

# An Unsupervised Remote Sensing Change Detection Method Based on Multiscale Graph Convolutional Network and Metric Learning

Xu Tang<sup>ID</sup>, *Member, IEEE*, Huayu Zhang, *Graduate Student Member, IEEE*, Lichao Mou<sup>ID</sup>,  
Fang Liu<sup>ID</sup>, *Member, IEEE*, Xiangrong Zhang<sup>ID</sup>, *Senior Member, IEEE*,  
Xiao Xiang Zhu<sup>ID</sup>, *Fellow, IEEE*, and Licheng Jiao<sup>ID</sup>, *Fellow, IEEE*

**Abstract**—As a fundamental application, change detection (CD) is widespread in the remote sensing (RS) community. With the increase in the spatial resolution of RS images, high-resolution remote sensing (HRRS) image CD tasks receive growing attention. The change information hidden in multitemporal HRRS images could help discover our planet comprehensively. In the current deep learning era, convolutional neural networks (CNNs) have become one of the most powerful tools for a wide range of RS tasks including HRRS image CD, due to their superb feature learning capacity. However, most of them need a large amount of labeled data to accomplish the CD process, which is challenging or even impractical in many RS applications. Also, given the limited valid receptive field, CNNs can only capture short-range context within HRRS images, which is probably not enough to fully explore change information from the images. To overcome these limitations, in this article, we propose an unsupervised CD method, termed GMCD, based on graph convolutional network (GCN) and metric learning. GMCD consists of a Siamese fully convolution network (FCN), a multiscale dynamic GCN (Mlt-GCN), and a pseudolabel generation mechanism based on metric learning. The Siamese FCN contains a Siamese encoder and a pyramid-shaped decoder, aiming to extract multiscale features and integrate them to generate reliable difference

images (DIs). Mlt-GCN focuses on capturing the short- and long-range contextual patterns at feature map level to extract changed and unchanged areas completely. The pseudolabel generation mechanism aims to produce reliable pseudolabels (changed, unchanged, and uncertain) to help accomplish the model training in an unsupervised way. Experiments on four HRRS image CD datasets demonstrate that GMCD outperforms the existing state-of-the-art methods.

**Index Terms**—Change detection (CD), graph convolution network (GCN), high resolution remote sensing (RS) images, metric learning, unsupervised.

## I. INTRODUCTION

CHANGE detection (CD) is an important and basic research topic in the remote sensing (RS) community. It is a process of discovering changed pixels/regions by comparing multitemporal RS images which cover the same locations but are collected at different times. In recent years, with the increase in the type and number of satellites and the development of Earth observation (EO) technologies, a growing number of high-resolution RS (HRRS) images are generated every day. As a useful content interpretation tool, CD draws more and more attention from the community and plays an important role in many applications, such as land cover monitoring [1], disaster assessment [2], and urban planning [3]. However, since HRRS images are complex in contents, diverse in types, and huge in volume, the CD is still a tough and challenging task.

In the last few decades, an ocean of RS image CD methods have been proposed, and they can be divided into supervised and unsupervised models roughly according to whether the labeled change maps are used in the training phase or not [4], [5]. For the supervised methods, the ground truth data are available. Although they can achieve satisfactory results, collecting the labeled data is a time-consuming, laborious, and even impractical task in the RS community [6], which limits the generalization of supervised methods. Therefore, unsupervised RS CD methods receive growing attention.

Traditional unsupervised CD approaches are usually developed based on difference images (DIs). To get useful DIs, many practical algorithms have been proposed, such as principal component analysis (PCA) [7] and slow feature analysis

Manuscript received March 15, 2021; revised June 17, 2021; accepted August 14, 2021. This work was supported in part by the National Natural Science Foundation of China under Grant 61801351, Grant 62171332, Grant 61802190, and Grant 61772400; in part by the Key Research and Development Program of Shaanxi under Grant 2021GY-035; in part by the Key Laboratory of National Defense Science and Technology Foundation Project under Grant 6142A010301; in part by the China Postdoctoral Science Foundation Funded Project under Grant 2017M620441; in part by the Fundamental Research Funds for the Central Universities under Grant 30919011281 and Grant JSGP202101; and in part by the Xidian University Artificial Intelligence School Innovation Fund Project under Grant YJS2115. (*Corresponding author: Xu Tang.*)

Xu Tang, Huayu Zhang, Xiangrong Zhang, and Licheng Jiao are with the Key Laboratory of Intelligent Perception and Image Understanding of Ministry of Education, School of Artificial Intelligence, Xidian University, Xi'an 710071, China (e-mail: tangxu128@gmail.com).

Lichao Mou and Xiao Xiang Zhu are with the Remote Sensing Technology Institute, German Aerospace Center, 82234 Wessling, Germany, and also with the Data Science in Earth Observation (SiPEO, former: Signal Processing in Earth Observation), Technical University of Munich, 80333 Munich, Germany.

Fang Liu is with the Key Laboratory of Intelligent Perception and Systems for High-Dimensional Information of Ministry of Education, School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China.

This article has supplementary downloadable material available at <https://doi.org/10.1109/TGRS.2021.3106381>, provided by the authors.

Digital Object Identifier 10.1109/TGRS.2021.3106381

(SFA) [8]. After getting DI, a change map can be identified via thresholding [9] or clustering [10] schemes. At the very beginning, to make full use of pixels in multitemporal RS images, scholars treated individual pixels within the images as elementary units to design their CD approaches. For example, Chen *et al.* [11] proposed a Markov random field-based approach, which considers contextual information among neighbor pixels to obtain the change map. Nowadays, with the rapid increase in image resolution [12], an increasing number of object-/region-based HRRS image CD methods [13] have been developed, in which homogeneous or heterogeneous regions rather than pixels are regarded as basic units in the CD process. Although the traditional methods get successes in their applications, their behavior is limited by the hand-crafted visual features extracted from HRRS images.

Recently, deep learning techniques, especially convolutional neural networks (CNNs) [14], bring computer vision into a new era. Due to the strong nonlinear fitting capacity and hierarchical structure of CNNs, the learned features can obtain high-level semantics and rich spatial context information simultaneously. Therefore, CNNs have benefited many image processing tasks as diverse as object detection [15], image semantic segmentation [16], and CD [17]. In the RS community, CNNs are also getting popular. Many researchers have used CNNs to develop unsupervised HRRS CD methods, and they achieved impressive results [4]. For example, Saha *et al.* [18] proposed an unsupervised context-sensitive framework, named deep change vector analysis (DCVA), to accurately capture change information through employing deep spatial context to complete change vector analysis. These deep-learning-based methods perform well with deep features, but their CD results have a high false alarm rate in general. The reasons behind this can be attributed as follows [19]. First, in HRRS images, there are many pseudochanges (e.g., shadow and vegetation color change) that negatively impact CD results. Second, many approaches simply deem CNNs as feature extractors, and in this regard, some characteristics of HRRS images are not fully considered.

To enhance the feature representation and consider specific properties (e.g., complex scene and changing areas' chaotic distribution) of HRRS images thoroughly, we propose an unsupervised CD model based on a trained Siamese fully convolutional network (FCN) [20], a multiscale dynamic graph convolutional network (Mlt-GCN), and metric learning. We name it GCN and metric learning-based CD (GMCD) for short.<sup>1</sup> Specifically, the main framework of the proposed model is a Siamese FCN, including a Siamese FCN encoder and a pyramid-shaped decoder. The former aims to extract deep features from HRRS images, and the latter focuses on making full use of multiscale features to predict dual-channel DIs at various scales. Then, Mlt-GCN uses deep features to capture multiple long-range contextual patterns to grasp the relationships among pixels in the HRRS images fully and convey more comprehensive feature information. Meanwhile, to enhance the generalization capacity of our model for

different types of HRRS images and make full use of the extracted features, a novel dynamic pseudolabel generating mechanism is proposed. It combines spatial-spectral feature analysis and metric learning to ensure that the resulting CD maps are satisfactory. Furthermore, training with the joint CD loss can effectively highlight the changed areas and alleviate the problem of pseudochanges. Note that the spatial and spectral features in this article denote properties of a feature map representing the relationships among pixels in the same channel and the correlations among channels, respectively. In addition, the proposed method can handle red, green, and blue (RGB) data or RGB and near-infrared data.

The main contributions of our work can be summarized as follows.

- 1) We propose an unsupervised CD network GMCD. It makes use of GCN and a metric learning algorithm to learn rich contextual information for CD tasks. We find that our method is applicable to a wide variety of imagery types.
- 2) We propose an Mlt-GCN module whose adjacency matrices are generated through an attention diagonalization procedure, eliminating the projecting process and reserves more spatial information. On the one hand, due to the dense connection characteristic of GCN, Mlt-GCN can adequately capture multiple long-range contextual patterns in deep feature maps. On the other hand, the attention diagonalization procedure is able to extract the spatial-spectral feature, which helps generate reliable pseudolabels for unsupervised training.
- 3) A simple yet effective mechanism for generating dynamic pseudolabels is developed. It combines the analysis of spatial-spectral features and metric learning. The former can explore rich spatial and spectral information from HRRS images, while the latter is able to mine the semantic similarity of unlabeled pixels in the images. They can significantly improve the reliability of the pseudolabels and help get promising CD results.
- 4) The comprehensive experiments are conducted on several HRRS datasets, including QuickBird (QB), Zi-Yuan 3 (ZY3) [19], SZADA/2 [21], and Montpellier [22]. The encouraging results demonstrate that our method is effective for the CD task.

The remainder of this article is organized as follows. Section II briefly reviews the related works on the deep-learning-based CD and GCN. In Section III, the proposed CD method is introduced in detail. The experiments and discussion are presented in Section IV. Finally, Section V concludes this article.

## II. RELATED WORK

### A. Deep-Learning-Based CD

The existing deep-learning-based RS CD methods can be divided into two categories: supervised and unsupervised.

In the first category, many successful supervised CD methods have been proposed in recent years. For example, Gong *et al.* [23] introduced a deep-learning-based CD method for synthetic aperture radar (SAR) images. Some traditional

<sup>1</sup>Our source codes are available at <https://github.com/TangXu-Group/Unsupervised-Remote-Sensing-Change-Detection>

approaches and a shallow CNN work together to generate positive CD results. However, the quality of the produced CD maps would be influenced by conventional methods. Also, the developed shallow CNN cannot ensure that the CD results are satisfactory as the speckle noise of SAR images is not fully considered and SAR images' contextual information and semantics are not fully explored. To take more RS image properties (e.g., multiscale) into account, researchers pay more attention to developing various specific CNNs. For instance, Zhan *et al.* [24] presented a deep Siamese convolutional network for RS CD tasks. Due to the particular structure of the Siamese convolutional network, the pixel-wise similarity between the input bi-temporal images can be automatically learned, which helps identify changed pixels in images. Convolutional coupling network [25] is another typical deep model for RS CD tasks. It maps two heterogeneous RS images into a common feature space to estimate the changed information. Apart from diverse CNNs, autoencoders (AEs) [26] and generative adversarial networks (GANs) [27] have also been widely used in RS CD tasks. Through adversarial learning, both rich semantic information and high-quality DIs can be obtained.

Although the supervised deep RS CD methods achieve promising results, they need a large number of labeled samples to train models. It is well known that ground-truth data in the RS community are expensive to acquire. To deal with this issue, a series of unsupervised deep CD methods have been proposed. For example, Jong *et al.* [28] developed an unsupervised CD framework under the paradigm of semantic segmentation. With a specific DI construction approach, any pre-trained CNN model for semantic segmentation can be used for CD tasks. To mine more useful information from RS images for good CD results, some specifically designed CNNs were developed. An unsupervised RS CD method was introduced in [29], in which a noise modeling block is added on top of an FCN-based feature learning module. By modeling the noise within RS images, the method enhances the ability to distinguish noise information and further improves the robustness of CD results significantly. Besides, Chen *et al.* [30] proposed a deep Siamese multiscale convolutional network for RS CD tasks. A multiscale feature convolution unit (MFCU) is designed to extract multiscale information from RS images for obtaining positive CD results. To accomplish the task of CD for polarimetric SAR (POLSAR) Images, Liu *et al.* [31] introduced a local restrict CNN (LRCNN). It takes the local similarity into account and conducts finetune based on the pseudo-labeled pixels obtained from discriminative enhanced layered difference images (DELIDs). Furthermore, Looking-Around-and-Into model [32] combined an attention proposal network and a recurrent CNN for large-scope POLSAR image CD. Hyperspectral image CD has also developed rapidly in recent years with CNN. For example, Yuan *et al.* [33] proposed a robust PCA network through integrating deep feature with the traditional PCA method and Wang *et al.* [34] present an end-to-end 2-D CNN framework to mine cross-channel gradient features and enhance the result and generalization ability of hyperspectral image CD with the features extraction of multisource data. Due to complex contents within HRRS images, not only global but also local information

should be explored, so the visual attention mechanism [35] draws considerable attention from the community. A pyramid feature-based attention-guided Siamese network [36] was proposed to improve CD results by adding a global co-attention model, which emphasizes the importance of the correlation between the input feature pairs. Another popular direction in unsupervised CD is transfer learning [37]. For instance, Yang *et al.* [38] designed a transferred deep-learning-based CD algorithm. The source domain labels can be transferred to the unlabeled target data so that CD results can be obtained in an unsupervised manner. Saha *et al.* [39] proposed an unsupervised CD method based on transfer learning, in which the changed information between SAR and optical images can be captured accurately.

### B. GCNs in RS Image Processing Tasks

In the general CNN model, convolutional kernels only convey the regular structured area of data, which cannot fully reflect the context information hidden in the data. To solve the drawback, GCN is proposed, which can build the connection between data and capture global structure information via message propagation [40]. In the beginning, GCN is widely used to handle tasks involving unstructured data [41], such as text classification [42], network architecture search [43], and 3-D point cloud classification [44]. Afterward, due to the outstanding performance of diverse GCNs, they are becoming popular in image processing. The pixels/regions within images are regarded as graph nodes, and then the local and global information of the images can be captured simultaneously [45]. For example, Joan *et al.* [46] proposed a deep locally connected network based on the spectrum of graph Laplacian to recognize image and audio data. For semantic segmentation tasks, GCN has also been utilized [47], in which the multilayer graph structure and features of nodes can be effectively learned for extracting adequate deep features to improve segmentation results. In the RS community, GCN is also widely used. Comparing with natural images, HRRS images have various targets with diverse scales. It is a tough task to mine relationships hidden in the complex content. Thanks to the fact that GCN can explicitly model multiple long-range contextual patterns in HRRS images, we can make use of it to understand the content of an HRRS image in a semi-supervised or unsupervised way. For instance, Khan *et al.* [48] presented a multilabel RS scene recognition method with the help of GCN. This method can extract discriminative features from RS images using the region adjacency graph (RAG), which can explore the true semantics of RS scenes and boost the ability of scene recognition. An RS image classification algorithm was designed by combining CNN and GCN in [49], where CNN aims at extracting deep spatial features from images, and GCN focuses on capturing dependencies among diverse objects. In this way, the visual information and spatial locations can be fully used to produce satisfactory results. Chaudhuri *et al.* [50] introduced a Siamese graph convolutional network (SGCN) for RS image retrieval tasks. The resemblance between two images can be obtained by measuring the similarity between their corresponding graphs.



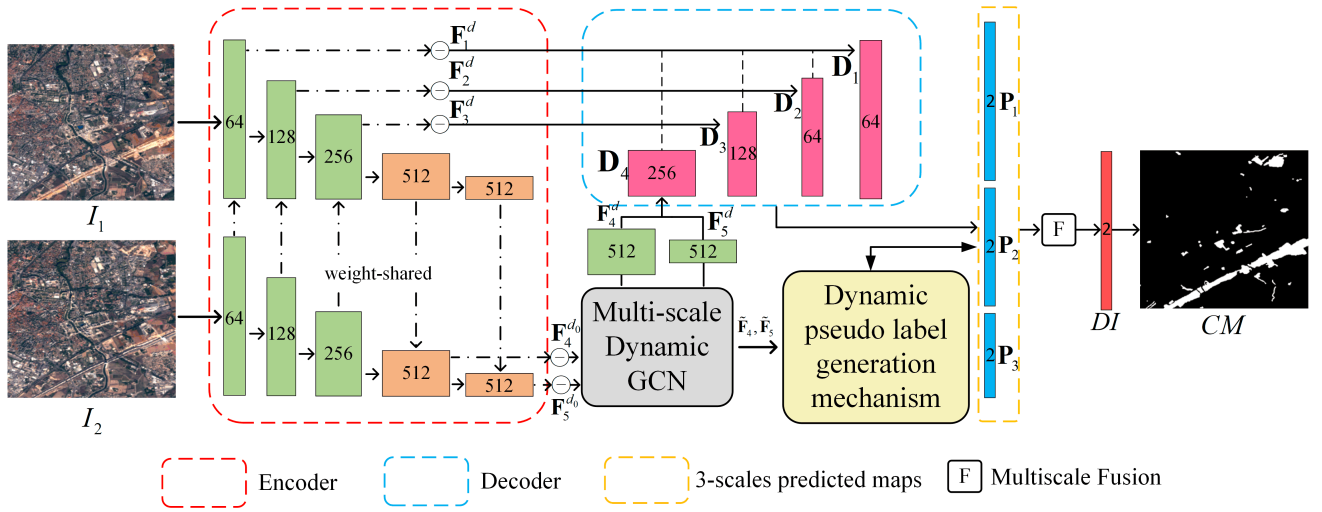


Fig. 1. Architecture of GMCD, which consists of a siamese FCN encoder, a pyramid-shaped decoder, a multiscale dynamic GCN, and a dynamic pseudolabel generation mechanism.

Besides the scene-level tasks mentioned above, pixel-level tasks can also be solved by GCN. For example, You *et al.* [51] introduced a model constructed by a sliced recurrent neural network (SRNN) and an attention-treated GCN. In this model, the GCN with attention mechanism makes full use of deep features and contextual semantics to accomplish pixel-level RS image recognition. For hyperspectral image (HSI) classification tasks, a mini-batch GCN method [52] was designed. The large-scale GCN is constructed to represent HSI at pixel-level, and the mini-batch scheme is developed to train the large-scale GCN with low computational costs. Wan *et al.* [53] presented a context-aware dynamic GCN for HSI classification. The parcels obtained by a super-pixel segmentation algorithm are regarded as graph nodes, and relationships among the parcels can be updated dynamically through graph convolutions. The result of this approach is competitive in HSI classification. Although GCN is applied to many RS applications, the number of GCN-based RS CD methods is few. Fortunately, researchers are paying attention to this field. For example, Saha *et al.* [54] proposed a semi-supervised CD method with a multilayer GCN, which can obtain CD results by exploring multiscale information deeply. Compared with other conventional semi-supervised methods, the utilization of GCN pushes this model to achieve superior CD maps.

### III. METHODOLOGY

#### A. Framework of Our Model

Our model is developed under the encoder-decoder paradigm, and its flowchart is shown in Fig. 1. It consists of a pre-trained Siamese FCN encoder, a pyramid-shape decoder, and an Mlt-GCN model. Suppose there is a pair of temporal HRRS images  $I_1 \in \mathbb{R}^{H \times W \times C}$  and  $I_2 \in \mathbb{R}^{H \times W \times C}$  that have been pre-processed by some common operations, such as image registration [55], radiometric relative normalization [56], and pansharpening [57]. When they are fed into our model, the trained Siamese FCN encoder is used to extract the deep spatial features from them. Then, the followed multiscale dynamic GCN is used to enrich them by capturing nonlocal

and spectral information. Next, the pyramid-shaped decoder integrates the resulting feature maps obtained by the encoder and multiscale dynamic GCN to generate dual-channel predicted DIs with three scales. Besides, to accomplish the CD task in an unsupervised manner, a dynamic pseudolabel generation mechanism, which contains the spatial-spectral feature analysis and metric learning, is developed to get the reliable and effective pseudolabels using the spatial-spectral features extracted from Mlt-GCN. Afterward, the pseudolabels and multiscale DIs are used to train our model with a specific CD loss function. When our model is trained, the multiscale DIs are fused for the final DI and change map.

#### B. Siamese Fully Convolutional Network

To fully extract multiscale spatial information from HRRS images, we proposed a Siamese encoder and a pyramid-shaped decoder.

The main structure of the Siamese down-sampling encoder is a dual-branch weight-shared FCN, which is made up of the first five convolution blocks of visual geometry group (VGG) 16. For the first two blocks, each of them consists of two convolution layers. The other three blocks are composed of three convolution layers. Besides, there are four max-pooling layers embedded between convolution blocks, which are used to reduce the resolution of the input image pairs. On account of the fact that our task is unsupervised, the proposed FCN is pre-trained on a building extraction dataset [58]. From the encoder, we can obtain five difference feature maps (DFMs)  $\{F_1^d, F_2^d, F_3^d, F_4^d, F_5^d\}$  with various scales through performing element-wise differencing on each scale. Note that,  $F_4^d$  and  $F_5^d$  will further be input to Mlt-GCN to generate new feature representations  $F_4^{d0}$  and  $F_5^{d0}$ . Note that the selection of the building extraction dataset is not a limitation of our method. Some other datasets, such as the Northwestern Polytechnique University (NWPU) very high resolution (VHR)-10 dataset [59], can be chosen to generate the pre-trained parameters. Furthermore, the effect of pre-trained parameters is discussed in the supplementary material.



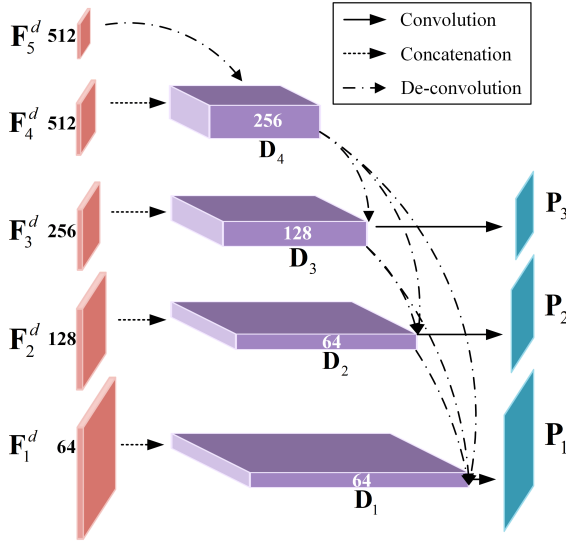


Fig. 2. Framework of pyramid-shaped decoder. It takes 5-scale feature maps generated by encoder as input and generates dual-channel predicted DIs with three different scales.

To further explore the multiscale information for our task, we make use of a pyramid-shaped decoder. The specific structure of the pyramid-shaped decoder is exhibited in Fig. 2. The five DFM s  $\{F_1^d, F_2^d, F_3^d, F_4^d, F_5^d\}$  can be fused up through top-down dense connections which are good at integrating semantic information. In specific, for each block in the decoder, the corresponding DFM (except  $F_5^d$ ) is channel-wisely concatenated with feature maps from the previous decoder blocks which have been processed by deconvolution (i.e., up-sampling, convolution, and dropout) operations. Then, the output is convolved by a  $1 \times 1$  kernel with a stride of  $1 \times 1$ . This process can be formulated as

$$\begin{aligned} D_i &= f_{\text{Conv}}(\text{concat}(F_i^d, D_{\text{prev}}), W_{\text{Conv}}) \\ D_{\text{prev}} &= \text{concat}(f_{\text{DConv}}(D_{i+1}), \dots, f_{\text{DConv}}(D_4)) \\ &\quad \times i = 1, 2, 3, 4 \end{aligned} \quad (1)$$

where  $f_{\text{Conv}}(\cdot)$  and  $f_{\text{DConv}}(\cdot)$  denote the convolution and deconvolution, respectively, and  $\text{concat}(\cdot)$  means the channel-wise concatenation. Finally, the dual-channel predicted DIs  $\{P_1, P_2, P_3\}$  with three scales can be generated using  $\{D_1, D_2, D_3\}$  by a  $1 \times 1$  convolution with a stride of  $1 \times 1$ .

### C. Multiscale Dynamic GCN

Although the FCN encoder could extract multiscale features from HRRS images through a series of convolutions, it can only model short-range relations due to the limited valid receptive field of convolution operations. The long-range relationships within HRRS images are not fully exploited but very important to CD tasks. To overcome this issue, Mlt-GCN is introduced, and its framework is shown in Fig. 3.

Before explaining Mlt-GCN in detail, we first introduce GCN briefly. In general, GCN can be regarded as a generalization of CNN to the graph domain, which significantly boosts mining relations among image features in the spatial domain. In GCN, assume that there are  $N$  nodes, and we

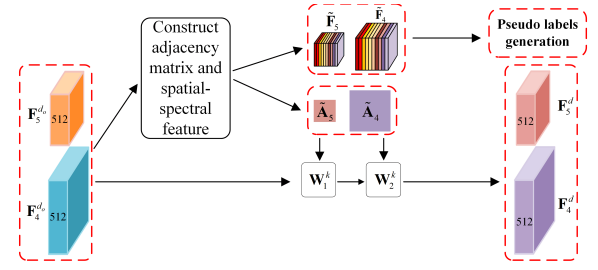


Fig. 3. Process of multiscale dynamic GCN module.

use  $\mathcal{G} = (\mathcal{V}, \mathcal{E})$  to represent an undirected graph, where  $\mathcal{V} = \{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_N\}$  indicates the nodes within  $\mathcal{G}$  and  $\mathcal{E} = \{e_{ij}, i = 1, \dots, N, j = 1, \dots, N\}$  denotes the edges between the nodes of  $\mathcal{G}$ . Also, the adjacency matrix  $\mathbf{A} \in \mathbb{R}^{N \times N}$  is used to describe the weights of edges between each pair of nodes. In general,  $\mathbf{A}$  can be calculated as

$$A = \begin{cases} \exp(-\gamma \cdot \text{dis}(\mathbf{v}_i, \mathbf{v}_j)), & (\mathbf{v}_i, \mathbf{v}_j) \in \mathcal{N} \\ 0, & \text{otherwise} \end{cases} \quad (2)$$

where  $\gamma$  is an empirical parameter,  $\text{dis}(\mathbf{v}_i, \mathbf{v}_j)$  means the distance between nodes  $\mathbf{v}_i$  and  $\mathbf{v}_j$ , and  $\mathcal{N}$  indicates a neighborhood set. To generalize the convolution to graph signals, a degree matrix  $\mathbf{D} \in \mathbb{R}^{N \times N}$  should be computed by  $D_{ii} = \sum_j A_{ij}$  first. Then, the normalized Laplacian matrix  $\mathbf{L} = \mathbf{I} - \mathbf{D}^{-1/2} \mathbf{A} \mathbf{D}^{-1/2}$  can be calculated, where  $\mathbf{I}$  denotes the identity matrix. What is more, the Laplacian matrix can be further improved with a normalization trick, and its definition is

$$\tilde{\mathbf{L}} = \tilde{\mathbf{D}}^{-1/2} \tilde{\mathbf{A}} \tilde{\mathbf{D}}^{-1/2} \quad (3)$$

where  $\tilde{\mathbf{A}} = \mathbf{A} + \mathbf{I}$  and  $\tilde{D}_{ii} = \sum_j \tilde{A}_{ij}$ . By doing so, the propagation of a multilayer GCN can be formulated as

$$\mathbf{H}^{(l+1)} = \sigma(\tilde{\mathbf{D}}^{-1/2} \tilde{\mathbf{A}} \tilde{\mathbf{D}}^{-1/2} \mathbf{H}^{(l)} \mathbf{W}^{(l)}) \quad (4)$$

where  $\sigma(\cdot)$  is the activation function, and  $\mathbf{H}^{(l)}$  and  $\mathbf{W}^{(l)}$  illustrate the outputs and learnable weights of the  $l$ th layer, respectively.

As shown in Fig. 3, the inputs of Mlt-GCN are  $F_4^{d_o} \in \mathbb{R}^{h_4 \times w_4 \times 512}$  and  $F_5^{d_o} \in \mathbb{R}^{h_5 \times w_5 \times 512}$  which contain rich semantic and spatial information. The outputs of Mlt-GCN include two parts. First, two new DFMs  $F_4^d \in \mathbb{R}^{h_4 \times w_4 \times 512}$  and  $F_5^d \in \mathbb{R}^{h_5 \times w_5 \times 512}$  are generated under the paradigm of GCN. Unlike many existing GCN models, which regard parcels of images (which can be obtained by an over-segment algorithm [60]) as graph nodes, our Mlt-GCN model treats feature-map vectors in  $F_4^{d_o}$  and  $F_5^{d_o}$  as graph nodes. Each feature-map vector corresponds to a region within the input HRRS images. Both short- and long-range relationships among diverse regions in the input images can be explored by mining relations among feature-map vectors within DFMs. Here, the feature-map vector mentioned above indicates the basic unit corresponding to specific feature maps. Second, two new feature maps  $\tilde{F}_4$  and  $\tilde{F}_5$  are generated by a simple channel-wise attention scheme, consisting of a global average pooling and two  $1 \times 1$  convolutions. In this way,  $\tilde{F}_4$  and  $\tilde{F}_5$  contain both the spatial and spectral

information, which are beneficial to the following pseudolabel generation. Note that the reason why we only feed  $\mathbf{F}_4^{d_0}$  and  $\mathbf{F}_5^{d_0}$  into Mlt-GCN is that compared with  $\{\mathbf{F}_1^d, \mathbf{F}_2^d, \mathbf{F}_3^d\}$  they contain much more semantic information. Besides, processing only  $\mathbf{F}_4^{d_0}$  and  $\mathbf{F}_5^{d_0}$  could reduce the computational cost.

$$\mathbf{F}_k^d = f_{\text{ReLU}}\left(\tilde{\mathbf{L}}_k \cdot f_{\text{ReLU}}\left(\tilde{\mathbf{L}}_k \cdot \mathbf{F}_k \cdot \mathbf{W}_1^k\right) \cdot \mathbf{W}_2^k\right), \quad k = 4, 5 \quad (5)$$

In general, the adjacency matrix can be calculated by (2). However, the proper distance metric and empirical parameter  $\gamma$  are hard to select. Consequently, we develop an adjacency matrix construction method based on nonlocal block [61] to calculate adjacency matrices using  $\mathbf{F}_k$ . The flowchart is exhibited in Fig. 4. First,  $\mathbf{F}_k$  is convolved by a  $3 \times 3$  convolution layer followed by  $\text{ReLU}(\cdot)$  nonlinearity to produce a new feature representation  $\mathbf{F}'_k \in \mathbb{R}^{h_k \times w_k \times 64}$ . This step can decrease the number of parameters in the following computation and simulate the projecting process of traditional GCN, which is beneficial to calculate the similarity between different positions. Then, to measure the relationships among points in the feature map,  $\mathbf{F}'_k$  is reshaped into a pair of matrices  $\phi(\mathbf{F}_k) \in \mathbb{R}^{h_k w_k \times 64}$  and  $\phi(\mathbf{F}_k)^T \in \mathbb{R}^{64 \times h_k w_k}$ , where  $h_k w_k$  means the

$$\tilde{\Lambda}(\mathbf{F}_k) = \text{diag}(\text{Conv3}(\text{Conv2}(\text{GlobalAvgPool}(\mathbf{F}_k)))). \quad (6)$$
$$\widetilde{\mathbf{A}}_k = \phi(\mathbf{F}_k) \times \widetilde{\Lambda}(\mathbf{F}_k) \times \phi(\mathbf{F}_k)^T. \quad (7)$$

For two new feature maps  $\tilde{\mathbf{F}}_4$  and  $\tilde{\mathbf{F}}_5$ , by applying the global average pooling and  $1 \times 1$  convolution operations to  $\mathbf{F}_k$ , a one-dimension vector with the same number of channels as  $\mathbf{F}_k$ ,  $k = 4, 5$  is obtained, which can be seemed like the weight of each channel. Then, the spatial-spectral features  $\tilde{\mathbf{F}}_4$  and  $\tilde{\mathbf{F}}_5$  are generated by assigning weights to the corresponding channels so that the correlations among channels can be inserted into  $\mathbf{F}_k$ ,  $k = 4, 5$ . This process can be expressed as

$$\tilde{\mathbf{F}}_k = f_{\text{scale}}(\mathbf{F}_k, \gamma(\mathbf{F}_k)) \quad (8)$$

#### D. Dynamic Pseudolabel Generation Mechanism

From the Mt-GCN model, we obtain two feature maps  $\tilde{\mathbf{F}}_4 \in \mathbb{R}^{h_4 \times w_4 \times 512}$ ,  $\tilde{\mathbf{F}}_5 \in \mathbb{R}^{h_5 \times w_5 \times 512}$  with spatial-spectral information, where each value ( $v_k(i, j) \in v_k, 0 < i < h_k, 0 < j < w_k, k = 4, 5$ ) indicates the degree of change in the corresponding area. Besides, to further distinguish changed areas from unchanged ones, we divide pixels of the feature maps into three classes according to their values with the analysis of spatial-spectral features. The three classes are un-changed class ( $\omega_n$ ), changed class ( $\omega_c$ ), and uncertain class ( $\omega_u$ ). As described in [63], the probability density function  $p(v_k)$  of each element  $v_k(i, j)$  can be modeled as a Gaussian mixture distribution

$$p(v) = p(v_k|\omega_n)p(\omega_n) + p(v_k|\omega_c)p(\omega_c) + p(v_k|\omega_u)p(\omega_u) \quad (9)$$

$$p(v_k|w_l) = \frac{1}{\sqrt{2\pi}\sigma_{w_l}} \exp\left(-\frac{(v_k - \mu_{w_l})^2}{2\sigma_{w_l}^2}\right) \quad (10)$$

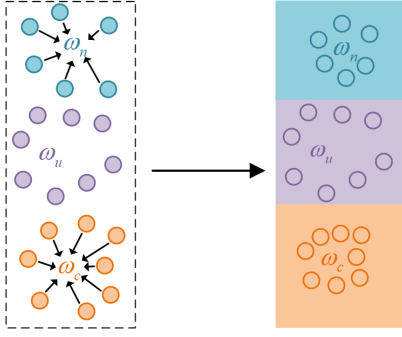


Fig. 5. Mechanism of metric learning network.

where  $l = 3$ , and  $\mu_{\omega_l}$  and  $\sigma_{\omega_l}^2$  are the mean and variance values of the corresponding regions, respectively. As mentioned in [64], the expectation-maximization (EM) algorithm based on the Bayesian decision theory is proposed. Because of the curve fitting capability for discrete values of the EM algorithm, three estimated probability density curves of  $\omega_n$ ,  $\omega_c$  and  $\omega_u$  are presented. As a result, thresholds  $T_1$  and  $T_2$  that separate the three classes can be obtained, which is the intersection of two adjacent curves. Finally, the initial pseudolabels can be generated as follows:

$$\begin{cases} v_{k,n}(i, j) \in \omega_n, & \text{if } v_k(i, j) < T_1 \\ v_{k,u}(i, j) \in \omega_u, & \text{if } T_1 \leq v_k(i, j) < T_2 \\ v_{k,c}(i, j) \in \omega_c, & \text{if } v_k(i, j) \geq T_2. \end{cases} \quad (11)$$

However, the obtained initial pseudolabels are not reliable enough and hard to update during the model training. Therefore, we propose a metric learning mechanism based on a similarity network [65] to solve these problems and improve the credibility of pseudolabels, and its schematic is shown in Fig. 5. The mechanism mainly includes the calculation of the centroid of each class and label assignment. First, for pixels  $(x_k^d(i, j) \in \mathbf{x}_k^d, k = 4, 5)$  of the DFMs ( $\mathbf{F}_4^d$  and  $\mathbf{F}_5^d$ ), the centroid of each class in a DFM is calculated as

$$c_l^d = \frac{1}{|x_{k,l}^d|} \sum_{x_{k,l}^d} x_{k,l}^d(i, j) \quad (12)$$

$$x_{k,l}^d(i, j) \in v_{k,n}(i, j), \quad v_{k,c}(i, j), \quad v_{k,u}(i, j)$$

where  $|x_{k,l}^d|$  is the number of chosen pixels of each class, and  $l = 1, 2, 3$  represents class  $\omega_n$ ,  $\omega_c$ , and  $\omega_u$ , respectively. Then, similarities between unlabeled nodes and class centroids can be calculated. After that, the unlabeled nodes can be assigned to the class with the highest probability value. The probability value  $V(x_k^d)$  is computed by

$$V(x_k^d) = \frac{\exp[\text{dis}(x_k^d, c_l^d)]}{\sum_{i=1}^l \exp[\text{dis}(x_k^d, c_i^d)]} \quad (13)$$

where  $\text{dis}(\cdot)$  indicates the distance metric. In this article, instead of using the Euclidean distance metric, we adopt a Mahalanobis distance metric (which can be learned during the model training) to measure the resemblance between the unlabeled nodes and centroids with the consideration of data

distribution, and its function can be present as

$$\text{dis}(\mathbf{x}_k^d, \mathbf{c}_l^d) = \sqrt{(\mathbf{x}_k^d - \mathbf{c}_l^d)^T \mathbf{M} (\mathbf{x}_k^d - \mathbf{c}_l^d)} \quad (14)$$

where  $\mathbf{M} = \mathbf{W}_m \mathbf{W}_m^T$  is a symmetric positive semidefinite matrix and  $\mathbf{W}_m$  is a trainable weight matrix. After that, we can divide pixels from  $\mathbf{F}_4^d$  and  $\mathbf{F}_5^d$  into three categories (i.e.,  $\omega_n$ ,  $\omega_c$ ,  $\omega_u$ ), which can be regarded as pseudolabels of this training epoch. Note that the pseudolabels and predicted results can be updated through a joint CD loss.

#### E. Joint CD Loss and Changed Pixel Classification

To train our model, we develop a joint CD loss with two terms: a metric-based cross-entropy loss and the Tversky loss [66]. As mentioned before, we can get the centroids of pseudolabels, and every pixel  $x_{k,i}^d$  from  $\mathbf{F}_4^d$  and  $\mathbf{F}_5^d$  is assigned to a label  $i$ . Therefore, the metric-based loss can be formulated as

$$L_{\text{Me}} = - \sum_{i=1,2} \log \frac{\exp[\text{dis}(x_{k,i}^d, c_i)]}{\sum_{i=1}^2 \exp[\text{dis}(x_{k,i}^d, c_i)]}. \quad (15)$$

The metric-based cross-entropy loss can compact the centroids and nodes belonging to the same pseudolabel so that our model could highlight characteristics of  $\omega_n$  and  $\omega_c$ . In addition, since  $\omega_u$  is not involved in training, the influence of the uncertain class can be decreased, and at the same time, the extent of uncertain areas can also be reduced during training. The Tversky loss is exploited for DIs to measure the difference between the predicted results and the pseudolabels. The Tversky loss is good at dealing with the issue of extreme class unbalance, and its definition is

$$L_{\text{Tv}} = \frac{|\mathcal{P}_k \cap \mathcal{L}_k|}{|\mathcal{P}_k \cap \mathcal{L}_k| + \alpha |\mathcal{P}_k - \mathcal{L}_k| + \beta |\mathcal{L}_k - \mathcal{P}_k|} \times \begin{cases} \mathcal{L}_k(i, j) = 0, & \mathcal{L}_k(i, j) \in \omega_n, \mathcal{L}_k(i, j) = 1 \\ \mathcal{L}_k(i, j) \in \omega_c \end{cases} \quad (16)$$

where  $\mathcal{P}_k$  and  $\mathcal{L}_k$  denotes the predicted results and the pseudolabels of the  $k$ th layer, respectively, and  $|\mathcal{L}_k \cap \mathcal{P}_k|$ ,  $|\mathcal{P}_k - \mathcal{L}_k|$ , and  $|\mathcal{L}_k - \mathcal{P}_k|$  represent the numbers of true positives (TPs), false positives (FPs), and false negatives (FNs), respectively. In addition, tradeoffs between  $|\mathcal{P}_k - \mathcal{L}_k|$  and  $|\mathcal{L}_k - \mathcal{P}_k|$  are controlled by  $\alpha$  and  $\beta$ .

In sum, the joint CD loss can be formulated as

$$L_{\text{CD}} = L_{\text{Tv}} + \lambda L_{\text{Me}} \quad (17)$$

where  $\lambda$  is a hyperparameter that controls contributions of two terms. After the optimization, the expected dual-channel predicted DIs could be generated. To fuse results of different scales and refine region boundaries, a multiscale decision fusion is utilized in this work. It uses a majority voting method to determine whether an area changes or not. Finally, we can obtain a binary change map  $\text{CM} \in \mathbb{R}^{H \times W}$ .

## IV. EXPERIMENT AND DISCUSSION

### A. Datasets and Experimental Settings

To verify the effectiveness of our model, we select four public datasets collected by different sensors, including QB,















Data Set	Pre-changed Images	Post-changed Images	Ground-truth
QB			
ZY3			
SZADA/2 SZTAKI			
Montpellier OSCD			

Fig. 6. Testing datasets.

ZY3 [19], Számítástechnikai és Automatizálási Kutatóintézet (Institute for Computer Science and Control) (SZTAKI) [21], and Onera Satellite Change Detection dataset (OSCD) [22]. The QB dataset contains a pair of HRRS images with RGB bands covering an area of Wuhan, China, and was collected in 2014 and 2016, respectively. The size and spatial resolution of the images are  $1154 \times 740$  and 2.4 m/pixel. The ZY3 dataset also covers Wuhan, China. The RGB HRRS images within ZY3 were obtained in 2009 and 2014, and their size is  $458 \times 559$  and the spatial resolution is 5.8 m/pixel. The SZTAKI dataset contains 12 pairs of HRRS images (acquisition times: 2000 and 2005) provided by the Hungarian Institute of Geodesy Cartography and RS. We select a pair of images (SZADA/2) with a size of  $952 \times 640$  and a spatial resolution of 1.5 m/pixel. The OSCD dataset contains 24 pairs of multispectral HRRS images with RGB and near-infrared bands captured by the Sentinel-2 satellites. In the following experiments, we choose a pair of images from Montpellier captured in 2015 and 2017. The size and spatial resolution of them are  $451 \times 426$  and 10 m/pixel. The images of the four datasets and their corresponding ground-truth maps are shown in Fig. 6.

GMCD is implemented by Pytorch [67], and all experiments are completed on a workstation with GeForce RTX 2080 Ti and 11G memory. The input HRRS images are resized into  $640 \times 640$ . Also, we select the Adam algorithm to train our model. The number of epochs and learning rate are set to be 40 and  $1 \times 10^{-3}$ , respectively. Note that when the number of epochs reaches 20, we reduce the learning rate to  $1 \times 10^{-4}$ . Besides, the values of  $T_1$  and  $T_2$  can be obtained by the Bayesian-based EM algorithm [64]. An example is shown in Fig. 7, which is counted by the QB dataset. For hyperparameters ( $\alpha$ ,  $\beta$ , and  $\lambda$ ) in the CD joint loss, we set  $\alpha = 0.3$ ,  $\beta = 0.7$  to highlight the importance of FN as described in [66], and set  $\lambda = 1$  empirically. The influence of  $\lambda$  will be discussed in Section IV-D. There is another point we want to touch on, that is, the memory costs. As mentioned

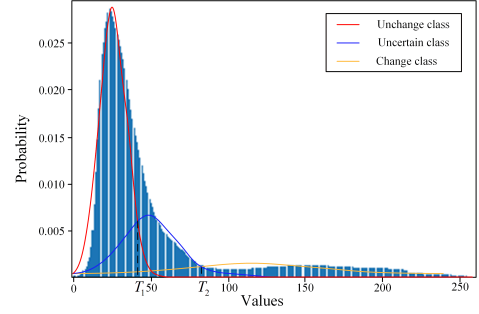


Fig. 7. Real histogram of different values and estimated probability density curves of different classes fit by the bayesian-based EM algorithm, which are counted by the QB dataset.

before, we resize the input data into  $640 \times 640$ . In this case, the memory requirement of the proposed scheme is 8271 MB.

To evaluate the proposed method quantitatively, five assessment criteria are used: precision (Pre), recall (Rec), F1 score, overall accuracy (OA), and Kappa coefficient. The higher the values of these metrics are, the better the CD results. Assume that we get TP, true negative (TN), FP, and FN. Then, the five metrics can be calculated as follows:

$$\begin{aligned}
 \text{Pre} &= \frac{TP}{TP + FP} \\
 \text{Rec} &= \frac{TP}{TP + FN} \\
 \text{F1} &= \frac{2 \times \text{Pre} \times \text{Rec}}{\text{Pre} + \text{Rec}} \\
 \text{OA} &= \frac{TP + TN}{TP + FP + TN + FN} \\
 \text{Kappa} &= \frac{\text{OA} - \text{PC}}{1 - \text{PC}} \\
 \text{where PC} &= \frac{(TP + FP) \times (TP + FN)}{(TP + FP + TN + FN)^2} \\
 &\quad + \frac{(TN + FN) \times (FP + TN)}{(TP + FP + TN + FN)^2}. \quad (18)
 \end{aligned}$$

### B. Performance of GMCD

To evaluate the performance of GMCD, seven popular unsupervised CD methods, including two traditional approaches and five deep-learning-based methods, are used as competitors. They are summarized as follows.

- 1) Iterative slow feature analysis (ISFA) [8], which is an unsupervised CD method based on SFA.
- 2) PCA-Kmeans [7], which extracts eigenvectors with PCA and accomplishes CD through k-means.
- 3) Symmetric CNN (SCCN) [25], a typical unsupervised CD method for heterogeneous radar images with a pretrained convolutional coupling network embedded by AE.
- 4) Principal component analysis network (PCANet) [68], which utilizes Gabor wavelets and fuzzy c-means to select desired patches and trains a PCANet with these patches for CD tasks.
- 5) Deep slow feature analysis (DSFA) [69], which is an unsupervised CD method that combines deep networks with SFA. Two deep symmetric networks are used to

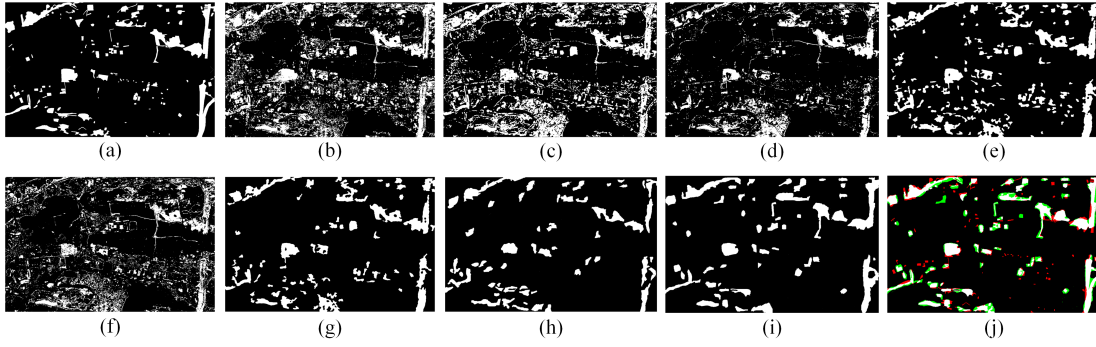


Fig. 8. Change maps obtained by different methods on the QB dataset. (a) ground truth. (b) ISFA. (c) PCA-Kmeans. (d) SCCN. (e) PCA-Net. (f) DSFA. (g) DCVA. (h) MSDRL. (i) GMCD. (j) Confusion map of GMCD (TP: white; TN: black; FP: green; FN: red).

TABLE I  
QUANTITATIVE COMPARISON OF CD RESULTS OBTAINED BY  
DIFFERENT METHODS ON THE QB DATASET (%)

Method	Pre	Rec	F1	OA	Kappa
ISFA	38.91	78.89	52.11	84.08	43.86
PCA-Kmeans	38.53	<b>79.42</b>	51.89	83.83	43.55
SCCN	47.59	59.84	53.02	88.36	46.47
PCANet	52.44	72.07	60.71	89.76	54.99
DSFA	41.43	60.48	49.17	86.27	41.55
DCVA	62.00	67.33	64.55	91.88	59.98
MSDRL	<b>73.03</b>	65.97	69.32	93.59	65.76
<b>GMCD</b>	70.20	76.08	<b>73.02</b>	<b>93.82</b>	<b>69.54</b>

learn useful features, and then the SFA module highlights changed components in the features.

- 6) DCVA [18], which is an unsupervised context-sensitivity CD method. It adopts a pretrained CNN to extract spatial contextual information and execute DCVA to identify changed pixels.
- 7) Multiscale difference representation learning (MSDRL) [63], which conducts an uncertainly analysis of spatial-spectral change information and trains a support vector machine (SVM) classifier with multiscale patches to obtained good CD results.

For the sake of fairness, all the input image pairs have been pre-processed by the same operations, including image registration, radiometric relative normalization, and pansharpening.

1) *Results on the QB Dataset:* In the QB dataset, farmland, woodland, and grassland are the main land covers, and the main change appears in the vegetation. The visual and numerical CD results of different methods on the QB dataset are shown in Fig. 8 and Table I.

From the observation of Fig. 8, we can find that compared with the ground-truth, CD results of traditional methods [cf. Fig. 8(b) and (c)] roughly cover changed areas, but include a large number of false alarms. For SCCN and PCA-Net, due to the introduction of deep features, the differences between changed and unchanged regions are widened, and the number of FPs is significantly decreased [see Fig. 8(d) and (e)]. The CD map of DSFA [cf. Fig. 8(f)] is not as good as expected. We can see some distinct false detections, which is because the simple classification strategy used in DSFA would lead to incorrect results. DCVA and MSDRL perform better than the

abovementioned methods [see Fig. 8(g) and (h)]. Moreover, the two methods can not only detect changes but also suppress noise in images. Nevertheless, boundaries of changed areas are not clear enough. Owe to GCN and metric learning, the context within RS images can be fully explored so that irregular changed areas can be detected precisely by GMCD [see Fig. 8(i)]. Furthermore, the confusion map [Fig. 8(j)] illustrates that our method is able to detect main changed areas more completely.

As displayed in Table I, the performance of GMCD is the best in terms of F1 (73.02%), OA (93.82%), and Kappa (69.54%). Compared with other competitors, the improvements in Kappa delivered by our model are 25.68% (over ISFA), 25.99% (over PCA-Kmeans), 23.07% (over SCCN), 14.55% (over PCANet), 27.99% (over DSFA), 9.56% (over DCVA), and 3.78% (over MSDRL). However, our precision and recall values are not the highest. The reason is that GMCD pays more attention to balance precision and recall. Although our method performs slightly weaker in these two metrics, overall, its performance is the best among all methods.

Fig. 8 and Table I confirm the effectiveness of our method on the QB dataset. However, an interesting observation is that GMCD fails in accurately detecting small areas. Also, the unsatisfactory precision reported in Table I indicates that the FP rate of our model is relatively large. These drawbacks are mainly caused by the resized operation. Since the original images in the QB dataset are large, the resize operation would lead to the loss of details. This point can be further confirmed by CD results on other smaller datasets (e.g., SZADA/2 dataset).

2) *Results on the ZY3 Dataset:* In the ZY3 dataset, the primary change type is construction. The shapes of changed regions are mostly polygonal, and changed areas are large. The visual and numerical results of different methods are shown in Fig. 9 and Table II.

As shown in Fig. 9(b) and (c), CD results of traditional methods (ISFA and PCA-Kmeans) suffer from salt-and-pepper noise, which indicates that they may be not suitable to process this dataset. With deep learning technologies, SCCN, PCA-Net, and DSFA can suppress false detections effectively, and changed regions can be depicted clearly [see Fig. 9(d)–(f)]. DCVA and MSDRL outperform other competing methods [Fig. 9(g) and (h)]. They can detect more

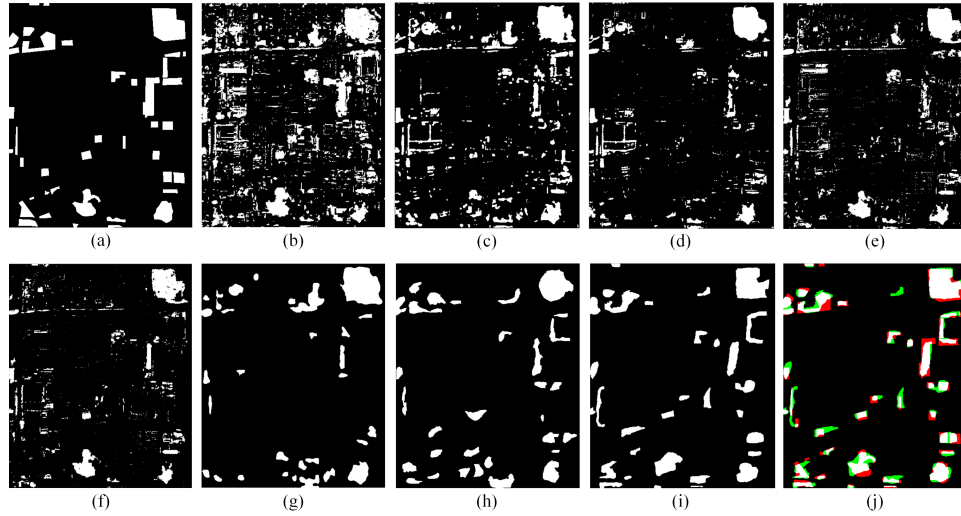


Fig. 9. Change maps obtained by different methods on the ZY3 dataset. (a) Ground truth. (b) ISFA. (c) PCA-Kmeans. (d) SCCN. (e) PCA-Net. (f) DSFA. (g) DCVA. (h) MSDRL. (i) GMCD. (j) Confusion map of GMCD (TP: white; TN: black; FP: green; FN: red).

TABLE II

QUANTITATIVE COMPARISON OF CD RESULTS OBTAINED BY DIFFERENT METHODS ON THE ZY3 DATASET (%)

Method	Pre	Rec	F1	OA	Kappa
ISFA	40.03	<b>71.79</b>	51.40	85.58	43.73
PCA-Kmeans	50.36	59.50	54.55	89.46	48.64
SCCN	60.23	45.75	52.00	91.02	47.15
PCANet	58.60	54.98	56.74	91.09	51.78
DSFA	73.18	48.99	58.69	92.67	54.85
DCVA	54.73	41.72	57.35	90.14	42.02
MSDRL	73.58	61.76	67.15	93.28	63.62
<b>GMCD</b>	<b>73.68</b>	71.57	<b>72.61</b>	<b>94.26</b>	<b>69.41</b>

TABLE III

QUANTITATIVE COMPARISON OF CD RESULTS OBTAINED BY DIFFERENT METHODS ON THE SZADA/2 DATASET (%)

Method	Pre	Rec	F1	OA	Kappa
ISFA	44.71	57.69	50.38	93.43	46.92
PCA-Kmeans	18.71	64.04	28.95	81.85	21.98
SCCN	37.94	60.33	46.58	92.00	42.50
PCANet	38.47	58.44	46.40	92.20	42.39
DSFA	47.48	56.45	51.58	93.87	48.34
DCVA	40.00	55.86	46.62	92.61	42.77
MSDRL	59.67	57.02	58.31	94.53	55.50
<b>GMCD</b>	<b>61.25</b>	<b>67.98</b>	<b>64.44</b>	<b>95.39</b>	<b>61.98</b>

changed areas. However, the shapes of these areas are not satisfactory compared with the ground-truth [cf. Fig. 9(a)]. In contrast with all competitors, both changed regions and their boundaries can be well identified by our approach, which can be verified in CD map [Fig. 9(i)] and confusion map [Fig. 9(j)]. These encouraging results show that our model works pretty well on the ZY3 dataset.

As can be seen in Table II, GMCD achieves the best performance in terms of most assessment criteria. For instance, compared with other methods, increases in Kappa gained by our model are 25.68% (over ISFA), 20.77% (over PCA-Kmeans), 22.26% (over SCCN), 17.63% (over PCANet), 14.56% (over DSFA), 27.39% (over DCVA), and 5.79% (over MSDRL). The enhancements in OA are 8.68% (over ISFA), 4.8% (over PCA-Kmeans), 3.24% (over SCCN), 3.17% (over PCANet), 1.59% (over DSFA), 4.12% (over DCVA), and 0.98% (over MSDRL). However, the recall of our model is a little lower than that of ISFA. This is the price of a great increase in precision that the value of GMCD is 23.65% higher than that of ISFA. The positive numerical results demonstrate the effectiveness of our method again.

3) *Results on the SZADA/2 Dataset:* The SZADA/2 dataset is gathered by different sensors and contains many kinds of changes, such as roadway construction, house building,

and fresh plowland, which make changed areas irregular [see Fig. 10(a)]. The visual and numerical CD results of different methods on this dataset are shown in Fig. 10 and Table III.

From the observation of visible results, we can see that CD results of the two traditional methods [cf. Fig. 10(b) and (c)] are extremely different, which indicates the instability of conventional methods. Fig. 10(d)–(g) show the results generated by SCCN, PCANet, DSFA, and DCVA. SCCN and DSFA tend to reveal all possible changes, so their CD results are refined. PCANet and DCVA miss many discrete changed areas because they focus more on highlighting large changes. As shown in Fig. 10(h), MSDRL further improves the CD result at the cost of losing small changes. Comparing with them, the proposed method is capable of simultaneously emphasizing changed regions and preserving details [see Fig. 10(i) and 10(j)]. This is because of the use of Mlt-GCN and metric learning in our model. These promising results illustrate that our method is superior to the SZADA/2 dataset.

As reported in Table III, we can see that our model performs the best. Taking F1 score as an example, compared with other methods, the improvements in it obtained by our model are 14.06% (over ISFA), 35.49% (over PCA-Kmeans),



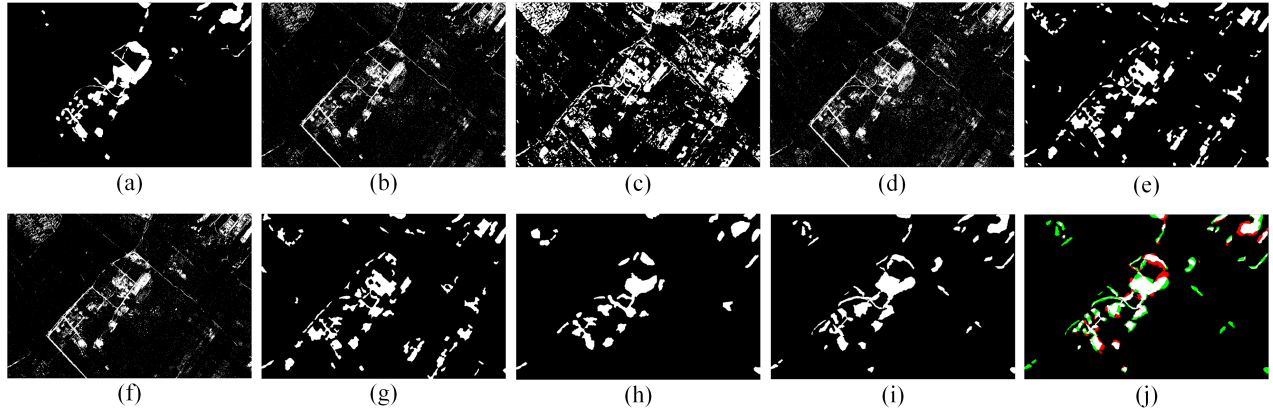


Fig. 10. Change maps obtained by different methods on the SZADA/2 dataset. (a) Ground truth. (b) ISFA. (c) PCA-Kmeans. (d) SCCN. (e) PCA-Net. (f) DSFA. (g) DCVA. (h) MSDRL. (i) GMCD. (j) Confusion map of GMCD (TP: white; TN: black; FP: green; FN: red).

TABLE IV  
QUANTITATIVE COMPARISON OF CD RESULTS OBTAINED BY  
DIFFERENT METHODS ON THE MONTPELLIER DATASET (%)

Method	Pre	Rec	F1	OA	Kappa
ISFA	34.66	60.19	43.99	89.58	38.70
PCA-Kmeans	64.28	68.24	66.2	95.26	63.66
SCCN	69.73	58.00	63.33	95.43	60.91
PCANet	57.54	<b>79.14</b>	66.63	94.61	63.79
DSFA	52.02	43.32	47.28	93.43	43.81
DCVA	62.20	73.19	67.25	95.15	64.65
MSDRL	<b>83.40</b>	66.73	74.13	96.83	72.47
<b>GMCD</b>	82.28	73.64	<b>77.72</b>	<b>97.13</b>	<b>76.19</b>

17.86% (over SCCN), 18.04% (over PCANet), 12.86% (over DSFA), 17.82% (over DCVA), and 6.13% (over MSDRL).

These positive results prove the usefulness of our model in CD tasks. Also, through the above three experiments, the effectiveness of GMCD using 3-channel images for CD has been fully verified.

4) *Results on the Montpellier Dataset*: The Montpellier dataset is a pair of 4-channel images, and the main content is urban, which indicates that the scene is complicated and the changed areas are tanglesome [see Fig. 11(a)]. The visual and numerical CD results of different methods on this dataset are shown in Fig. 11 and Table IV.

Different from the SZADA/2 dataset, PCA-Kmeans outperforms ISFA on this dataset, which further exhibits the instability of the traditional methods. In addition, some deep-learning-based methods, such as SCCN and DSFA, do not perform well on this scene as shown in Fig. 11(d) and (f). Besides, due to the large diversity of CD regions in the dataset, PCA-Net, DCVA, and MSDRL miss a lot of detailed regions [cf. Fig. 11(e), (g), and (h)]. On account of introducing the spectral information into the generation of pseudolabels, our method is able to detect more tiny changed areas, as shown in Fig. 11(i) and (j). As shown in Table IV, the precision and recall of our method are 82.28% and 73.64%, respectively. Although the precision is slightly lower than that of MSDRL and the recall is slightly lower than that of PCANet,

the proposed method is the best in terms of primary metrics, i.e., F1 score, OA, and Kappa.

### C. Ablation Studies

To fully evaluate contributions of the different components of GMCD, we conduct the following ablation studies. We treat the proposed Siamese FCN as the basic CD model composed of a pretrained Siamese encoder and a pyramid-shaped decoder. The pseudolabels are generated by the spatial-spectral analysis, where the learned features and input images are regarded as spatial and spectral features, respectively. The specific experimental settings can be found in [63]. Then, we introduce three different modules, namely Mlt-GCN, metric learning, and fusion module, into the basic model step by step. For the sake of simplicity, we name “Baseline + Mlt-GCN” Model-1, record “Baseline + Mlt-GCN + metric learning” Model-2, represent “Baseline + metric learning + fusion” Model-3, and call “Baseline + Mlt-GCN + metric learning + fusion” Model-4. In fact, Model-4 equals GMCD. Three overall assessment criteria (F1, OA, and Kappa) are used to evaluate CD results, and they are reported in Table V.

By using the Mlt-GCN module, Model-1 is capable of better-capturing irregular and long-range contextual information. Besides, it can detect small CD areas accurately in the image pairs and avoid missed detections effectively. Accordingly, F1, OA, and Kappa values of Model-1 on the four datasets are increased distinctly. For example, on the QB dataset, the enhancement in F1 is 3.43%, the increase in OA is 0.89%, and the improvement in Kappa is 3.95%. These results demonstrate the effectiveness of the Mlt-GCN module.

When adding the metric learning module, Model-2 can dynamically and adaptively generate pseudolabels and provides a helpful loss function for optimizing the network, which can enlarge the difference between the change and unchanged classes and highlight change areas. The significant gains in terms of F1 score, OA, and Kappa coefficient on the four datasets prove the advantages of this module.

Model-3 contains all components within GMCD except Mlt-GCN. Thus, it can only capture the information and short-range relations from the fixed rectangular areas. Also,

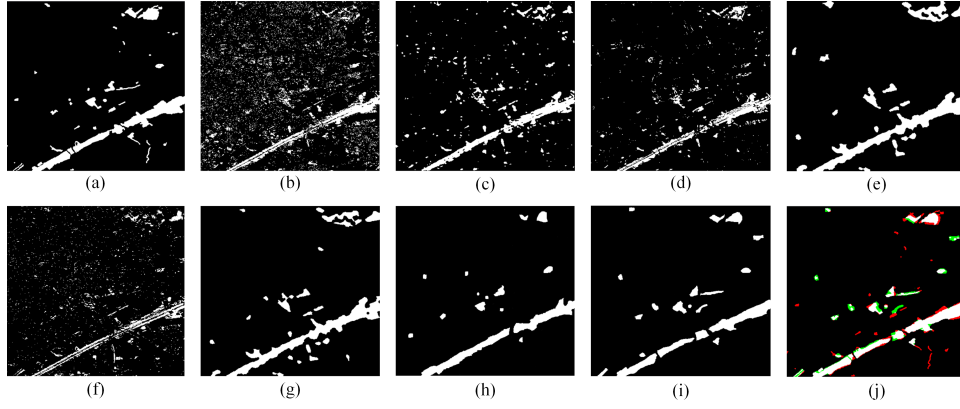


Fig. 11. Change maps obtained by different methods on the Montpellier\*\*\* dataset. (a) Ground truth. (b) ISFA. (c) PCA-Kmeans. (d) SCCN. (e) PCA-Net. (f) DSFA. (g) DCVA. (h) MSDRL. (i) GMCD. (j) Confusion map of GMCD (TP: white; TN: black; FP: green; FN: red).

TABLE V  
ABLATION STUDIES OF GMCD WITH DIFFERENT MODULES ON FOUR DATASETS (%)

Method	F1				OA				Kappa			
	QB	ZY3	SZADA/2	Montpellier	QB	ZY3	SZADA/2	Montpellier	QB	ZY3	SZADA/2	Montpellier
Baseline	65.08	64.89	55.80	69.31	91.98	92.31	93.82	95.60	60.55	60.58	52.48	66.95
Model-1	68.51	68.61	59.37	72.41	92.87	93.38	94.71	96.37	64.50	64.91	56.55	70.46
Model-2	72.38	72.12	63.80	76.79	93.55	94.21	95.23	97.04	68.74	68.89	61.25	75.21
Model-3	68.46	68.02	59.02	71.70	92.74	93.25	94.61	96.28	64.36	64.25	56.12	69.71
Model-4	<b>73.02</b>	<b>72.61</b>	<b>64.44</b>	<b>77.72</b>	<b>93.82</b>	<b>94.26</b>	<b>95.39</b>	<b>97.13</b>	<b>69.54</b>	<b>69.41</b>	<b>61.98</b>	<b>76.19</b>

due to the absence of Mlt-GCN, the input data of the pseudolabel generation block does not contain the necessary spectral knowledge. Accordingly, its behavior is the weakest among different models. This demonstrates the importance of Mlt-GCN and the effectiveness of long-range contextual patterns.

Through the fusion operation, Model-4 can integrate DIs with different scales together and enhance the model's ability to recognize more details. Therefore, the F1 score, OA, and Kappa coefficient are further increased on four datasets. For example, on the QB dataset, the improvement in F1 is 0.64%, the enhancement in OA is 0.27%, and the increase in Kappa is 0.8%. The effectiveness of the fusion operation has been adequately verified through these experiments.

#### D. Influence of $\lambda$

For the proposed joint CD loss [see (17)], the hyperparameter  $\lambda$  controls contributions of  $L_{Me}$  and  $L_{Tv}$ .  $L_{Me}$  focuses on compacting samples and their corresponding class centers and highlighting change and unchanged labels to generate useful pseudolabels.  $L_{Tv}$  aims to mitigate the negative impact of the issue of unbalanced classes. To dive into the influence of  $\lambda$ , we set it as 0.5, 0.8, 1, 1.2, and 1.5, respectively, and observe variations in the performance of GMCD on four different datasets (see Fig. 12). As can be seen in this figure, GMCD achieves the best results when  $\lambda = 1$  for all datasets. Upon  $\lambda$  is less than 1, its performance reduces dramatically, which means that the effect of  $L_{Me}$  is weakened, and the accuracy of pseudolabels decrease to a certain extent. When  $\lambda$  is larger than 1, the classification loss function  $L_{Tv}$  has a decreased

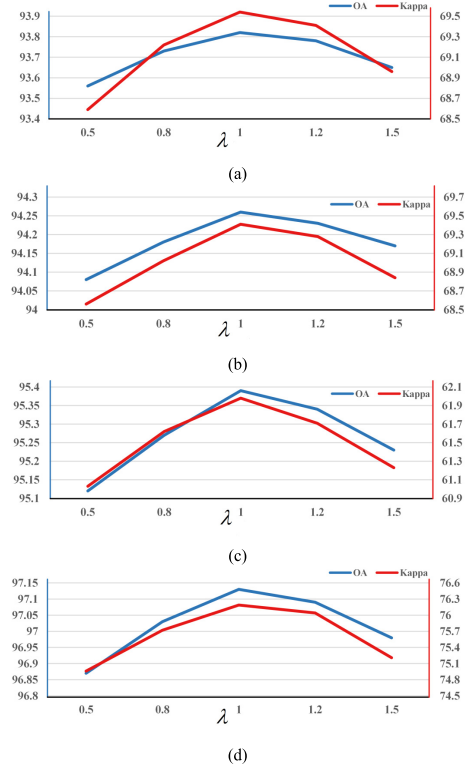


Fig. 12. Relationship between  $\lambda$  and the CD results (OA and kappa) on four datasets. (a) QB. (b) ZY3. (c) SZADA/2. (d) Montpellier.

impact, resulting in the drop of the CD results. Therefore, we set  $\lambda$  as 1 to ensure that functions of the two components ( $L_{Me}$  and  $L_{Tv}$ ) can be fully utilized simultaneously.

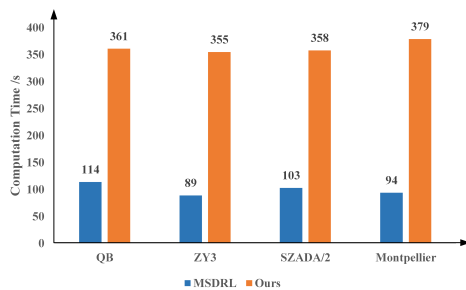


Fig. 13. Computational costs of MSDRL and GMCD.

### E. Computational Costs

Fig. 13 shows the computational cost of GMCD, and that of MSDRL is also exhibited for reference because MSDRL is the second best model in our experiments, and its training process is similar to ours. Note that times displayed in Fig. 13 are computational costs of whole processes without pretraining. It is evident that GMCD is slower than MSDRL, which is mainly because of the use of computing resources by Mlt-GCN and the time consumption during the generation of pseudolabels. Besides, our method consumes a little more time on the Montpellier dataset than on others, as the images of the Montpellier scene include one more channel, which increases the computational cost. However, on the whole, the maximum computational time, 379 s, is also acceptable.

## V. CONCLUSION

This article proposes an unsupervised CD method, GMCD, using HRRS images based on GCN and metric learning. GMCD consists of a pretrained Siamese FCN encoder and a pyramid-shaped decoder. The encoder is responsible for extracting discriminative features from the bi-temporal input images, and the decoder aims to integrate multiscale feature maps and generate dual-channel DIs. To capture short- and long-range contextual patterns within HRRS images, we devise Mlt-GCN and embed it into the encoder. We propose a dynamic pseudolabel generation mechanism, in order to train our method in an unsupervised way. By analyzing spatial-spectral features and using metric learning, not only can change and unchanged areas be highlighted, the quality of pseudolabels is also improved. Also, a joint CD loss is used to balance the pseudolabel generation and change area detection. Finally, a multiscale decision fusion is used to integrate DIs. The effectiveness of our GMCD is demonstrated by extensive experiments on four datasets.

The visual and numerical results on four different datasets show that our method outperforms the other competitors, including traditional and deep-learning-based methods. However, due to the use of GCN and the generation of dynamic pseudolabels in our method, the detection speed of GMCD is relatively slow. Therefore, our further work will aim to reduce the computational complexity of our model to speed up detection.

## REFERENCES

- [1] R. S. Lunetta, J. F. Knight, J. Ediriwickrema, J. G. Lyon, and L. D. Worthy, "Land-cover change detection using multi-temporal MODIS NDVI data," *Remote Sens. Environ.*, vol. 105, no. 2, pp. 142–154, 2006.
- [2] 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.
- [3] Y. Ban and O. A. Yousif, "Multitemporal spaceborne SAR data for urban change detection in China," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 5, no. 4, pp. 1087–1094, Aug. 2012.
- [4] Z. Zhu, "Change detection using Landsat time series: A review of frequencies, preprocessing, algorithms, and applications," *ISPRS J. Photogramm. Remote Sens.*, vol. 130, pp. 370–384, Aug. 2017.
- [5] A. Asokan and J. Anitha, "Change detection techniques for remote sensing applications: A survey," *Earth Sci. Informat.*, vol. 12, no. 2, pp. 143–160, Jun. 2019.
- [6] A. J. H. Meddens, J. A. Hicke, L. A. Vierling, and A. T. Hudak, "Evaluating methods to detect bark beetle-caused tree mortality using single-date and multi-date landsat imagery," *Remote Sens. Environ.*, vol. 132, pp. 49–58, May 2013.
- [7] T. Celik, "Unsupervised change detection in satellite images using principal component analysis and *k*-means clustering," *IEEE Geosci. Remote Sens. Lett.*, vol. 6, no. 4, pp. 772–776, Apr. 2009.
- [8] 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.
- [9] F. Bovolo and L. Bruzzone, "An adaptive thresholding approach to multiple-change detection in multispectral images," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Jul. 2011, pp. 233–236.
- [10] Y. T. Solano-Correa, F. Bovolo, and L. Bruzzone, "An approach to multiple change detection in VHR optical images based on iterative clustering and adaptive thresholding," *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 8, pp. 1334–1338, Aug. 2019.
- [11] Y. Chen and Z. Cao, "An improved MRF-based change detection approach for multitemporal remote sensing imagery," *Signal Process.*, vol. 93, no. 1, pp. 163–175, 2013.
- [12] M. Hussain, D. Chen, A. Cheng, H. Wei, and D. Stanley, "Change detection from remotely sensed images: From pixel-based to object-based approaches," *ISPRS J. Photogramm. Remote Sens.*, vol. 80, pp. 91–106, Jun. 2013.
- [13] L. Li, X. Li, Y. Zhang, L. Wang, and G. Ying, "Change detection for high-resolution remote sensing imagery using object-oriented change vector analysis method," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2016, pp. 2873–2876.
- [14] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 25, 2012, pp. 1097–1105.
- [15] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.
- [16] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Cham, Switzerland: Springer, 2015, pp. 234–241.
- [17] A. Varghese, J. Gubbi, A. Ramaswamy, and P. Balamuralidhar, "ChangeNet: A deep learning architecture for visual change detection," in *Proc. Eur. Conf. Comput. Vis. (ECCV) Workshops*, 2018, pp. 129–145.
- [18] S. Saha, F. Bovolo, and L. Bruzzone, "Unsupervised deep change vector analysis for multiple-change detection in VHR images," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 6, pp. 3677–3693, Jun. 2019.
- [19] M. Zhang and W. Shi, "A feature difference convolutional neural network-based change detection method," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 10, pp. 7232–7246, Oct. 2020.
- [20] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 3431–3440.
- [21] C. Benedek and T. Szirányi, "Change detection in optical aerial images by a multilayer conditional mixed Markov model," *IEEE Trans. Geosci. Remote Sens.*, vol. 47, no. 10, pp. 3416–3430, Oct. 2009.
- [22] R. C. Daudt, B. L. Saux, A. Boulch, and Y. Gousseau, "Urban change detection for multispectral Earth observation using convolutional neural networks," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2018, pp. 2115–2118.
- [23] 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, 2017.
- [24] 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, Oct. 2017.



- [25] 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, Mar. 2018.
- [26] Y. Li, L. Zhou, G. Lu, B. Hou, and L. Jiao, "Change detection in synthetic aperture radar images based on log-mean operator and stacked auto-encoder," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2017, pp. 3090–3096.
- [27] 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. 2019.
- [28] K. L. de Jong and A. S. Bosman, "Unsupervised change detection in satellite images using convolutional neural networks," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2019, pp. 1–8.
- [29] X. Li, Z. Yuan, and Q. Wang, "Unsupervised deep noise modeling for hyperspectral image change detection," *Remote Sens.*, vol. 11, no. 3, p. 258, 2019.
- [30] H. Chen, C. Wu, B. Du, and L. Zhang, "Change detection in multi-temporal VHR images based on deep Siamese multi-scale convolutional networks," 2019, *arXiv:1906.11479*. [Online]. Available: <http://arxiv.org/abs/1906.11479>
- [31] F. Liu, L. Jiao, X. Tang, S. Yang, W. Ma, and B. Hou, "Local restricted convolutional neural network for change detection in polarimetric SAR images," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 3, pp. 818–833, Mar. 2019.
- [32] F. Liu, X. Tang, X. Zhang, L. Jiao, and J. Liu, "Large-scope PolSAR image change detection based on looking-around-and-into mode," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 1, pp. 363–378, Jan. 2021.
- [33] Z. Yuan, Q. Wang, and X. Li, "ROBUST PCANet for hyperspectral image change detection," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2018, pp. 4931–4934.
- [34] 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.
- [35] F. Wang and D. M. J. Tax, "Survey on the attention based RNN model and its applications in computer vision," 2016, *arXiv:1601.06823*. [Online]. Available: <http://arxiv.org/abs/1601.06823>
- [36] H. Jiang, X. Hu, K. Li, J. Zhang, J. Gong, and M. Zhang, "PGA-SiamNet: Pyramid feature-based attention-guided Siamese network for remote sensing orthoimagery building change detection," *Remote Sens.*, vol. 12, no. 3, p. 484, Feb. 2020.
- [37] O. Sener, H. O. Song, A. Saxena, and S. Savarese, "Learning transferable representations for unsupervised domain adaptation," in *Proc. 30th Int. Conf. Neural Inf. Process. Syst.*, 2016, pp. 2118–2126.
- [38] M. Yang, L. Jiao, F. Liu, B. Hou, and S. Yang, "Transferred deep learning-based change detection in remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 9, pp. 6960–6973, Sep. 2019.
- [39] S. Saha, F. Bovolo, and L. Bruzzone, "Building change detection in VHR SAR images via unsupervised deep transcoding," *IEEE Trans. Geoecon. Remote Sens.*, vol. 59, no. 3, pp. 1917–1929, Mar. 2021.
- [40] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, and P. S. Yu, "A comprehensive survey on graph neural networks," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 32, no. 1, pp. 4–24, Jan. 2021.
- [41] J. Zhou et al., "Graph neural networks: A review of methods and applications," 2018, *arXiv:1812.08434*. [Online]. Available: <http://arxiv.org/abs/1812.08434>
- [42] F. Monti, D. Boscaini, J. Masci, E. Rodola, J. Svoboda, and M. M. Bronstein, "Geometric deep learning on graphs and manifolds using mixture model CNNs," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 5115–5124.
- [43] P. Lin, P. Sun, G. Cheng, S. Xie, X. Li, and J. Shi, "Graph-guided architecture search for real-time semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2020, pp. 4203–4212.
- [44] Y. Yang, C. Feng, Y. Shen, and D. Tian, "FoldingNet: Point cloud auto-encoder via deep grid deformation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 206–215.
- [45] X. Wang, Y. Ye, and A. Gupta, "Zero-shot recognition via semantic embeddings and knowledge graphs," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6857–6866.
- [46] J. Bruna, W. Zaremba, A. Szlam, and Y. LeCun, "Spectral networks and locally connected networks on graphs," 2013, *arXiv:1312.6203*. [Online]. Available: <http://arxiv.org/abs/1312.6203>
- [47] X. Li, Y. Yang, Q. Zhao, T. Shen, Z. Lin, and H. Liu, "Spatial pyramid based graph reasoning for semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 8950–8959.
- [48] N. Khan, U. Chaudhuri, B. Banerjee, and S. Chaudhuri, "Graph convolutional network for multi-label VHR remote sensing scene recognition," *Neurocomputing*, vol. 357, pp. 36–46, May 2019.
- [49] J. Liang, Y. Deng, and D. Zeng, "A deep neural network combined CNN and GCN for remote sensing scene classification," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 13, pp. 4325–4338, Jul. 2020.
- [50] U. Chaudhuri, B. Banerjee, and A. Bhattacharya, "Siamese graph convolutional network for content based remote sensing image retrieval," *Comput. Vis. Image Understand.*, vol. 184, pp. 22–30, Jul. 2019.
- [51] H. You, S. Tian, L. Yu, and Y. Lv, "Pixel-level remote sensing image recognition based on bidirectional word vectors," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 2, pp. 1281–1293, Feb. 2020.
- [52] D. Hong, L. Gao, J. Yao, B. Zhang, A. Plaza, and J. Chanussot, "Graph convolutional networks for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 7, pp. 5966–5978, Jul. 2021.
- [53] S. Wan, C. Gong, P. Zhong, S. Pan, G. Li, and J. Yang, "Hyperspectral image classification with context-aware dynamic graph convolutional network," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 1, pp. 597–612, May 2020.
- [54] S. Saha, L. Mou, X. X. Zhu, F. Bovolo, and L. Bruzzone, "Semisupervised change detection using graph convolutional network," *IEEE Geosci. Remote Sens. Lett.*, vol. 18, no. 4, pp. 607–611, Apr. 2021.
- [55] B. Zitová and J. Flusser, "Image registration methods: A survey," *Image Vis. Comput.*, vol. 21, pp. 977–1000, Oct. 2003.
- [56] F. Pacifici, N. Longbotham, and W. J. Emery, "The importance of physical quantities for the analysis of multitemporal and multiangular optical very high spatial resolution images," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 10, pp. 6241–6256, Oct. 2014.
- [57] F. Bovolo, L. Bruzzone, L. Capobianco, A. Garzelli, S. Marchesi, and F. Nencini, "Analysis of the effects of pansharpening in change detection on VHR images," *IEEE Geosci. Remote Sens. Lett.*, vol. 7, no. 1, pp. 53–57, Jan. 2010.
- [58] J. Ma, L. Wu, X. Tang, F. Liu, X. Zhang, and L. Jiao, "Building extraction of aerial images by a global and multi-scale encoder-decoder network," *Remote Sens.*, vol. 12, no. 15, p. 2350, Jul. 2020.
- [59] G. Cheng, J. Han, P. Zhou, and L. Guo, "Multi-class geospatial object detection and geographic image classification based on collection of part detectors," *ISPRS J. Photogramm. Remote Sens.*, vol. 98, pp. 119–132, Dec. 2014.
- [60] 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, Nov. 2012.
- [61] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7794–7803.
- [62] Y. Zou, Z. Yu, B. Kumar, and J. Wang, "Unsupervised domain adaptation for semantic segmentation via class-balanced self-training," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 289–305.
- [63] T. Zhan, M. Gong, X. Jiang, and M. Zhang, "Unsupervised scale-driven change detection with deep Spatial-Spectral features for VHR images," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 8, pp. 5653–5665, Aug. 2020.
- [64] M. Zanetti and L. Bruzzone, "A theoretical framework for change detection based on a compound multiclass statistical model of the difference image," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 2, pp. 1129–1143, Feb. 2018.
- [65] W. Lin, Z. Gao, and B. Li, "Shoestring: Graph-based semi-supervised classification with severely limited labeled data," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 4174–4182.
- [66] N. Abraham and N. M. Khan, "A novel focal Tversky loss function with improved attention U-Net for lesion segmentation," in *Proc. IEEE 16th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2019, pp. 683–687.
- [67] A. Paszke et al., "PyTorch: An imperative style, high-performance deep learning library," 2019, *arXiv:1912.01703*. [Online]. Available: <http://arxiv.org/abs/1912.01703>
- [68] 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.
- [69] 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 Sens.*, vol. 57, no. 12, pp. 9976–9992, Dec. 2019.



**Xu Tang** (Member, IEEE) received the B.Sc., M.Sc., and Ph.D. degrees in electronic circuit and system from Xidian University, Xi'an, China, in 2007, 2010, and 2017, respectively.

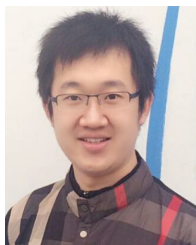
From 2015 to 2016, he was a Joint Ph.D. along with Prof. W. J. Emery at the University of Colorado at Boulder, Boulder, CO, USA. He is an Associate Professor with the Key Laboratory of Intelligent Perception and Image Understanding, Ministry of Education, Xidian University. His research interests include remote sensing (RS)

image content-based retrieval and reranking, hyperspectral image processing, RS scene classification, object detection, etc. For more details, please refer to: <https://web.xidian.edu.cn/tangxu/>.



**Huayu Zhang** (Graduate Student Member, IEEE) received the B.E. degree in automation from Xidian University, Xi'an, China, in 2019, where he is pursuing the M.S. degree in computer science and technology with the Key Laboratory of Intelligent Perception and Image Understanding of the Ministry of Education.

His research interests include image understanding, remote sensing image processing, and deep learning.



**Lichao Mou** received the bachelor's degree in automation from the Xi'an University of Posts and Telecommunications, Xi'an, China, in 2012, the master's degree in signal and information processing from the University of Chinese Academy of Sciences (UCAS), Beijing, China, in 2015, and the Dr.-Ing. degree from the Technical University of Munich (TUM), Munich, Germany, in 2020.

He is a Guest Professor at the Munich Artificial Intelligence (AI) Future Lab AI4EO, TUM and the Head of Visual Learning and Reasoning team at

the Department "EO Data Science," Remote Sensing Technology Institute (IMF), German Aerospace Center (DLR), Wessling, Germany. Since 2019, he has been the Research Scientist at DLR-IMF and an AI Consultant for the Helmholtz Artificial Intelligence Cooperation Unit (HAICU). In 2015 he spent six months at the Computer Vision Group at the University of Freiburg in Germany. In 2019 he was a Visiting Researcher with the Cambridge Image Analysis Group (CIA), University of Cambridge, U.K.

Dr. Mou was a recipient of the first place in the 2016 IEEE Geoscience and Remote Sensing Society (GRSS) Data Fusion Contest and the Finalists for the Best Student Paper Award at the 2017 Joint Urban Remote Sensing Event and 2019 Joint Urban Remote Sensing Event.



**Fang Liu** (Member, IEEE) was born in China, in 1990. She received the B.S. degree in information and computing science from Henan University, Kaifeng, China, in 2012, and the Ph.D. degree in intelligent information processing from Xidian University, Xi'an, China, in 2018.

She is a Lecturer with the Nanjing University of Science and Technology, Nanjing, China. Her research interests include deep learning, object detection, polarimetric SAR image classification and change detection.



**Xiangrong Zhang** (Senior Member, IEEE) received the B.S. and M.S. degrees from the School of Computer Science, Xidian University, Xi'an, China, in 1999 and 2003, respectively, and the Ph.D. degree from the School of Electronic Engineering, Xidian University, in 2006.

She is a Professor with the Key Laboratory of Intelligent Perception and Image Understanding of the Ministry of Education, Xidian University. From January 2015 to March 2016, she was a Visiting Scientist with the Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA, USA. Her research interests include pattern recognition, machine learning, and remote sensing image analysis and understanding.



**Xiao Xiang Zhu** (Fellow, IEEE) received the M.Sc., Dr.-Ing., and Habilitation degrees in signal processing from the Technical University of Munich (TUM), Munich, Germany, in 2008, 2011, and 2013, respectively.

She is a Professor for Data Science in Earth Observation (former: Signal Processing in Earth Observation), Technical University of Munich (TUM), and the Head of the Department "EO Data Science," Remote Sensing Technology Institute, German Aerospace Center (DLR). Since 2019,

she has been a Co-Coordinator of the Munich Data Science Research School ([www.mu-ds.de](http://www.mu-ds.de)). Since 2019, she also heads the Helmholtz Artificial Intelligence–Research Field "Aeronautics, Space and Transport." Since May 2020, she has been the Director of the International Future Artificial Intelligence (AI) Lab "AI4EO–Artificial Intelligence for Earth Observation: Reasoning, Uncertainties, Ethics and Beyond," Munich, Germany. Since October 2020, she has also been serves as a Co-Director of the Munich Data Science Institute (MDSI), TUM. She was a Guest Scientist or a Visiting Professor at the Italian National Research Council (CNR-IREA), Naples, Italy, Fudan University, Shanghai, China, The University of Tokyo, Tokyo, Japan and University of California, Los Angeles, CA, USA, in 2009, 2014, 2015, and 2016, respectively. She is a Visiting AI Professor at ESA's Phi-Laboratory. Her main research interests are remote sensing and Earth observation, signal processing, machine learning, and data science, with a special application focus on global urban mapping.

Dr. Zhu is a member of Young Academy (Junge Akademie/Junges Kolleg) at the Berlin-Brandenburg Academy of Sciences and Humanities and the German National Academy of Sciences Leopoldina and the Bavarian Academy of Sciences and Humanities. She serves on the Scientific Advisory Board in several research organizations, among others the German Research Center for Geosciences (GFZ) and Potsdam Institute for Climate Impact Research (PIK). She is an Associate Editor of IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING and serves as the Area Editor responsible for special issues of *IEEE Signal Processing Magazine*.



**Licheng Jiao** (Fellow, IEEE) received the B.S. degree in high voltage from Shanghai Jiao Tong University, Shanghai, China, in 1982, and the M.S. and Ph.D. degrees in electronic engineering from Xi'an Jiaotong University, Xi'an, China, in 1984 and 1990, respectively.

From 1984 to 1986, he was an Assistant Professor with the Civil Aviation Institute of China, Tianjin, China. From 1990 to 1991, he was a Post-Doctoral Fellow with the Key Laboratory for Radar Signal Processing, Xidian University, Xi'an, where he is

the Director of the Key Laboratory of Intelligent Perception and Image Understanding of Ministry of Education of China. He has authored or coauthored more than 200 scientific articles. His research interests include signal and image processing, nonlinear circuits and systems theory, wavelet theory, natural computation, and intelligent information processing.

Dr. Jiao is also a member of the IEEE Xian Section Executive Committee and an Executive Committee Member of the Chinese Association of Artificial Intelligence. He is the Chairman of the Awards and Recognition Committee.