated objects, we can construct the graph data to represent the structural information. For prechange image \( X \), the structural information contained in \( O^X_i \) can be represented by a graph as

\[
\begin{align*}
&\mathcal{G}_{O^X_i} = \{\mathcal{V}_{O^X_i}, \mathcal{E}_{O^X_i}\} \\
&\mathcal{V}_{O^X_i} = \{\tilde{x}(h, w) \mid (h, w) \in \Omega(i)\} \\
&\mathcal{E}_{O^X_i} = \{\{\tilde{x}(h, w), \tilde{x}(m, n)\) \mid (h, w), (m, n) \in \Omega(i)\}
\end{align*}
\]

where \( \mathcal{V}_{O^X_i} \) is the set of feature vectors of pixels in \( O^X_i \) and \( \mathcal{E}_{O^X_i} \) is the set of edge between two pixels in \( O^X_i \). Here, we assume that the local structural relationship exists between any two pixels within an object. Thus, \( \mathcal{G}_{O^X_i} \) is designed as a fully connected graph, where an edge exists between any two vertices for connection. To quantitatively measure the relationship between pixels within \( O^X_i \), we further construct an adjacency matrix \( A_{O^X_i} \) as

\[
A_{O^X_i} = \{\exp(-\phi_1 \text{dist}^X(h, w), (xi(m, n))) | (h, w), (m, n) \in \Omega(i)\} \quad (12)
\]

where \( \phi_1 \) is the bandwidth parameter and \( \text{dist}^X(\cdot, \cdot) \) means a metric measuring the distance between two pixels in modality \( X \).

C. Learning Structural Relationship Graph Representations

After constructing structural graphs for prechange and postchange images, we can exploit the two structural relationships for change detection. A feasible way is to refer to the processing flows in these similarity measure-based methods [54], [57]. However, these methods only use the low-level information of images, and the structural information contained in the graphs is not fully explored, which is inadequate to cope with the complex conditions in multimodal change detection. Therefore, we present an SR-GCAE to fully learn robust and representative information from structural graphs, as shown in Fig. 5.

Given a structural graph \( \mathcal{G}_{O^X_i} \) of prechange image, SR-GCAE learns the structural information in \( \mathcal{G}_{O^X_i} \) with several stacked graph convolutional layers as

\[
\begin{align*}
&\mathcal{F}^{(L)}_{O^X_i} = \mathcal{F}(\mathcal{G}_{O^X_i}) \\
&= \text{GCL}^{(L)}\left(\text{GCL}^{(L-1)}\left(\ldots, \text{GCL}^{(1)}\left(\mathcal{G}_{O^X_i}\right)\right)\right) \ldots
\end{align*}
\]

where \( L \) is the number of graph convolutional layers; \( \mathcal{F}^{(L)}_{O^X_i} \in \mathbb{R}^{N_{Ax} \times C_F} \) is the deep graph representations with \( C_F \) channels learned by graph convolutional layers; and \( \text{GCL}^{(i)} \) is the \( i \)th graph convolutional layer, learning representations following the rule:

\[
\mathcal{F}^{(L+1)}_{O^X_i} = \sigma\left(\mathcal{D}_{O^X_i}^{-1/2} \tilde{A}_{O^X_i} \mathcal{D}_{O^X_i}^{-1/2} \mathcal{F}^{(L)}_{O^X_i} W^{(L)}\right).
\]

After learning \( \mathcal{F}^{(L)}_{O^X_i} \), the decoder of SR-GCAE reconstructs the information of \( \mathcal{G}_{O^X_i} \) from \( \mathcal{F}^{(L)}_{O^X_i} \). The reconstruction process can be formulated as

\[
\hat{\mathcal{G}}_{O^X_i} = \mathcal{F}^{-1}\left(\mathcal{F}^{(L)}_{O^X_i}\right)
\]

where \( \hat{\mathcal{G}}_{O^X_i} \) is the reconstructed information of \( \mathcal{G}_{O^X_i} \).
be achieved by reconstructing the vertex information and edge information. This aim can be encoded in $F_{O_{E}^1}$.

Through optimizing $\mathcal{L}$ and $\mathcal{L}_{\text{ver}}$, we only adopt a simple decoder with one graph convolutional layer to reconstruct the vertex information. An intuitive idea for this design is that the weak reconstruction ability of the decoder can enforce the encoder to learn more representative semantic features.

The local structural relationship is the relationship between pixels within an area. Thus, we reconstruct the edge information from $F_{O_{E}^1}$ directly as follows:

$$
\hat{AO} = \mathcal{F}^{-1} \left( F_{O_{E}^1} \right) = \sigma \left( F_{O_{E}^1} \left( F_{O_{E}^1}^T \right)^T \right)
$$

(18)

where $\hat{AO}$ is the reconstructed adjacent matrix of $G_{O_{E}^1}$ and $\mathcal{L}_{\text{ver}}$ is the optimization objective for reconstructing vertex information. Through optimizing $\mathcal{L}_{\text{ver}}$, the vertex information can be encoded in $F_{O_{E}^1}$. We denote it as $F_{O_{E}^1}^{\text{ver}}$. Note that, different from the symmetric structure commonly seen in the standard autoencoder and autoencoder-based work [50], [79], [80], we only adopt a simple decoder with one graph convolutional layer to reconstruct the vertex information.
D. Change Information Mapping

After learning \(F^{eg}_{O^x_i} \), \(F^{ver}_{O^x_i} \), \(F^{eg}_{O^y_i} \), and \(F^{ver}_{O^y_i} \) from \(X \) and \(Y \), we can perform change detection using the rich robust structural information contained in these features.

If change event happens in the area of \(\Omega_i \), the structural relationship between pixels in \(\Omega_i \) cannot preserve consistent in \(O^x_i \) and \(O^y_i \). Thus, this change can be reflected in \(F^{eg}_{O^x_i} \) and \(F^{eg}_{O^y_i} \). Therefore, the intuitive and simple idea is to calculate the distance between \(F^{eg}_{O^x_i} \) and \(F^{eg}_{O^y_i} \)

\[
d_{\Omega_i}^{\text{cl}} = \text{dist}\left(F^{eg}_{O^x_i}, F^{eg}_{O^y_i}\right)
\]

(19)

where \(\text{dist}(F^{eg}_{O^x_i}, F^{eg}_{O^y_i}) \) is the distance between deep edge representations \(F^{eg}_{O^x_i} \) and \(F^{eg}_{O^y_i} \). We apply \(L_1\)-distance here, i.e.,

\[
\text{dist}\left(F^{eg}_{O^x_i}, F^{eg}_{O^y_i}\right) = \frac{1}{N_{\Omega_i}} \sum_{j=1}^{N_{\Omega_i}} \left| F^{eg}_{O^x_i}(j) - F^{eg}_{O^y_i}(j) \right|
\]

(20)

The local difference image \(DI^{\text{kl}} \in \mathbb{R}^{H \times W} \) can be obtained by assigning \(d_{\Omega_i}^{\text{cl}} \) to the specific pixels according to the segmentation map \(\Omega \)

\[
DI^{\text{kl}}(h, w) = d_{\Omega_i}^{\text{cl}}
\]

(21)

where \((h, w) \in \Omega_i, i = 1, 2, \ldots, N_{\text{cls}}\).

If a change event happens in the area of \(\Omega_i \), the structural relationship between \(\Omega_i \) and its similar objects is not consistent in \(X \) and \(Y \). Accordingly, we further construct a nonlocal structural graph for \(O^x_i \) by finding its most similar \(K \) objects and calculating their similarities based on deep graph representations to represent the nonlocal structural relationship as

\[
\begin{align*}
G^{X}_{O^x_i} &= \{\gamma^{X}_{O^x_i}, \varphi^{X}_{O^x_i}, A^{X}_{O^x_i}\} \\
\gamma^{X}_{O^x_i} &= \{F^{ver}_{O^x_i}, F^{ver}_{O^y_i}, k = 1, 2, \ldots, K\} \\
\varphi^{X}_{O^x_i} &= \{F^{ver}_{O^x_i}, F^{ver}_{O^y_i} \in \gamma^{X}_{O^x_i}\} \\
A^{X}_{O^x_i} &= \exp(-\phi \text{ dist}^{X}(\gamma^{X}_{O^x_i}, \gamma^{X}_{O^y_i}))|F^{ver}_{O^x_i}, F^{ver}_{O^y_i} \in \varphi^{X}_{O^x_i}|
\end{align*}
\]

(22)

where \(G^{X}_{O^x_i} \) is the nonlocal structural graph constructed for \(O^x_i \), and \(A^{X}_{O^x_i} \) is the \(k\)th object that is most similar to \(O^x_i \).

In postchange image \(Y \), we can also construct such a nonlocal graph \(G^{Y}_{O^y_i} = \{\gamma^{Y}_{O^y_i}, \varphi^{Y}_{O^y_i}, A^{Y}_{O^y_i}\} \) based on the nonlocal structural relationship found for \(O^y_i \). Then, we can compare the difference between \(G^{X}_{O^x_i} \) and \(G^{Y}_{O^y_i} \) to determine whether change event happens in the area \(\Omega_i \) or not

\[
d_{\Omega_i}^{\text{nlcl}} = \frac{1}{K} \sum_{k=1}^{K} \sum_{c=1}^{C_{\text{cls}}} \left\lvert \exp(-\phi_2 \text{ dist}^{X}(\|F^{ver}_{O^x_i}(c)\|_p, \|F^{ver}_{O^y_i}(c)\|_p)) - \exp(-\phi_2 \text{ dist}^{Y}(\|F^{ver}_{O^x_i}(c)\|_p, \|F^{ver}_{O^y_i}(c)\|_p)) \right\rvert
\]

(23)

where \(K \) is the number of the most similar objects found for \(O^x_i \) and \(C^F \) is the channel number of \(F^{ver}_{O^x_i} \). Since the dimension of \(F^{ver}_{O^x_i} \) and \(F^{ver}_{O^y_i} \) could be different, we cannot calculate the distance between them directly. To overcome this problem, the \(p\)-norm is first employed on each channel of \(F^{ver}_{O^x_i} \) and \(F^{ver}_{O^y_i} \) before calculating \(\text{dist}^{X}(-,-) \). The rationale for this operation is that the message-passing step of graph convolution enforces neighboring vertices to get similar representations [67], [70]. Common values of \(p \) are one, two, and infinity. We use the \(L_1\) norm in this article.

The above step can also be executed by finding the most similar objects for \(O^y_i \) and mapping this relationship to \(X \)

\[
d_{\Omega_i}^{\text{nlcl}} = \frac{1}{K} \sum_{k=1}^{K} \sum_{c=1}^{C_{\text{cls}}} \left\lvert \exp(-\phi_2 \text{ dist}^{Y}(\|F^{ver}_{O^y_i}(c)\|_p, \|F^{ver}_{O^y_i}(c)\|_p)) - \exp(-\phi_2 \text{ dist}^{X}(\|F^{ver}_{O^y_i}(c)\|_p, \|F^{ver}_{O^y_i}(c)\|_p)) \right\rvert
\]

(24)

After these two steps, the nonlocal difference image can be denoted as

\[
DI^{\text{nlcl}}(h, w) = d_{\Omega_i}^{\text{X-Y}} + d_{\Omega_i}^{\text{Y-X}}
\]

(25)

where \((h, w) \in \Omega_i, i = 1, 2, \ldots, N_{\text{cls}}\).

Once \(DI^{\text{kl}} \) and \(DI^{\text{nlcl}} \) are obtained, we can fuse them to get a more robust difference image. Compared to simply adding two difference images together, it is more appropriate to let the difference image with higher quality take a greater weight in the fusion process. In this article, we present an effective adaptive fusion strategy to fuse \(DI^{\text{kl}} \) and \(DI^{\text{nlcl}} \)

\[
DI^{\text{final}} = \text{Fuse}(DI^{\text{kl}}, DI^{\text{nlcl}}) = \frac{\mathcal{V}(DI^{\text{kl}})DI^{\text{kl}} + \mathcal{V}(DI^{\text{nlcl}})DI^{\text{nlcl}}}{\mathcal{V}(DI^{\text{kl}}) + \mathcal{V}(DI^{\text{nlcl}})}
\]

(26)

where \(DI^{\text{final}} \) is the final difference image and \(\mathcal{V}(DI) \) is the variance of change intensity in \(DI \) calculated as

\[
\mathcal{V}(DI) = \frac{1}{HW} \sum_{h=1}^{H} \sum_{w=1}^{W} (DI(h, w))^2
\]

(27)

The idea of this adaptive fusion strategy is very simple and straightforward: changed and unchanged pixels are very discriminative in the difference image with high quality. This high discrimination would be reflected in the variance of the change intensity of the difference image.

E. Morphological Postprocessing

After \(DI^{\text{final}} \) is generated, the change detection problem can be treated as a binary classification problem. The threshold segmentation methods can be performed on \(DI^{\text{final}} \) to map pixels to change class \(\omega_c \) and nonchange class \(\omega_{nc} \)

\[
CM(h, w) = \begin{cases} 
\omega_{nc}, & DI^{\text{final}}(h, w) \leq T \\
\omega_c, & DI^{\text{final}}(h, w) > T 
\end{cases}
\]

(28)


**Algorithm 1 Process of Unsupervised Multimodal Change Detection via SR-GCAE**

**Require:**
- Pre-change image $X$ and Post-change $Y$ with different modalities;

**Ensure:**
- The difference image $D_I^{\text{final}}$ and binary change map $\widehat{CM}$;

1: Align $X$ and $Y$ geometrically;
2: Normalize $X$ and $Y$ according to their modality;
3: Construct structural graphs $G_{O^x}$ and $G_{O^y}$;
4: Learn graph representations $F_{O^x}$ and $F_{O^y}$ via SR-GCAE with the optimization objective of edge reconstruction:
   $$L_{\text{edge}} = \frac{1}{N_{c_x}} \sum_{i=1}^{N_{c_x}} \frac{1}{N_{a_x}} \sum_{m=1}^{N_{a_x}} \sum_{n=1}^{N_{a_x}} (\hat{A}_{O^x}(m, n) - A_{O^x}(m, n))^2;$$
5: Learn graph representations $F_{O^x}$ and $F_{O^y}$ via SR-GCAE with the optimization objective of vertex reconstruction:
   $$L_{\text{ver}} = \frac{1}{N_{c_x}} \sum_{i=1}^{N_{c_x}} \frac{1}{N_{a_x}} \sum_{n=1}^{N_{a_x}} (\hat{V}_{O^x}(n) - V_{O^x}(n))^2;$$
6: Calculate local similarity difference image $D_I^{\text{lcl}}$ using $F_{O^x}$ and $F_{O^y}$;
7: Calculate nonlocal similarity difference image $D_I^{\text{ncl}}$ using $F_{O^x}$ and $F_{O^y}$;
8: Fuse local and nonlocal difference images to get the final difference image:
   $$D_I^{\text{final}} = \text{Fuse}(D_I^{\text{lcl}}, D_I^{\text{ncl}});$$
9: Perform threshold segmentation method to get the binary change map $\widehat{CM}$;
10: Perform morphological filtering to refine the change map:
   $$\widehat{CM} = \text{MF}(\widehat{CM}, K_c, K_o);$$
11: return $D_I^{\text{final}}$ and $\widehat{CM}$;

where $CM$ is the binary change map and $T$ is the threshold obtained by threshold segmentation methods, such as Otsu’s method [81] and expectation maximization [82].

In addition, the difficult conditions of unsupervised multimodal change detection often cause inaccurate detection pixels. Therefore, applying some postprocessing methods to refine the detection results is necessary. Some approaches choose probability graph models for postprocessing [29], [53], [57]. However, the process of optimizing such models is often complex and computationally intensive. To keep our framework simple and practical, morphological filtering is applied as our postprocessing method.

Specifically, a close operation is first performed to fill possible voids within changed areas

$$\widehat{CM} = CM \bullet K_c = (CM \oplus K_c) \ominus K_c$$ (29)

where $\bullet$ is the close operator composed of a dilation operator $\oplus$ and an erosion operator $\ominus$, and $K_c$ is the structuring element (filtering kernel) of the close operator.

Then, an open operation is performed to erase those isolated changed pixels

$$\widehat{CM} = CM \circ K_o = (CM \ominus K_o) \circ K_o$$ (30)

where $\circ$ is the open operator, $K_o$ is the structuring element of the open operator, and $\widehat{CM}$ is the final refined change map.

The above process can be formed as

$$\widehat{CM} = \text{MF}(CM, K_c, K_o) = (CM \bullet K_c) \circ K_o.$$ (31)

Summarizing all of the aforementioned contents, the overall framework of our SRGRL-CD is elaborated on in Algorithm 1.

**V. EXPERIMENT**

**A. Data Description**

To verify the effectiveness of our method, five heterogeneous and one homogeneous change detection datasets are used in the experimental part.

The first dataset is the Shuguang dataset, consisting of an SAR image and an optical aerial image. Fig. 6(a)–(c) shows the two images and the reference map, respectively. The preprocessed SAR and optical images have a size of 921 $\times$ 593 pixels and were acquired in 2008 and 2012, respectively, covering a part of a village in Shandong province, China.

The second dataset is the River dataset, consisting of an SAR image and a panchromatic image with a size of 291 $\times$ 343 pixels. The prechange SAR image was captured by Radarsat-2 at the Yellow River Estuary in 2008. The postchange panchromatic image was captured by Landsat-7 in September 2010. The main change event in this dataset was bank erosion caused by flooding. Fig. 7(a)–(c) shows the two images and the reference map, respectively.

The third dataset called the Farmland dataset has two SAR images with the same size of 306 $\times$ 291 pixels, as shown in Fig. 8(a)–(c). These two SAR images were captured by Radarsat-2 in 2008 and 2009, respectively, covering farmland along the Yellow River in China. Although the two images...
TABLE I
INFORMATION OF SIX MULTIMODAL CHANGE DETECTION DATASETS

| Dataset   | Sensor                     | Size          | Location         | Change Event   |
|-----------|----------------------------|---------------|------------------|----------------|
| Shuguang  | Radarsat-2/Google Earth    | 593×921×1/3   | Dongying, China  | Constructions  |
| River     | Radarsat-2/Landsat-7       | 291×343×1/1   | Yellow River, China | River flood    |
| Farmland  | Radarsat-2 (single/four looks) | 306×291×1/1 | Eastern China    | Farm land reclaim |
| Gloucester| SPOT/NDVI                  | 990×554×3/1   | Gloucester, England| River flood    |
| Texas     | Landsat-5/EO-1 ALI        | 1534×808×7/10 | Texas, USA       | Forest fire    |
| Hanyang   | GaoFen-2/Gaofen-2         | 1000×1000×4/4 | Wuhan, China     | Urban constructions / water bloom |

Fig. 8. Farmland dataset. (a) Prechange SAR image (SAR image with one look). (b) Postchange SAR image (SAR image with four looks). (c) Reference map.

Fig. 9. Gloucester dataset. (a) Prechange multispectral image. (b) Postchange NDVI image. (c) Reference map.

were imaged by the same sensor, the prechange image and postchange image are single-look and four-look, respectively, thereby showing different modalities.

The fourth dataset is the Gloucester dataset. As shown in Fig. 9, the prechange image was imaged by SPOT, and the postchange image is a normalized difference vegetation index (NDVI) image. Both images cover the Gloucester area in the U.K. before and after a flooding event, respectively, with a size of 990 × 554.

The fifth dataset is composed of two multispectral images, as shown in Fig. 10(a)–(c). The prechange image was captured by Landsat-5 with seven bands, and the postchange image was captured by EO-1 ALI with ten bands. Both images have a size of 1534 × 808 pixels, showing the changes in a forest area in Texas, USA, caused by wildfire.

Besides, we also perform the experiments on a homogeneous dataset, the Hanyang dataset, to further prove the generalization of our method in homogeneous change detection. The two images of the Hanyang dataset consist of four spectral bands with 1000 × 1000 pixels, and they have a spatial resolution of 4 m/pixel. The pseudocolor images and the reference map are shown in Fig. 11.

We could see that these six multimodal change detection datasets have different spatial and spectral resolutions, cover different modality combinations, and reflect different types of change events. It can, therefore, fully evaluate the effectiveness and generalizability of the proposed method in diverse conditions. The information of these six datasets is summarized in Table I.

B. Experiment Settings

The proposed SR-GCAE1 is implemented with the Pytorch library. The Adam optimizer with $1e^{-4}$ learning rate and $1e^{-6}$ weight decay parameters is applied to optimize the

1Source code for this work will be available at https://github.com/ChenHongruixuan/SRGCAE
network. In the training process, the maximum number of epochs is set to 20. Two graph convolutional layers with 16 and 32 convolutional kernels are set in the encoder network. In the step of calculating the similarity degree of nonlocal structural relationship, the number of most similar graphs is set to 50. The Otsu algorithm is selected as the threshold segmentation approach to get the binary change map from the generated difference image.

To demonstrate the superiority of our method in unsupervised multimodal change detection, we compare it with some SOTA approaches. First, we choose six recently proposed methods as comparison methods since they are representative and their code is open-sourced.

1) **RIF** [39]: Recursive inverse filter (RIF) projects the prechange image to the modality of the postchange image by an original concentric circular invariant convolution model, whose parameters are estimated in the least squares sense using a conjugate gradient routine.

2) **CCLMRF** [83]: The Markov model based on a neighborhood adaptive class conditional likelihood (CCLMRF) relies on a spatially adaptive class conditional likelihood. A change detection segmentation step is performed after estimating the parameters of the likelihood model.

3) **M3CD** [29]: The Markov model for multimodal change detection (M3CD) relies on an observation field built up from a pixel pairwise modeling on heterogeneous image pair. It first estimates the likelihood model parameters using the preliminary iterative estimation technique. Then, the change detection map is computed with a stochastic optimization process.

4) **FPMS** [54]: The fractal projection and Markovian segmentation (FPMS)-based method projects the prechange image to the modality of the postchange image by fractal projection. After the projection, pixelwise differencing is performed. The difference image is binarized by an MRF segmentation model.

5) **NPSG** [56]: The nonlocal patch similarity graph (NPSG)-based method builds a graph for each image patch based on the self-similarity and then calculates the change level by mapping the graph structure from one image to the other.

6) **IRGMcS** [57]: The iterative robust graph and Markovian cosegmentation (IRGMcS) method builds a robust KNN graph to represent the structure of each image and compare the graph to measure the change level. The Markovian cosegmentation model is applied to refine the change maps.

Also, we further collect the accuracy obtained by the SOTA methods in each dataset reported in their original paper and list them in Table III. In this way, the superiority of our method can be fully verified.

In the accuracy assessment step, to evaluate the performance of the change maps generated by the proposed approach and comparison methods, the following three commonly used evaluation criteria are employed.

1) **Overall Accuracy (OA):** OA is defined as the number of pixels correctly detected as a percentage of the total number of pixels

\[ \text{OA} = \frac{TP + TN}{TP + TN + FP + FN} \times 100\% \]  

where TP, TN, FP, and FN are the true positives, true negatives, false positives, and false negatives, respectively.

2) **F1 Score (F1):** F1 is defined as the harmonic mean of the precision and recall rates and can be calculated using the following formula:

\[ F1 = \frac{2TP}{2TP + FP + FN} \]  

3) **Kappa Coefficient (KC):** KC is a similarity measurement between the detection result and the change reference map, given by

\[ \text{KC} = \frac{\text{OA} - \text{PRE}}{1 - \text{PRE}} \]  

where PRE is defined as PRE = ((TP + FP) · (TP + FN) + (TN + FN) · (FP + TN))/N² and N is the total number of pixels.

In addition, the empirical receiver operating characteristic (ROC) curve is drawn to evaluate the quality of the difference image obtained by our method. The corresponding area under the curve (AUC) is also calculated as the evaluation criterion.

All experiments are done on a single PC. The CPU used is an Intel Core i7-8750H with a clock rate of 2.2 GHz. The GPU used is a single NVIDIA GeForce GTX 1060.

**C. Change Detection Results**

First, the ROC curves of difference images obtained by our method on the six datasets are plotted in Fig. 13. We could see that the main areas of distribution for all six curves are on the top left of the figure. The AUCs of the six ROC curves are 0.930, 0.980, 0.970, 0.985, 0.991, and 0.935, respectively. These results suggest that the difference images generated by our method have a very good performance. The changed and unchanged areas would be well separated after the execution of the threshold segmentation algorithm, as shown in Fig. 12.

Fig. 12 shows the binary change maps generated by six comparison methods and the proposed method on the six change detection datasets. Due to the mere utilization of low-level information in remote sensing images, the six comparison methods are not robust enough on these six datasets with different modal combinations. For example, many changed and unchanged areas are not detected by IRGMcS on the Texas dataset. In addition, since these comparison methods are only designed for heterogeneous images, they have difficulty in achieving accurate detection results on the homogeneous Hanyang dataset. In comparison, our method obtains accurate change maps with only a few false positive pixels and false negative pixels on all six datasets. Then, Table II lists the OA, F1, and KC of change maps obtained by these different methods. It is clear that SRGRL-CD obtains the best values on the three metrics on each dataset. These results demonstrate its effectiveness and practicality in multimodal change detection.

Moreover, to further evaluate the performance of our framework, we also select SOTA methods for comparison...
Fig. 12. Change maps obtained by different methods on the six change detection datasets. (a) RIF. (b) CCLMRF. (c) M3CD. (d) FPMS. (e) NPSG. (f) IRGMcS. (g) SRGRL-CD. In change maps, white: TPs, red: FPs, black: TNs, and green: FNs.

| Method      | Shugang OA | Shugang F1 | Shugang KC | River OA | River F1 | River KC | Farmland OA | Farmland F1 | Farmland KC | Gloucester OA | Gloucester F1 | Gloucester KC | Texas OA | Texas F1 | Texas KC | Hanyang OA | Hanyang F1 | Hanyang KC |
|-------------|------------|------------|------------|----------|----------|----------|------------|------------|------------|---------------|---------------|---------------|---------|----------|---------|-----------|-----------|-----------|
| RIF [39]    | 96.22      | 0.6699     | 0.6507     | 78.44    | 0.6599   | 0.0157   | 94.60      | 0.6722     | 0.6653     | 84.95        | 0.4742        | 0.3892        | 79.53   | 0.1348   | 0.0200  | 89.85     | 0.6385   | 0.5804    |
| CCLMRF [83] | 92.61      | 0.3937     | 0.3569     | 90.62    | 0.3477   | 0.3137   | 95.86      | 0.7360     | 0.7149     | 86.33        | 0.6036        | 0.5303        | 87.63   | 0.1370   | 0.0869  | 89.01     | 0.6095   | 0.5466    |
| M3CD [29]   | 95.64      | 0.5794     | 0.5569     | 71.17    | 0.1388   | 0.0865   | 95.55      | 0.7164     | 0.6938     | 90.10        | 0.5782        | 0.5221        | 92.97   | 0.6644   | 0.6252  | 91.05     | 0.5514   | 0.5049    |
| FPMS [56]   | 92.43      | 0.5464     | 0.5141     | 87.22    | 0.3150   | 0.2768   | 96.91      | 0.7805     | 0.7644     | 94.21        | 0.7437        | 0.7112        | 86.49   | 0.0824   | 0.0249  | 86.88     | 0.5252   | 0.4500    |
| NPSG [56]   | 95.02      | 0.6387     | 0.6151     | 89.11    | 0.3572   | 0.3220   | 92.45      | 0.5970     | 0.5614     | 79.27        | 0.4380        | 0.3298        | 89.93   | 0.4826   | 0.4273  | 87.94     | 0.5472   | 0.4780    |
| IRGMcS [57] | 98.18      | 0.7791     | 0.7698     | 97.66    | 0.6665   | 0.6644   | 97.86      | 0.8204     | 0.8490     | 93.56        | 0.7402        | 0.7035        | 93.84   | 0.6456   | 0.6137  | 93.38     | 0.6380   | 0.6067    |
| SRGRL-CD    | 98.55      | 0.8239     | 0.8214     | 98.54    | 0.7760   | 0.7685   | 98.70      | 0.8330     | 0.8762     | 96.93        | 0.8756        | 0.8582        | 98.69   | 0.9379   | 0.9306  | 96.89     | 0.8634   | 0.8461    |

**TABLE II**

Accuracy assessment on change maps obtained by different methods on the six change detection datasets. The highest accuracy is highlighted in bold, and the second highest accuracy is underlined.
TABLE III

KC OF DIFFERENT METHODS ON THE SIX MULTIMODAL CHANGE DETECTION DATASETS. THE ACCURACY REPORTED HERE COMES FROM THEIR ORIGINAL PAPERS AND IS LISTED RANKED FROM HIGHEST TO LOWEST. OUR METHOD IS HIGHLIGHTED IN BOLD

| Dataset   | SRGRL-CD | SGRL-CD | pt-CDN [26] | HPT [52] | SRGRL-CD | UIR [53] | KPCA-MNet [20] |
|-----------|----------|---------|-------------|----------|----------|----------|----------------|
| Shuguang  | 0.8214   | 0.7585  | 0.8883      | 0.8883   | 0.8582   | 0.9306   | 0.8461         |
| River     | 0.6595   | 0.7620  | 0.8070      | 0.8762   | 0.5942   | 0.7671   | 0.7556         |
| Farmland  | 0.7320   | 0.6154  | 0.8438      | 0.8526   | 0.7204   | 0.689    | 0.6530         |
| Gloucester| 0.7292   | 0.5064  | 0.8438      | 0.8526   | 0.7204   | 0.689    | 0.6530         |
| Texas     | 0.6595   | 0.7620  | 0.8070      | 0.8762   | 0.5942   | 0.7671   | 0.7556         |
| Hanyang   | 0.6595   | 0.7620  | 0.8070      | 0.8762   | 0.5942   | 0.7671   | 0.7556         |

TABLE IV

GAIN IN KC OF THE DIFFERENT STEPS IN THE PROPOSED FRAMEWORK ON SIX DATASETS. HERE, LSR/NLSR ALONE REFERS TO UTILIZING THE LOW-LEVEL SPECTRAL FEATURES INSTEAD OF DEEP GRAPH REPRESENTATIONS TO COMPUTE LOCAL/NONLOCAL STRUCTURAL RELATIONSHIP

| Method | LSR | NLSR | SR-GCAE | Fuse | MF | Shuguang | River | Farmland | Gloucester | Texas | Hanyang |
|--------|-----|------|---------|------|----|----------|-------|----------|------------|-------|---------|
|        | ✓   | ✓    |         |      |    | 0.4053   | 0.2548 | 0.6811   | 0.3931     | 0.5453 | 0.3173  |
|        | ✓   |      | Lrc     |      |    | 0.7687   | 0.5991 | 0.8349   | 0.6480     | 0.8418 | 0.6927  |
|        |     |      | Lrc     | ✓    |    | 0.4325   | 0.6178 | 0.8270   | 0.5141     | 0.8111 | 0.6550  |
|        | ✓   | ✓    |         | ✓    |    | 0.5962   | 0.7361 | 0.8586   | 0.7599     | 0.8673 | 0.8178  |
|        | ✓   | ✓    |         | ✓    | ✓  | 0.7923   | 0.7466 | 0.8729   | 0.8270     | 0.8950 | 0.8213  |
|        | ✓   | ✓    |         | ✓    | ✓  | 0.8214   | 0.7685 | 0.8762   | 0.8582     | 0.9306 | 0.8461  |

Fig. 13. ROC curves of the difference images generated by SRGRL-CD on different datasets.

and report the KC obtained by these methods in Table III. Among these methods, CACD [50], X-Net [42], ACE-Net [42], SCCN [45], cGAN [38], LTFL [48], DCNet [75], pt-CDN [26], DCNet [84], SARDNN [24], DBN [87], CAAE [41], KPCA-MNet [20], DSMSCN [17], SiamCRNN [31], and DSFANet [10] are deep learning-based methods. It can be seen that our method outperforms almost all SOTA methods on all six datasets. Note that, even compared to some methods specializing in SAR image change detection, like pt-CDN, and optical image change detection, like KPCA-MNet and DSFANet, SRGRL-CD can still show competitive results. The comparison in Table III further demonstrates the superiority of our method.

D. Discussion

To validate our motivation for introducing graph representation learning and the role of each part in the proposed method, we present the contribution of different steps to the final accuracy in Table IV. In this table, LSR and NLSR alone refer to utilizing the low-level spectral information to calculate local and nonlocal structural relationships, respectively. Combined with the specific loss function of SR-GCAE, it means calculating the structural relationship using deep graph representations learned by SR-GCAE. Fuse means the adaptive fusion strategy. MF indicates the morphological filtering-based postprocessing.

First, compared with utilizing low-level information for multimodal change detection, introducing graph representation learning to explore graph information can obviously improve accuracy. For example, in the Shuguang dataset, applying the proposed SR-GCAE to learn vertex and edge information can bring KC increments of 0.3634 and 0.1637 for multimodal change detection. We also visualize the distribution of original data and deep edge representation learned by SR-GCAE on two datasets in Fig. 15. It can be seen that, although the distributions of the multimodal data vary considerably in the original spectral domain, they become close after SR-GCAE models the edge information from structural graphs.

Then, learning local and nonlocal structure relationships are more suitable on different datasets, respectively. For example, learning local structure relationships can achieve better...
Fig. 14. Difference image obtained by learning different graph information on the two datasets. (a) Prechange image. (b) Postchange image. (c) Local structural difference image. (d) Nonlocal structural difference image. (e) Fused difference image.

| Method          | Datasets          |
|-----------------|-------------------|
|                 | Shuguang | River | Farmland | Gloucester | Texas | Hanyang |
| (1, 0)          | 0.7687   | 0.6011 | 0.8349   | 0.6480     | 0.8418 | 0.6927  |
| (0.7, 0.3)      | 0.7909   | 0.6833 | 0.8700   | 0.7389     | 0.8802 | 0.7673  |
| (0.5, 0.5)      | 0.7853   | 0.7420 | 0.8709   | 0.8042     | 0.8948 | 0.8035  |
| (0.3, 0.7)      | 0.6391   | 0.7362 | 0.8683   | **0.8286** | **0.8961** | **0.8189** |
| (0, 1)          | 0.5962   | 0.7106 | 0.8586   | 0.7599     | 0.8673 | 0.8178  |
| Adaptive fusion | **0.7923** | **0.7466** | **0.8729** | **0.8270** | **0.8950** | **0.8213** |

Fig. 15. Distribution comparison of original images and deep edge representation on (a) Shuguang and (b) Texas datasets.

detection results than learning nonlocal ones on the Shuguang dataset. Besides, the number of most similar graphs $K$ will influence the performance of nonlocal structural similarity measurement. Fig. 16 shows the effect of $K$ on the accuracy of change detection. It can be seen that a value of around 50 for $K$ is an appropriate choice.

After fusing the two difference images, the detection accuracy is further improved. Fig. 14 illustrates the local and nonlocal difference images and fused ones. It can be seen that local and nonlocal structural difference images have different
The final step is morphological filtering. Compared to the Markovian model-based postprocessing methods in some approaches, the morphological filtering used in our method is simple and efficient. In the six datasets, morphological filtering can further improve the detection performance. Fig. 17 compares the change maps before and after morphological filtering. It can be seen that some noises are eliminated, and some changed areas become more intact in refined change maps.

Furthermore, the running time of SRGRL-CD is evaluated on three datasets with different scales. We report the running time of the six comparison methods and our approach in Table VI. Note that, in the C++ code of RIF, CCLMRF, M3CD, and FPMS, the resampling algorithm is performed on the input images to reduce the computational overhead. It can be seen that the running time of our method is an order of magnitude smaller than that of the M3CD and NPSG algorithms. However, the remaining four methods are more efficient than our method. This running time is acceptable considering the good performance of SRGRL-CD and the possibility that we can execute our algorithm on more advanced computers and hardware.

VI. CONCLUSION

In this article, we propose an unsupervised multimodal change detection framework based on two structural relationships of multimodal data and graph convolutional networks. In particular, we construct structural graphs to represent the structural information of multimodal images. Then, we propose an SR-GCAE to learn the structural information from constructed graphs. Reconstructing the edge information and vertex information is designed as the optimization objectives of SR-GCAE, respectively, so that the learned features can meet the requirement of change detection. Subsequently, we simultaneously explore the local structural relationship and nonlocal structural relationship in multimodal images and utilize the deep graph representations learned by SR-GCAE to detect land-cover changes based on these two relationships. Once the difference images are obtained, a simple and effective fusion strategy based on the variance of change intensity is proposed to fuse the difference images. Finally, a postprocessing method based on morphological filtering is applied to refine the detection result.

The visual and quantities results on four heterogeneous and one homogeneous change detection datasets show that our method outperforms the other competitors, including traditional and deep-learning-based methods. However, the learning stage of SR-GCAE is a bit time-consuming. Therefore, our further work includes, but is not limited to, reducing the computational overhead of our model to speed up detection.

ACKNOWLEDGMENT

The authors would like to thank the editors and reviewers for their valuable comments, which have helped a lot to improve this article.

REFERENCES

[1] A. Singh, “Review article digital change detection techniques using remotely-sensed data,” *Int. J. Remote Sens.*, vol. 10, no. 6, pp. 989–1003, Jun. 1989.
[2] B. Desclée, P. Bogaert, and P. Deformery, “Forest change detection by statistical object-based method,” Remote Sens. Environ., vol. 102, nos. 1–2, pp. 1–11, 2006.

[3] D. Brunner, G. Lemoine, and L. Bruzzone, “Earthquake damage assessment of buildings using VHR optical and SAR imagery,” IEEE Trans. Geosci. Remote Sens., vol. 48, no. 5, pp. 2403–2420, May 2010.

[4] C. Wu, B. Du, X. Cui, and L. Zhang, “A post-classification change detection method based on iterative slow feature analysis and Bayesian soft fusion,” Remote Sens. Environ., vol. 199, pp. 241–255, Sep. 2017.

[5] H. Luo, C. Liu, C. Wu, and X. Guo, “Urban change detection based on Dempster–Shafer theory for multitemporal very high-resolution imagery,” Remote Sens., vol. 10, no. 7, pp. 20–22, 2018.

[6] C. Wu et al., “An investigation of traffic density changes inside Wuhan during the COVID-19 epidemic with GF-2 time-series images,” Int. J. Appl. Earth Observ. Geoinf., vol. 103, pp. 102503, 2021.

[7] 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), Jul. 2022, pp. 1600–1603.

[8] L. Bruzzone and D. F. Prieto, “Automatic analysis of the difference in multispectral imagery for change detection in heterogeneous optical and SAR remote sensing images,” IEEE Geosci. Remote Sens. Lett., vol. 17, no. 12, pp. 1711–1718, Mar. 2020.

[9] C. Wu, B. Du, and L. Zhang, “Slow feature analysis for change detection in multispectral imagery,” IEEE Trans. Geosci. Remote Sens., vol. 52, no. 5, pp. 2858–2874, May 2014.

[10] B. Du, L. Wu, and L. Zhang, “Unsupervised deep slow feature analysis for change detection in multi-temporal remote sensing images,” IEEE Trans. Geosci. Remote Sens., vol. 57, no. 12, pp. 9976–9992, Dec. 2019.

[11] L. Mou, L. Bruzzone, and X. X. Zhu, “Learning spectral–spatial–temporal features via a recurrent convolutional neural network for change detection in multispectral imagery,” IEEE Trans. Geosci. Remote Sens., vol. 57, no. 2, pp. 924–935, Feb. 2019.

[12] S. Liu, D. Marinelli, L. Bruzzone, and F. Bovolo, “A review of change detection in multitemporal hyperspectral images: Current techniques, applications, and challenges,” IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens., vol. 7, no. 6, pp. 1456–1468, Jun. 2014.

[13] S. Liu, L. Bruzzone, F. Bovolo, and P. Du, “Hierarchical unsupervised change detection in multitemporal hyperspectral images,” IEEE Trans. Geosci. Remote Sens., vol. 53, no. 1, pp. 244–260, Jan. 2015.

[14] S. Liu, L. Bruzzone, F. Bovolo, M. Zanetti, and P. Du, “Sequential spectral change vector analysis for iteratively discovering and detecting multiple changes in hyperspectral images,” IEEE Trans. Geosci. Remote Sens., vol. 53, no. 3, pp. 3465–3478, Apr. 2015.

[15] Q. Wang, Z. Yuan, Q. Du, and X. Li, “GETNET: A general end-to-end 2-D CNN framework for hyperspectral image change detection,” IEEE Trans. Geosci. Remote Sens., vol. 57, no. 1, pp. 3–13, Jan. 2018.

[16] M. Hu, C. Wu, L. Zhang, and B. Du, “Hyperspectral anomaly change detection based on autoencoder,” IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens., vol. 14, pp. 3750–3762, 2021.

[17] H. Chen, C. Wu, B. Du, and L. Wang, “Change detection in multisource VHR images via deep Siamese convolutional multiplexers recurrent neural network,” IEEE Trans. Geosci. Remote Sens., vol. 58, no. 4, pp. 2848–2861, Apr. 2020.

[18] J. Wu et al., “A multiscale graph convolutional network for change detection in homogeneous and heterogeneous remote sensing images,” Int. J. Appl. Earth Observ. Geoinf., vol. 105, Dec. 2021, Art. no. 102615.

[19] 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, Jul. 2019.

[20] 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 Sens., vol. 57, no. 12, pp. 9941–9959, Dec. 2019.

[21] Z.-G. Liu, G. Mercier, J. Dezert, and Q. Pan, “Change detection in heterogeneous remote sensing images based on multidimensional evidential reasoning,” IEEE Geosci. Remote Sens. Lett., vol. 11, no. 1, pp. 168–172, Jan. 2014.

[22] 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, p. 4918, Dec. 2021.

[23] H. Li, M. Gong, M. Zhang, and Y. Wu, “Spatially self-paced convolutional networks for change detection in heterogeneous images,” IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens., vol. 14, pp. 4966–4979, Nov. 2020.

[24] 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, Jan. 2018.

[25] R. Touati, M. Mignotte, and M. Dahmame, “Multimodal change detection in remote sensing images using an unsupervised pixelwise-based Markov random field model,” IEEE Trans. Image Process., vol. 29, pp. 757–767, 2020.

[26] H. Chen, C. Wu, B. Du, L. Zhang, and L. Wang, “Change detection in multitemporal VHR images based on deep Siamese change detection,” ISPRS J. Photogramm. Remote Sens., vol. 185, pp. 16–31, Mar. 2022.

[27] M. Gong, J. Zhao, J. Liu, Q. Miao, and L. Jiao, “Change detection in synthetic aperture radar images based on deep neural networks,” IEEE Trans. Neural Netw. Learn. Syst., vol. 27, no. 1, pp. 125–138, Jan. 2015.

[28] F. Gao, J. Dong, B. Li, and Q. Xu, “Automatic change detection in synthetic aperture radar images based on PCANet,” IEEE Geosci. Remote Sens. Lett., vol. 13, no. 12, pp. 1792–1796, Dec. 2016.

[29] M. Yang, L. Jiao, F. Liu, B. Hou, and S. Yang, “Transferred deep learning for pervasive change detection in remote sensing images,” IEEE Trans. Geosci. Remote Sens., vol. 57, no. 9, pp. 6960–6973, Sep. 2019.

[30] Y. Sun, L. Lei, D. Guan, X. Li, and G. Kuang, “SAR image change detection based on nonlocal low-rank model and two-level clustering,” IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens., vol. 13, pp. 293–306, 2020.

[31] J. Wang, F. Gao, J. Dong, S. Zhang, and Q. Du, “Change detection from synthetic aperture radar images via graph-based knowledge supplementation network,” IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens., vol. 15, pp. 1823–1836, 2022.

[32] R. Touati, M. Mignotte, and M. Dahmame, “Multimodal change detection in remote sensing images using an unsupervised pixelwise-based Markov random field model,” IEEE Trans. Image Process., vol. 29, pp. 757–767, 2020.
Hongruixuan Chen (Student Member, IEEE) received the B.E. degree in surveying and mapping engineering from the School of Resources and Environmental Engineering, Anhui University, Hefei, China, in 2019, and the M.E. degree in photogrammetry and remote sensing from the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan, China, in 2022. He is currently pursuing the Ph.D. degree with the Graduate School of Frontier Science, The University of Tokyo, Chiba, Japan. He was a Trainee with the United Nations Satellite Centre (UNOSAT), United Nations Institute for Training and Research (UNITAR), Geneva, Switzerland. His research fields include deep learning, domain adaptation, and multimodal remote sensing image interpretation and analysis.

Mr. Chen also acts as a Reviewer of eight international journals, e.g., IEEE TRANSACTIONS ON IMAGE PROCESSING (TIP), IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING (TGRS), IEEE GEO- SCIENCE AND REMOTE SENSING LETTERS (GRSL), and IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING (JSTARS).

Naoto Yokoya (Member, IEEE) received the M.Eng. and Ph.D. degrees from the Department of Aeronautics and Astronautics, The University of Tokyo, Tokyo, Japan, in 2010 and 2013, respectively. He was an Assistant Professor with The University of Tokyo from 2013 to 2017. From 2015 to 2017, he was an Alexander von Humboldt Fellow with the German Aerospace Center (DLR), Oberpfaf- fenhofen, Germany, and the Technical University of Munich (TUM), Munich, Germany. Since 2018, he has been with the RIKEN Center for Advanced Intelligence Project, Tokyo, where he leads the Geoinformatics Unit. Since 2020, he has been with The University of Tokyo, where he is currently an Associate Professor. His research is focused on the development of image processing, data fusion, and machine learning algorithms for understanding remote sensing images, with applications to disaster management and environmental assessment.

Dr. Yokoya won the First Place in the 2017 IEEE Geoscience and Remote Sensing Society (GRSS) Data Fusion Contest organized by the Image Analysis and Data Fusion Technical Committee (IADF TC). He was the Chair and the Co-Chair of the IEEE GRSS IADF TC from 2019 to 2021 and 2017 to 2019, respectively, and the Secretary of the IEEE GRSS All Japan Joint Chapter from 2018 to 2021. He was an Associate Editor of the IEEE JOURNAL OF SELECTED TOPICS IN APPLIED EARTH OBSERVATIONS AND REMOTE SENSING (JSTARS) from 2018 to 2021. He has been an Associate Editor of the IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING since 2021.

Chen Wu (Member, IEEE) received the B.S. degree in surveying and mapping engineering from Southeast University, Nanjing, China, in 2010, and the Ph.D. degree in photogrammetry and remote sensing from the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan, China, in 2015. He is currently a Professor with the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University. His research interests include multitemporal remote sensing image change detection and analysis in multispectral and hyperspectral images.

Bo Du (Senior Member, IEEE) received the Ph.D. degree in photogrammetry and remote sensing from the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan, China, in 2010. He is currently a Professor with the School of Computer Science and the Institute of Artificial Intelligence, Wuhan University, where he is also the Director of the National Engineering Research Center for Multimedia Software. He has more than 80 research articles published in the IEEE TPAMI, IEEE TGRS, IEEE GRSL, and 13 of them are ESI hot papers or highly cited papers. His main research interests include pattern recognition, hyperspectral image processing, and signal processing.

Dr. Du regularly serves as a Senior PC Member of the International Joint Conference on Artificial Intelligence (IJCAI) and the Association for the Advancement of Artificial Intelligence. He was a recipient of the IJCAI Distinguished Paper Prize, the IEEE Data Fusion Contest Champion, and the IEEE Workshop on Hyperspectral Image and Signal Processing Best Paper Award in 2018. He served as the Area Chair of the International Conference on Pattern Recognition (ICPR). He won the Highly Cited Researcher 2019 by the Web of Science Group. He serves as an Associate Editor for Neural Networks, Pattern Recognition, and Neurocomputing. He serves as a reviewer for 20 Science Citation Index (SCI) magazines, including IEEE TPAMI, TCYB, TGRS, TIP, JSTARS, and GRSL.