Multiview Hypergraph Fusion Network for Change Detection in High-Resolution Remote Sensing Images

Xue Zhao, Kai Zhang, *Member, IEEE*, Feng Zhang, Jiande Sun, Wenbo Wan, and Huaxiang Zhang

***Abstract*—Currently, convolutional neural networks and transformers have been the dominant paradigms for change detection (CD) thanks to their powerful local and global feature extraction capabilities. However, with the improvement of resolution, spatial, spectral, and temporal relationships among objects in remote sensing images are becoming more complicated and cannot be modeled efficiently by the existing methods. To capture the high-order complex relationships in images, we propose a multiview hypergraph fusion network (MVHFNet) for CD, in which the high-order relationships along spatial, spectral, and temporal views are extracted by hypergraph learning. Specifically, this network is composed of three branches, including the spectral hypergraph learning branch, the spatial hypergraph learning branch, and the temporal hypergraph learning branch. In these branches, multiview features are extracted by different attention modules, and hypergraph learning consisting of hypergraph construction and hypergraph convolution is imposed on these features to model the high-order relationships. Then, to integrate the multiview features from different branches, a multiview feature fusion module is designed, in which the multiview features are fused and condensed for the following prediction. Finally, the change map is produced by a prediction head. We conduct extensive experiments on three datasets, such as LEVIR-CD, SYSU-CD, and CLCD. The experimental results demonstrate that the proposed MVHFNet achieves better CD performance compared to some state-of-the-art methods.**

***Index Terms*—Change detection (CD), hypergraph network, multiview fusion, remote sensing images.**

I. Introduction

MULTITEMPORAL high-resolution (HR) remote sensing images contain rich spatial, spectral, and temporal information. Effectively leveraging these abundant features is essential for monitoring changes on the earth's surface [1]. Change detection (CD) aims to infer the changed regions among multitemporal HR remote sensing images, which has been applied to many fields, including land cover analysis [2], urban expansion [3], and earthquake damage estimation [4], [5].

In recent years, deep neural networks (DNNs), particularly convolutional neural networks (CNNs), have demonstrated exceptional performance in the field of CD due to their strong ability to represent features effectively. For example, in [6], a CNN was proposed for the semantic CD of multitemporal HR images, in which a shared U-Net was adopted to encode the spatial and spectral features of changed areas. In [7], a W-Net was developed, which integrated the superpixel technique into the U-Net structure. In [8], a high-frequency attention-guided concatenated network was presented to better highlight the high-frequency information of buildings. In [9], an encoder–decoder architecture based on a CNN model was introduced. This model incorporated gating and self-attention modules to achieve efficient fusion of multimodal features. The features extracted by CNNs demonstrate strong performance in terms of capturing local details and characteristics. However, these networks may struggle to effectively model and represent the global information in images.

To address the need for capturing global relationships in images, researchers have increasingly turned their attention to transformer-based and graph convolutional neural network (GCN) methods for CD. For instance, a cross-temporal difference transformer was established in [10] to learn the local–global semantic features in images. In [11], a hybrid transformer was built for CD, in which the architecture was designed according to that of U-Net. In [12], a transformer-based contextual information aggregation network was proposed, effectively filtering out irrelevant changes while extracting rich semantic information. The CNN and the transformer were also jointly considered in [13] to infer the spatial and spectral dependencies. Besides, as a typical global learning tool, GCN-based models are also considered. Tang et al. [14] introduced a multiscale GCN that incorporated metric learning techniques to enhance CD performance. In [15], a multiscale framework was employed to model high-level semantic similarities at various levels using a GCN. In [16], autoencoders with bipartite graph attention were proposed to learn the global properties of different objects in

Manuscript received 21 November 2023; revised 18 December 2023; accepted 16 January 2024. Date of publication 31 January 2024; date of current version 15 February 2024. This work was supported in part by the Natural Science Foundation of Shandong Province of China under Grant ZR2023MF066, in part by the National Natural Science Foundation of China under Grant 61901246, in part by the China Postdoctoral Science Foundation under Grant 2019TQ0190 and Grant 2019M662432, in part by the Scientific Research Leader Studio of Ji'nan under Grant 2021GXRC081, and in part by the Joint Project for Smart Computing of Shandong Natural Science Foundation under Grant ZR2020LZH015. *(Corresponding author: Kai Zhang.)*

Xue Zhao, Kai Zhang, Jiande Sun, Wenbo Wan, and Huaxiang Zhang are with the School of Information Science and Engineering, Shandong Normal University, Jinan 250358, China (e-mail: zhaoxuesdnu@163.com; zhangkainuc@163.com; jiandesun@hotmail.com; wanwenbo@sdnu.edu.cn; huaxzhang@sdnu.edu.cn).

Feng Zhang is with the School of Information Science and Engineering, University of Jinan, Jinan 250024, China (e-mail: fengzhangpl@163.com).

Digital Object Identifier 10.1109/JSTARS.2024.3360431

images. Although these methods produced good CD results, the complicated correlations among different objects in multitemporal images cannot be well depicted because the methods mentioned above, especially GCN models, can only represent the pairwise low-order relationships. However, multitemporal images contain more complex change patterns in terms of spatial and spectral properties. It is difficult for these methods to formulate the high-order dependencies in images.

Recently, hypergraph convolutional networks (HCNs) have emerged as powerful tools for modeling complex high-order relationships. For example, HCNs were employed in [17] to exploit the correlation among multimodal data. Furthermore, HCNs were extended in [18] to deal with the underlying relationships by encoding them with degree-free hyperedges. In [19], hypergraphs were also applied to the CD task, where hypergraphs were constructed by estimating spatial coupling neighbors. In [20], multiscale segmentation was utilized to find similar neighbors, and hyperedges were calculated from the features extracted by the pretrained U-Net. However, the existing methods, including HCNs, simply concatenate features directly ignoring the differences between them and face challenges in exploring correlations in spatial, spectral, and temporal dimensions.

To address these limitations and capture the high-order relationships in multitemporal images more efficiently, we propose a multiview hypergraph fusion network (MVHFNet) for CD, in which hypergraphs are constructed along spatial, spectral, and temporal dimensions for the modeling of high-order relationships in images. The proposed MVHFNet is specifically divided into three branches, including the spectral hypergraph learning (SpeHGL) branch, the spatial hypergraph learning (SpaHGL) branch, and the temporal hypergraph learning (TemHGL) branch. In these three branches, the spectral attention module (SpeAM), the spatial attention module (SpaAM), and the temporal enhancement module (TemEM) are designed to comprehensively extract the features in multitemporal images. Then, hypergraphs are constructed for each view, and the hypergraph convolutional (HGC) module is utilized to obtain higher order relationships for efficient learning of complex relationships among objects. Finally, we design the multiview feature fusion (MVF) module, which effectively integrates information from multiple views for the prediction of changed areas. The proposed MVHFNet is validated on three datasets, i.e., LEVIR-CD, SYSU-CD, and CLCD, and the results illustrate the effectiveness of the proposed MVHFNet due to the introduction of hypergraph learning (HGL). In summary, the contributions of the MVHFNet are as follows.

1) We propose an MVHFNet for CD that captures complex higher order relationships from spatial, spectral, and temporal views for CD.
2) We design a three-branch network, including SpaHGL, SpeHGL, and TemHGL, to extract the higher order relationships in multitemporal images. In these branches, hypergraphs are constructed, and hypergraph convolution is utilized for the learning of these relationships.
3) The proposed MVHFNet produces state-of-the-art CD performance on three benchmark datasets, including LEVIR-CD, SYSU-CD, and CLCD.

The rest of this article is organized as follows. Section II provides a detailed introduction to DNN- and GCN-based CD methods. In Section III, we describe the proposed MVHFNet in detail. Then, the experimental results on three datasets are presented in Section IV to demonstrate the CD performance of the MVHFNet. Finally, Section V concludes this article.

II. Related Work

A. DNN-Based CD Methods

Due to their outstanding performance, methods based on DNNs have seen extensive use in the field of CD in recent years. For example, Daudt et al. [21] developed three fully convolutional (FC) networks for CD. These networks were named FC-early fusion (FC-EF), FC-Siamese-concatenation (FC-Siam-conc), and FC-Siamese-difference (FC-Siam-diff). Since there are some differences among multitemporal HR remote sensing images, the two-branch network or the multibranch network is generally considered to extract discriminative features [3], [22], [23], [24]. In [25], a CNN-based two-branch network was introduced to address edge ambiguity issues in CD results. In [26], a dual-branch multiscale FC neural network was proposed, which can effectively detect the detailed changes of ground objects due to the introduction of multiscale features. In [27], a dual-discriminative metric network was constructed to measure the distance among features and infer changes through the discriminative implicit metric module. In [28], an innovative three-branch network was developed to acquire multitemporal image features and capture changes within the images. In addition, multiscale or multilevel feature representation was also introduced to boost the CD performance of the existing networks [29], [30], [31]. For example, a multilevel feature constraint fusion network was designed in [32], which imposed multiattention modules on the extraction of multilevel features. In [33], a multiscale network was proposed, which combined pyramid pooling and an attention mechanism to effectively leverage feature information from the original image while focusing on the change area. In [34], multiple cascaded attention blocks were embedded into the two-branch network to integrate the multiscale features more efficiently.

Besides, recurrent neural networks (RNNs) were also applied to the CD task due to their suitability for processing sequence data. In [35], a recurrent convolutional neural network was built to learn the temporal dependence in images. In [36], multilayer RNNs were incorporated into CNNs to simultaneously capture the spatial–spectral features in multitemporal images. Moreover, a semisupervised generative adversarial network (GAN) was developed by Jiang et al. [37] to jointly use the labeled and unlabeled data. In [38], the GAN was also employed to model spectral and spatial variations between multitemporal images. Transformer-based CD methods are also developed by exploiting the self- or cross-attention mechanisms in multitemporal images. For instance, a multilevel difference aggregation transformer was proposed in [39] to extract more informative deep features in terms of global properties. In [40], an asymmetric cross-attention network was presented, in which the CNN and the transformer were simultaneously introduced for the modeling of local and global features in images. Very recently, the

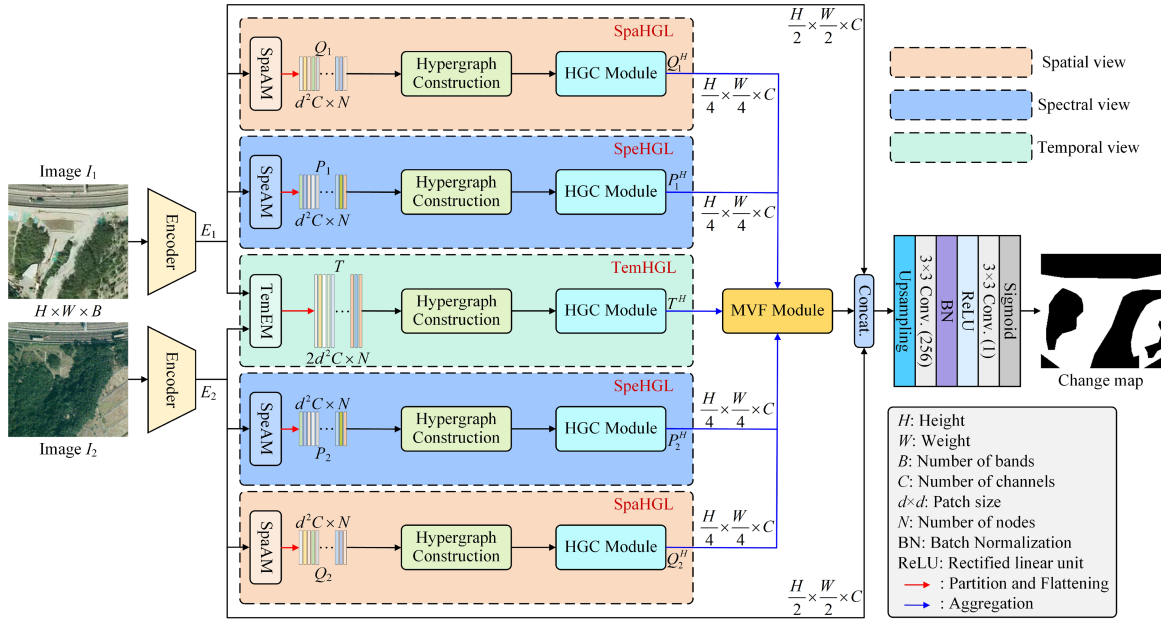

Fig. 1. Illustration of the proposed MVHFNet.

diffusion model has been considered in [41] and applied to the CD task, in which the excellent generative capacity of the diffusion model was mined.

B. GCN-Based CD Methods

In recent years, the GCN has begun to receive extensive interest from researchers due to its superior global modeling capability. For example, a multiscale fusion network model based on the GCN was proposed in [42], which learned richer features by aggregating the features of neighbors in the graph. In [43], a novel two-branch differential amplification GCN was designed to leverage the graph structure for capturing non-Euclidean features and preserving class boundaries. In [44], a robust graph mapping method was advanced for the heterogeneous CD problem. A two-branch framework based on spatiotemporal joint graph attention was proposed in [45], which extracted superpixel- and pixel-level features from multitemporal hyperspectral images. In [46], a stacked GCN was proposed to detect complex structural changes in multitemporal images. In [47], an unsupervised method based on autoencoders was developed for object-based CD, in which variational graphs were constructed for global modeling. A semisupervised GCN-based CD method was proposed in [48], which effectively captured spatial and temporal changes by segmenting multitemporal images into multiple patches and constructing graphs over these patches. Due to the more powerful global representation capability of hypergraphs, HCN-based CD methods also emerged. In [49], a hypergraph-based CD method was designed to extract change information by utilizing context-sensitive relationships among pixels. In [19], hypergraph matching and segmentation were applied to the CD of synthetic aperture radar images, in which a different image was generated by matching each vertex and hyperedge between two hypergraphs.

III. Multiview Hypergraph Fusion Network

In this section, we first introduce the proposed MVHFNet framework, as shown in Fig. 1. Then, we present the structures of TemHGL, SpaHGL, and SpeHGL modules in detail for the extraction of multiview features. Finally, these features from different views are integrated by the MVF module and decoded to obtain the estimated change map.

A. Overall Architecture

We denote the images from different phases as $I_1, I_2 \in {}^{H\times W\times B}$, where H and W are the height and width of the input images, respectively, and B is the number of bands in I_1 and I_2. In the proposed MVHFNet, we use the modified ResNet-18 as encoders for feature extraction. Through the corresponding encoders, we can obtain the features of I_1 and I_2, and they are denoted as $E_1 \in \mathbb{R}^{\frac{H}{2}\times\frac{W}{2}\times C}$ and $E_2 \in \mathbb{R}^{\frac{H}{2}\times\frac{W}{2}\times C}$, respectively. C is the number of channels in E_1 and E_2. For the modification of ResNet-18, we particularly remove the max-pooling layer and use pointwise convolution in the last layer to adjust the number of channels to C. Then, E_1 and E_2 are further fed into the proposed SpaHGL, SpeHGL, and TemHGL modules for multiview feature learning. Specifically, spatial features $Q_1^H, Q_2^H \in \mathbb{R}^{d^2C\times N}$ are obtained by the SpaHGL module. The features $P_1^H, P_2^H \in \mathbb{R}^{d^2C\times N}$ from spectral view are provided by the SpeHGL module. The temporal feature $T^H \in \mathbb{R}^{2\,d^2C\times N}$ is learned by the TemHGL module. To improve the network efficiency of training and inference, we perform downsampling on multiview features. Through these modules, multiview features are learned from I_1 and I_2 efficiently. Then, the MVF module is employed to aggregate these features, whose output is concatenated with E_1 and E_2 for the prediction of change map. Finally, the proposed MVHFNet is trained by minimizing the cross entropy between the predicted change map and the ground

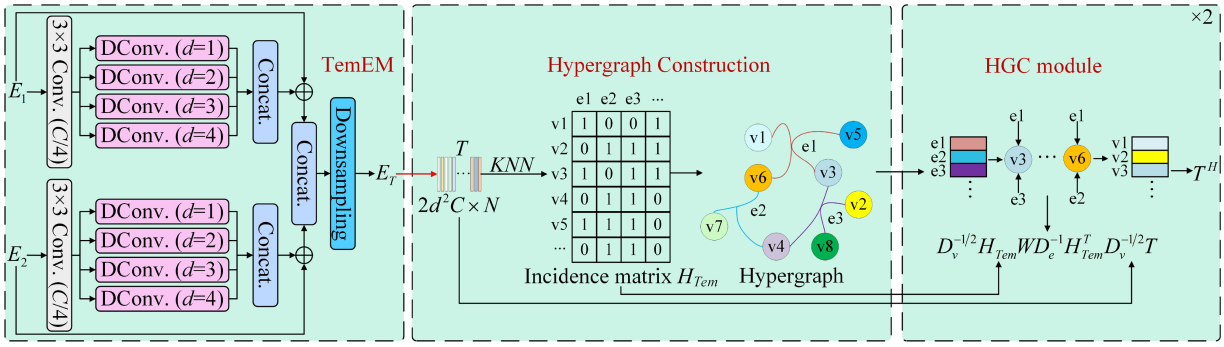

Fig. 2. Architecture of TemHGL.

truth (GT). The structures of all the modules in the proposed MVHFNet are detailed in the following subsections.

B. Temporal Hypergraph Learning

TemHGL is responsible for the learning of temporal information in I_1 and I_2, whose architecture is illustrated in Fig. 2. In TemHGL, the TemEM is first designed to extract the temporal information in images, which consists of two dilated convolution blocks. In the TemEM, the number of channels of E_1 and E_2 is condensed to $C/4$ by a 3×3 convolutional layer. Then, the convolutional layers with different dilation ratios are imposed on the condensed features of E_1 and E_2 for multiscale information extraction. The outputs of these dilated convolutional layers are concatenated and combined with the corresponding E_1 or E_2. In the TemEM, the two dilated convolution blocks share the same weights to reduce model size. Finally, all the features are concatenated and downsampled with a ratio of 2 to produce $E_T \in \mathbb{R}^{\frac{H}{4} \times \frac{W}{4} \times C}$.

1) Hypergraph Construction: For the enhanced temporal information E_T, we employ HGL to capture the high-order relationships along temporal dimension. To construct the temporal hypergraph, E_T is first segmented into nonoverlapping patches with the size of $d \times d$. Then, these patches are flattened and arranged as the matrix $T \in \mathbb{R}^{2d^2C \times N}$. According to the similarity among columns in T, the temporal hypergraph $\mathcal{G}_T(\mathcal{V}_T, \mathcal{E}_T, W_T)$ is constructed, in which $\mathcal{V}_T$ and $\mathcal{E}_T$ denote the vertex and hyperedge sets, respectively. A diagonal matrix W_T is utilized to assign a weight to each hyperedge. For simplicity, the diagonal elements in W_T are set as 1. Specifically, we consider each column in T as a vertex in $\mathcal{V}_T$. The similarity among vertices is measured by the Euclidean distance, and K-nearest neighbor is considered to select the neighbors. Here, we set K as 10. Then, the vertex and its K neighbors are connected by a hyperedge. Therefore, there are multiple vertices on one hyperedge. In this way, the high-order temporal relationships are modeled. Finally, we can obtain N hyperedges, and K+1 vertices are connected by each hyperedge. These high-order relationships are encoded in an incidence matrix $H_{\text{Tem}} \in \mathbb{R}^{N \times N}$, which is computed as

$$H_{\text{Tem}}(i,j) = \begin{cases} 1, & \text{if } v_i \in e_j \\ 0, & \text{else} \end{cases} \tag{1}$$

where v_i represents the ith vertex in $\mathcal{V}_T$, and e_j is the jth hyperedge in the hyperedge set $\mathcal{E}_T$. If the vertex v_i belongs to the hyperedge e_j, we set $H_{\text{Tem}}(i,j)$ as 1. Otherwise, it is 0. The hypergraph is contained in H_{Tem}.

2) Hypergraph Convolution: To embed the high-order temporal relationship into T, we construct an HGC module for the combination of T and H_{Tem}. Specifically, we first compute the degree matrix D_v of vertices and the degree matrix D_e of hyperedges from H_{Tem}. For D_v, it is defined as

$$D_v = \text{diag}\left(d\left(v_1\right), \ldots, d\left(v_i\right), \ldots, d\left(v_N\right)\right) \tag{2}$$

where $d(v_i) = \sum_{j=1}^{N} w_j H_{\text{Tem}}(i,j)$ represents the degree of each vertex, and w_j stands for the weight of the jth hyperedge. $\text{diag}(\cdot)$ denotes the diagonalization operation of a vector. In the same way, D_e is also calculated as

$$D_e = \text{diag}\left(d\left(e_1\right), \ldots, d\left(e_j\right), \ldots, d\left(e_N\right)\right) \tag{3}$$

where $d(e_j) = \sum_{i=1}^{N} H_{\text{Tem}}(i,j)$ is the degree of the jth hyperedge. By integrating D_v and D_e, the hypergraph convolution is defined as

$$T^H = D_v^{-1/2} H_{\text{Tem}} W D_e^{-1} H_{\text{Tem}}^T D_v^{-1/2} T. \tag{4}$$

Through the hypergraph convolution in (4), T is further enhanced by the high-order relationships at temporal view. To sufficiently depict the high-order relationships in images, two HGC layers are cascaded by which the ability to model complex change patterns in images is improved.

C. Spectral Hypergraph Learning

SpeHGL aims to learn the high-order relationships among multitemporal images along spectral dimensions. In SpeHGL, we design a SpeAM to extract the correlation among channels of E_1 and E_2. Fig. 3 presents the structure of SpeAM. In the SpeAM, max pooling and average pooling along spatial dimensions are performed on the input to produce $F_{\max}$ and F_{avg}. Then, a multilayer perception (MLP) is used to model the correlation in $F_{\max}$ and F_{avg}. To improve information interaction between $F_{\max}$ and F_{avg}, the outputs of the two MLPs are further combined to learn the correlation sufficiently. Through the interaction, we obtain $F_{\max}^a$ and F_{avg}^a, which are integrated as the attention map F. Finally, the spectral information in the input of SpeAM is

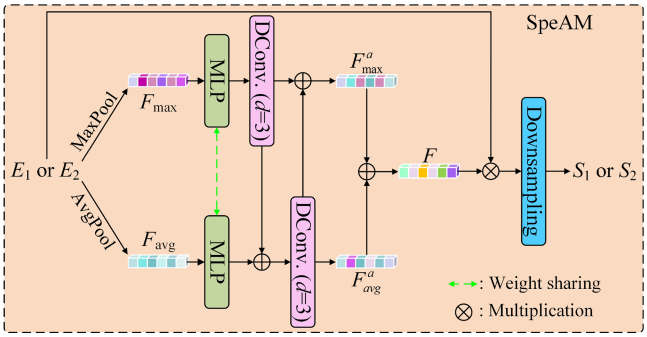

Fig. 3. Architecture of the SpeAM.

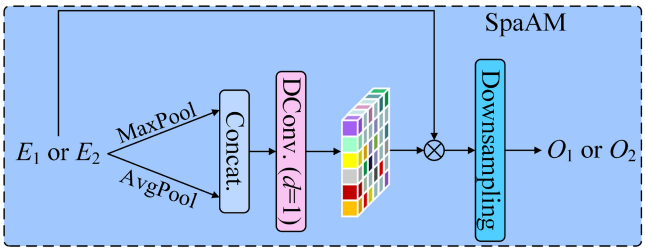

Fig. 4. Architecture of the SpaAM.

enhanced by F, and the outputs of SpeAM are denoted as S_1 and S_2, corresponding to E_1 and E_2.

To learn the high-order relationships at spectral view, we reshape S_1 and S_2 as $P_1 \in \mathbb{R}^{d^2C\times N}$ and $P_2 \in \mathbb{R}^{d^2C\times N}$, respectively, by partition and flattening. Here, the patch size is also set as $d \times d$. Similar to the formulation in TemHGL, HGL is also imposed on P_1 and P_2. Take P_1 for example; all the columns in P_1 consist of the vertex set, and an incidence matrix H_{Spe} can be inferred from these columns. Then, the hypergraph convolution on P_1 can be written as

$$P_1^H = A_v^{-1/2} H_{\text{Spe}} W_{\text{Spe}} A_e^{-1} H_{\text{Spe}}^T A_v^{-1/2} P_1 \tag{5}$$

where A_v and A_e are of the vertex and hypergraph degree matrices of H_{Spe}, respectively. W_{Spe} is the weight matrix of all the hyperedges. In the same way, the enhanced feature P_2^H also can be generated from P_2. Through SpeHGL, the high-order relationships along channel dimension are extracted, and complex patterns in terms of spectral changes can be modeled more efficiently.

D. Spatial Hypergraph Learning

In SpaHGL, the spatial high-order relationships are also learned by hypergraph convolution. To further enhance the spatial information in E_1 and E_2, a SpaAM is introduced into SpaHGL. Fig. 4. shows the structure of SpaAM. In the SpaAM, max pooling and average pooling are implemented on E_1 and E_2 along the channel dimension to extract the spatial features in E_1 and E_2. Then, the extracted features are concatenated and projected by a dilated convolutional layer for the generation of the spatial attention map. Finally, the spatial attention is combined with E_1 or E_2 to produce the corresponding $O_1 \in \mathbb{R}^{\frac{H}{4}\times\frac{W}{4}\times C}$ or $O_2 \in \mathbb{R}^{\frac{H}{4}\times\frac{W}{4}\times C}$. To meet the construction of hypergraphs, we divide O_1 and O_2 into N patches with the size of and assign them into the corresponding matrices $Q_1 \in \mathbb{R}^{d^2C\times N}$ and $Q_2 \in \mathbb{R}^{d^2C\times N}$. Then, hypergraphs are built by regarding all the columns in these matrices as vertices. Similar to the hypergraph convolution in Fig. 2, we can obtain $Q_1^H \in \mathbb{R}^{d^2C\times N}$ and $Q_2^H \in \mathbb{R}^{d^2C\times N}$, which contain the embedded high-order relationships in the spatial domain.

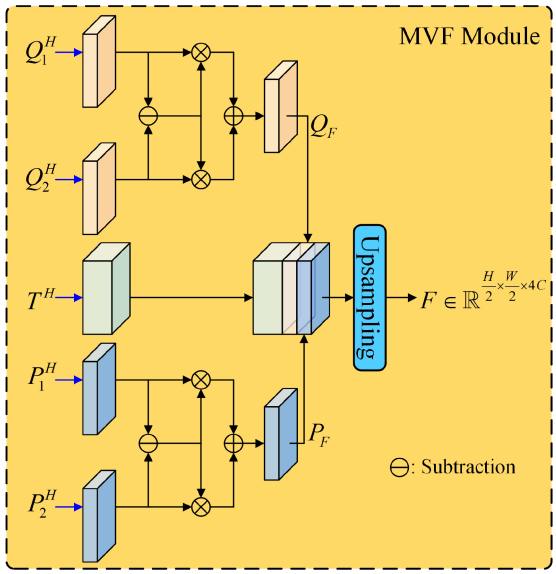

Fig. 5. Architecture of the MVF Module.

E. MVF Module

Through TemHGL, SpeHGL, and SpaHGL, we can obtain the higher order relationships at spectral, spatial, and temporal views jointly. Therefore, we design an MVF module as shown in Fig. 5 to integrate them efficiently. In this module, we combine the spatial features Q_1^H and Q_2^H, spectral features P_1^H and P_2^H, and temporal features T^H by the cross-temporal fusion strategy. Taking the spatial features as an example, the aggregated Q_1^H and Q_2^H are first subtracted to estimate the spatial difference feature. Then, the difference image is multiplied with the aggregated Q_1^H and Q_2^H to highlight the information in changed regions. Next, the information of changed regions in I_1 or I_2 is added to produce the fused spatial feature $Q_F \in \mathbb{R}^{\frac{H}{4}\times\frac{W}{4}\times C}$. In the same way, the fused spectral feature $P_F \in \mathbb{R}^{\frac{H}{4}\times\frac{W}{4}\times C}$ also can be acquired. Finally, all the fused features and the aggregated T^H are concatenated and upsampled as the output of the MVF module $F \in \mathbb{R}^{\frac{H}{2}\times\frac{W}{2}\times 4C}$. The fusion in Fig. 5 can be formulated as

$$F = \mathcal{U}\left(\text{Concat.}\left(Q_F, P_F, \mathcal{A}\left(T^H\right)\right)\right) \tag{6}$$

where $\mathcal{U}$ and $\mathcal{A}$ stand for the upsampling and aggregation operations, respectively. Through the MVF module, the features from different views are fused, in which the difference features between I_1 and I_2 are preserved.

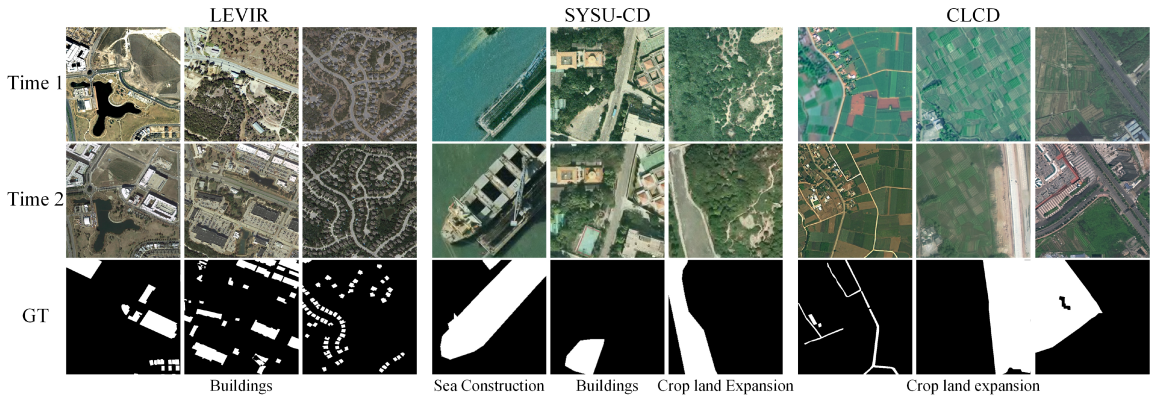

Fig. 6. Samples of the different change patterns in the three datasets.

F. Loss Function

In this work, we use the cross-entropy loss to evaluate model performance. For each sample i, the cross-entropy loss L_i can be calculated by the following equation:

$$L_{\mathrm{i}} = -\left[y_i \cdot \log\left(p_i\right) + \left(1 - y_i\right) \cdot \log\left(1 - p_i\right)\right] \tag{7}$$

where y_i is the GT, which takes the value 0 or 1 and indicates whether the sample is changed or not, and p_i is the predicted probability of the model. The overall cross-entropy loss is obtained by averaging the loss over all the samples

$$L = -\frac{1}{N}\sum_{i=1}^{N} L_i \tag{8}$$

where N is the number of samples.

IV. Experiments

In this section, the descriptions of all the datasets are first provided. Second, we introduce the evaluation indicators in detail and then briefly illustrate the implementation details of all the compared methods and the proposed MVHFNet. Subsequently, we analyze the comparative experimental results, and the effectiveness of MVHFNet is validated.

A. Datasets

In the experimental part, all the methods are implemented on three datasets with different change patterns. The datasets are LEVIR-CD [50], SYSU-CD [51], and CLCD [13]. As shown in Fig. 6, LEVIR-CD and CLCD datasets contain changes in terms of buildings and cropland, respectively. In the SYSU-CD dataset, there are three different kinds of change patterns, including buildings, construction within the sea area, and expansion of agricultural land.

1) LEVIR-CD: This dataset consists of 637 HR Google Earth image pairs, with a spatial resolution of 0.5 m and a size of 1024×1024 pixels. These images are composed of three bands: red (R), green (G), and blue (B). Considering the memory requirements, we downsampled the original images into patches with a size of 256×256. Then, the numbers of training, validation, and testing samples are 445, 64, and 128, respectively.

2) SYSU-CD: The dataset was collected in Hong Kong and contains 20 000 pairs of aerial images. For each image in this dataset, its size is 256×256, and the spatial resolution is 0.5 m. There are six different types of changes, and we select three main change patterns shown in Fig. 6 for training. In the training dataset, the number of images is 1200. Validation and testing datasets are 340 and 170, respectively.

3) CLCD: The CLCD dataset mainly covers the images of farmland changes collected in 2017 and 2019. This dataset consists of 600 images with a size of 512×512. The spatial resolutions of these images range from 0.5 to 2 m. Considering the insufficient GPU memory capacity, we downsampled the original images into patches with a size of 256×256. In the following experiments, these images are randomly divided into three groups: 360, 120, and 120 for training, validation, and testing of all the methods, respectively.

B. Evaluation Metrics

To evaluate the CD performance of all the methods, we use five quantitative indicators: precision (*P*), recall (*R*), *F1*-score (*F1*), intersection over union (IoU), and overall accuracy (OA). Their definitions are given as follows:

$$\begin{aligned} P &= (P_c + P_{uc})/2 \\ R &= (R_c + R_{uc})/2 \\ F1 &= P_c R_c (P_c + R_c) + P_{uc} R_{uc} (P_{uc} + R_{uc}) \\ \mathrm{IoU} &= \mathrm{TP}(\mathrm{TP} + \mathrm{FP} + \mathrm{FN}) + \mathrm{TN}(\mathrm{TN} + \mathrm{FN} + \mathrm{FP}) \\ \mathrm{OA} &= (\mathrm{TP} + \mathrm{TN})/\mathrm{TP} + \mathrm{TN} + m + \mathrm{FN} \end{aligned} \tag{9}$$

where P_c and P_{uc} represent the precision of detecting changed and unchanged regions, respectively. TP, TN, FP, and FN represent the quantities of true positives, true negatives, false positives, and false negatives, respectively. These metrics collectively reflect the overall detection performance of the model.

C. Compared Methods and Implementation Details

In the following experiments, the MVHFNet is compared with nine methods, including the FC-EF [21], the FC-Siam-conc [21], the FC-Siam-diff [21], the dual-task-constrained deep Siamese

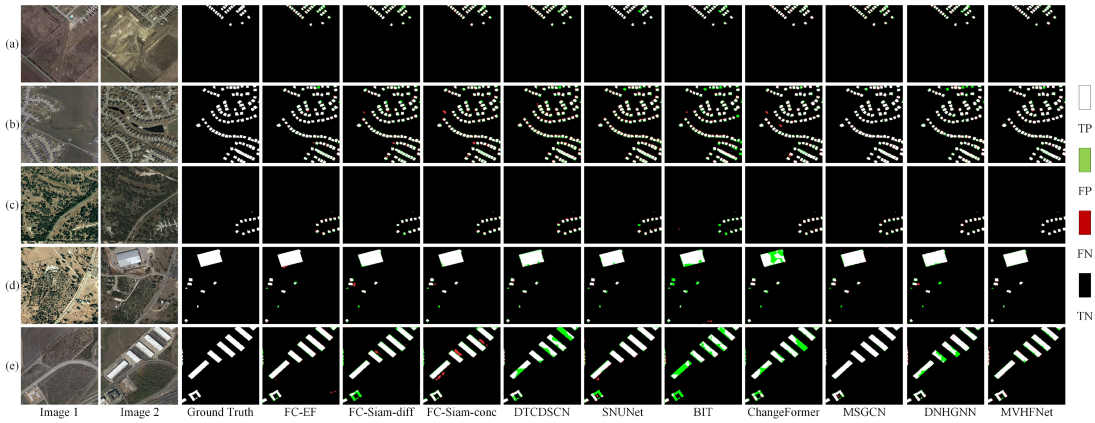

Fig. 7. (a)–(e) Visualization of all the methods on the LEVIR-CD dataset.

convolutional network (DTCDSCN) [52], the combination of Siamese network and NestedUNet (SNUNet) [53], the bitemporal image transformer (BIT) [54], the transformer-based CD (ChangeFormer) [55], a multiscale graph convolutional network (MSGCN) [15], and a dual-neighborhood hypergraph neural network (DNHGNN) [20]. A brief introduction to these methods is as follows.

1) *FC-EF:* This method treats the concatenated multitemporal images as the images with more channels, which is regarded as the input of the FC neural network.
2) *FC-Siam-diff:* The multitemporal images are processed separately by a dual-branch network with shared structures and parameters. Then, feature differences among dual-branch networks are utilized to infer the change areas.
3) *FC-Siam-conc:* This network uses the Siamese FC network to extract multilevel features, which are then fused by direct concatenation.
4) *DTCDSCN:* This network concatenates the multiscale features, which are extracted by channel attention and SpaAMs to improve their discriminative.
5) *SNUNet:* In SNUNet, Siamese network and NestedUNet are combined to extract high-level features. This method further applies channel attention and deep supervision to enhance the effectiveness of features.
6) *BIT:* BIT is proposed based on the transformer, which uses tokens to obtain deep semantic information and enhance the semantic information through the cross-attention mechanism.
7) *ChangeFormer:* This method embeds a hierarchical transformer encoder with an MLP into the Siamese network architecture to learn the global features in multitemporal images.
8) *MSGCN:* This network combines the GCN and multiscale features to overcome the difficulty of existing models in modeling ground object features with different modalities.
9) *DNHGNN:* In the DNHGNN, image segmentation is considered to construct the hypergraph, and the hypergraph neural network is used to extract changed features of nodes by combining feature maps at fine and high scales.

All the methods were trained using cross validation on NVIDIA 2080Ti GPU. The proposed MVHFNet is optimized through stochastic gradient descent. The initial learning rate is set to 0.01, and then, the learning rate is decayed every 100 epochs. Training for the proposed MVHFNet is considered complete when the number of epochs reaches 1000. For other compared methods, we also set the number of epochs to 1000. Besides, the patch size $d \times d$ during hypergraph construction is set to 4×4.

D. Comparison to State-of-the-Art Methods

1) Experiments on the LEVIR-CD Dataset: In this part, the experiments are conducted on the LEVIR-CD dataset. For visual analysis, we randomly select some samples, and their CD results are shown in Fig. 7. TP, TN, FP, and FN are represented in white, black, green, and red, respectively. As shown in Fig. 7(b) and (e), FC-Siam-diff and FC-Siam-conc have significantly more red regions, resulting in larger FN. We believe that multiview feature extraction enables our proposed MVHFNet to maintain the shape and edge information of changing buildings. For example, in Fig. 7(d), other methods have some issues in terms of incomplete internal structure of buildings, while FC-EF has more FN regions. Moreover, hypergraphs can capture high-order relationships among objects, so the results of MVHFNet produce fewer error regions. As shown in Fig. 7(a) and (b), it is particularly important to obtain high-order relationships among buildings because the buildings in these scenes are very close. It can be seen that the MVHFNet has fewer FP areas compared to other methods. Overall, it can be seen intuitively that the MVHFNet has achieved the best detection results.

Table I reports the quantitative results of all the methods on this dataset. The SNUNet achieved the highest *P* value, while the DNHGNN attained the highest OA. Our method obtained the highest values in the other three metrics. However, in general, these metrics can reflect that our method achieves fine-grained change detection.

2) Experiments on the SYSU-CD Dataset: In this part, we compare all the methods on the SYSU-CD dataset. Fig. 8 shows the visual results of some typical samples from the SYSU-CD

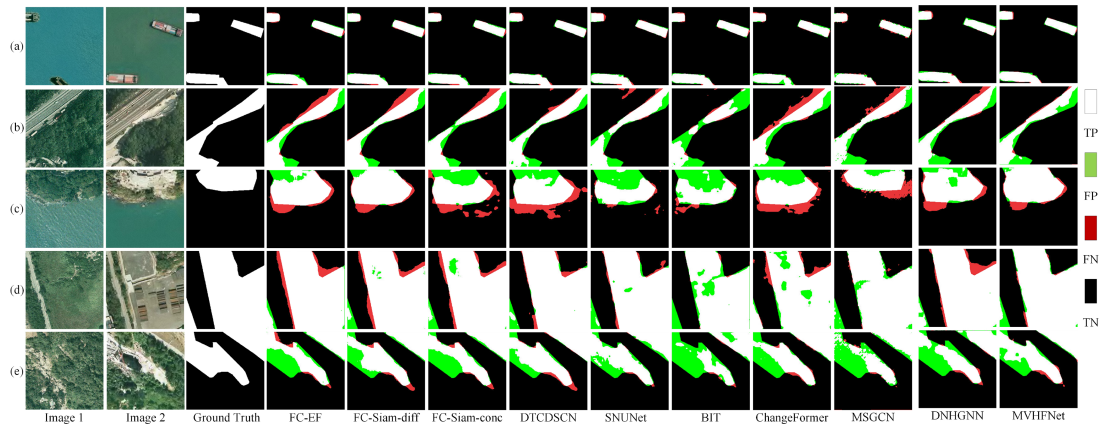

Fig. 8. (a)–(e) Visualization of all the methods on the SYSU-CD dataset.

TABLE I
QUANTITATIVE EVALUATION OF THE LEVIR-CD DATASET

Metric (%)	*P*	*R*	*F1*	IoU	OA
FC-EF	91.15	89.57	90.34	83.54	98.17
FC-Siam-diff	92.93	87.02	89.73	82.67	98.15
FC-Siam-conc	92.03	89.82	90.89	84.34	98.28
DTCDSCN	91.06	87.69	89.30	82.05	98.01
SNUNet	**93.83**	90.11	91.88	85.84	98.50
BIT	90.27	83.37	86.46	78.23	97.60
ChangeFormer	91.46	86.31	88.69	81.21	97.95
MSGCN	90.28	87.79	88.99	81.61	97.94
DNHGNN	93.46	90.86	92.11	86.18	**98.52**
MVHFNet	92.03	**92.98**	**92.49**	**86.72**	98.12

The best values are labelled in bold.

TABLE II
QUANTITATIVE EVALUATION OF THE SYSU-CD DATASET

Metric (%)	*P*	*R*	*F1*	IoU	OA
FC-EF	91.01	87.75	89.26	81.30	93.69
FC-Siam-diff	93.10	87.69	90.07	82.54	94.11
FC-Siam-conc	92.63	87.56	89.81	82.13	93.94
DTCDSCN	92.07	87.45	89.52	81.68	93.74
SNUNet	92.86	86.73	89.39	81.53	93.97
BIT	91.41	78.35	82.88	72.54	91.12
ChangeFormer	91.34	83.41	86.66	77.55	92.60
MSGCN	92.59	88.55	90.36	82.98	94.22
DNHGNN	**93.35**	88.91	90.91	83.85	94.55
MVHFNet	93.16	**89.27**	**90.99**	**84.06**	**95.21**

The best values are labelled in bold.

datasets. In Fig. 8, we present five change patterns. Fig. 8(a) contains the change of ships at sea area, and we can see that all the methods show good performance on this image pair. Meanwhile, our proposed MVHFNet is slightly better than other methods in terms of FP and FN. In Fig. 8(b) and (c), the change areas are focused on roads and the construction in the sea. In these results, it can be seen that the MVHFNet has the least amount of red areas, which corresponds to a smaller FN. This may be because the MVHFNet extracts more discriminative features by hypergraph modeling. In Fig. 8(d), there is a large area of building changes. We can find that the MVHFNet and the DNHGNN have fewer undetected parts thanks to hypergraph neural network (HNN). However, the result of DNHGNN has more FN regions. Fig. 8(e) shows the farmland change, and the proposed MVHFNet also has the least FP and FN. In conclusion, the MVHFNet also shows better detection performance on the SYSU-CD dataset with more complex change patterns.

All the evaluation values on the dataset are shown in Table II. This dataset contains more change types than the LEVIR-CD dataset. The results imply that the proposed MVHFNet can find more complicated change patterns than other compared methods due to the introduction of HGL.

3) Experiments on the CLCD Dataset: This part presents the experimental results of all the methods on the CLCD dataset quantitatively and qualitatively. Fig. 9 shows some CD results of all the methods on the CLCD dataset. In this dataset, the types of changes mainly include road and building changes. For road changes shown in Fig. 9(a) and (d), the road is more likely to merge with the background and cause pseudo changes. It can be seen that the MVHFNet produces fewer FP regions compared to FC-EF, FC-Siam-diff, and SNUNet. In Fig. 9(b), there is a wider road. Therefore, high-order information modeling is more important for the detection of this kind of change. It can be seen that the MVHFNet has lower FN and more complete detection performance results compared to BIT and ChangeFormer. Fig. 9(c) and (e) is related to building changes, and the MVHFNet has also achieved good performance in terms of FN. Table III provides the quantitative results of all the methods on this dataset. The OA value for the MVHFNet is not the highest but is very close to that of the DNHGNN. Thus, the proposed method has better overall performance.

E. Ablation Study

In this subsection, we evaluate the effectiveness of each module in the MVHFNet on three datasets, including encoders, multiview feature branches, and the MVF Module.

1) Effectiveness of Encoders: In the MVHFNet, we use the modified ResNet-18 as an encoder for feature extraction. In this part, ResNet-18 is replaced by ResNet-50 to analyze the performance of different backbones. For a more intuitive illustration, we randomly select some CD results from different datasets and

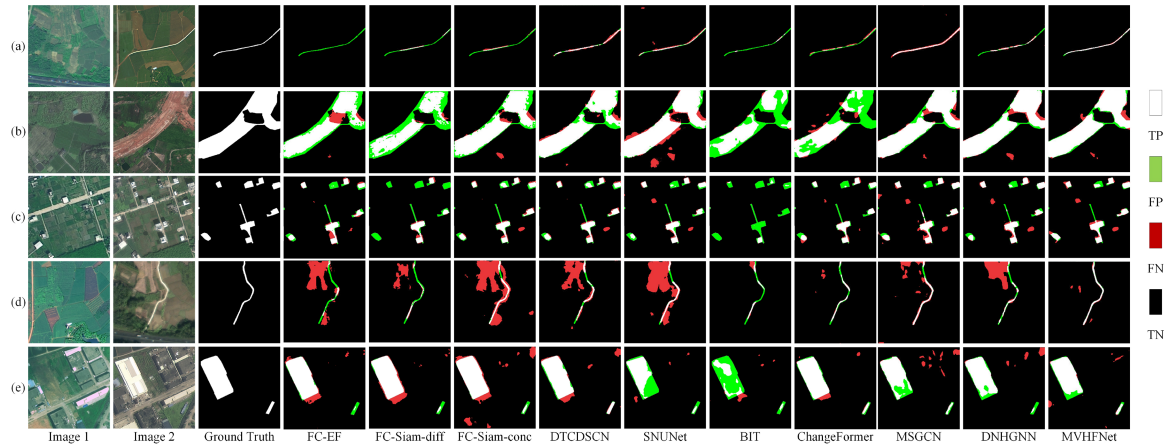

Fig. 9. (a)–(e) Visualization of all the methods on the CLCD dataset.

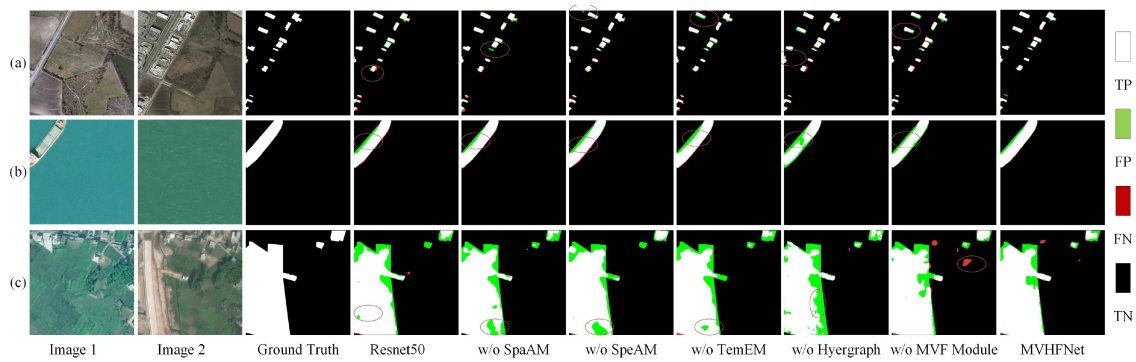

Fig. 10. Visualization of ablation experiments. (a) LEVIR-CD. (b) SYSU-CD. (c) CLCD.

TABLE III
QUANTITATIVE EVALUATION OF THE CLCD DATASET

Metric (%)	*P*	*R*	*F1*	IoU	OA
FC-EF	82.83	76.20	79.06	69.19	94.81
FC-Siam-diff	83.13	74.60	78.10	68.19	94.73
FC-Siam-conc	81.65	79.94	80.77	70.98	94.84
DTCDSCN	85.54	81.75	83.51	74.28	95.70
SNUNet	85.92	82.62	84.19	75.11	95.84
BIT	81.42	70.27	74.39	64.51	94.17
ChangeFormer	84.41	81.36	82.80	73.40	95.47
MSGCN	86.73	78.68	82.21	72.65	95.61
DNHGNN	86.77	82.28	84.34	75.31	**95.95**
MVHFNet	**87.45**	**82.92**	**85.01**	**76.05**	95.69

The best values are labelled in bold.

TABLE IV
ABLATION STUDY OF RESNET ON THE LEVIR-CD, SYSU-CD, AND CLCD DATASETS

Datasets	ResNet	*P*	*R*	*F1*	IoU	OA
LEVIR-CD	ResNet-50	91.83	92.87	92.34	86.48	98.07
	ResNet-18	**92.03**	**92.98**	**92.49**	**86.72**	**98.12**
SYSU-CD	ResNet-50	91.02	86.41	88.47	80.14	93.34
	ResNet-18	**93.16**	**89.27**	**90.99**	**84.06**	**95.21**
CLCD	ResNet-50	86.84	80.11	83.07	73.77	**95.75**
	ResNet-18	**87.45**	**82.92**	**85.01**	**76.05**	95.69

The best values are labelled in bold.

show them in Fig. 10. Some interesting regions are circled in red. From these regions, we can observe that the results of ResNet-50 suffer from performance degradation and have more FN areas.

The results in Table IV show that higher values in terms of *P*, *R*, *F1*, and IoU are achieved when ResNet-18 is considered. For ResNet-50, better OA values are provided on CLCD datasets. The reason for this may be that ResNet-50 is larger than ResNet-18 and involves more parameters to be learned. ResNet-50 may not be trained sufficiently on these datasets. Besides, considering the higher computational complexity of ResNet-50, ResNet-18 is regarded as the encoder to extract spatial and spectral features from the corresponding inputs.

2) Impact of Three Attention Modules: First, we verify the validity of SpaAM, SpeAM, and TemEM in all three branches. Specifically, we remove SpeAM, SpaAM, and TemEM sequentially. In this way, hypergraphs in the three HGL branches are directly constructed on the features from encoders E_1 and E_2. As we can see from the red circles in Fig. 10, there are more FP areas in CD results when SpeAM and SpaAM are ablated. Some changed regions cannot be inferred when we remove the TemEM. The reason for this is the lack of interaction among multitemporal images. Thus, the above experimental

TABLE V
ABLATION STUDY OF THE SPAAM, SPEAM, AND TEMEM ON LEVIR-CD, SYSU-CD, AND CLCD DATASETS

Datasets	Settings	*P*	*R*	*F1*	IoU	OA
LEVIR-CD	w/o SpaAM	89.57	93.25	91.31	84.89	97.74
	w/o SpeAM	89.45	**93.35**	91.29	84.85	97.73
	w/o TemEM	89.61	92.38	90.94	84.33	97.67
	MVHFNet	**92.03**	92.98	**92.49**	**86.72**	**98.12**
SYSU-CD	w/o SpaAM	91.32	87.24	89.09	81.06	93.66
	w/o SpeAM	91.07	87.66	89.23	81.26	93.69
	w/o TemEM	91.31	87.35	89.15	81.15	93.69
	MVHFNet	**93.16**	**89.27**	**90.99**	**84.06**	**95.21**
CLCD	w/o SpaAM	**88.93**	80.80	84.28	75.12	95.43
	w/o SpeAM	88.29	82.16	84.89	75.87	95.49
	w/o TemEM	87.97	82.60	**85.03**	76.04	95.49
	MVHFNet	87.45	**82.92**	85.01	**76.05**	**95.69**

The best values are labelled in bold.

TABLE VI
GRAPH VERSUS HYPERGRAPH ON LEVIR-CD, SYSU-CD, AND CLCD DATASETS

Datasets	Settings	*P*	*R*	*F1*	IoU	OA
LEVIR-CD	With graph	89.65	89.95	89.80	82.75	98.02
	MVHFNet	**92.03**	**92.98**	**92.49**	**86.72**	**98.12**
SYSU-CD	With graph	91.90	85.45	88.21	79.77	93.33
	MVHFNet	**93.16**	**89.27**	**90.99**	**84.06**	**95.21**
CLCD	With graph	83.97	77.16	80.01	70.10	94.19
	MVHFNet	**87.45**	**82.92**	**85.01**	**76.05**	**95.69**

The best values are labelled in bold.

results indicate that our proposed MVHFNet can obtain more effective spatial, spectral, and temporal features enabling the CD performance better. Table V reports the numerical results of MVHFNet with different architectures, where "w/o" indicates that the corresponding module is removed from the MVHFNet. It can be seen that the complete MVHFNet has better overall metrics although a few best values are not from the complete MVHFNet.

3) Graph Versus Hypergraph: To verify the effectiveness of HGL, we replace the hypergraph convolution with graph convolution and constructed an adjacency matrix to represent the relationships among vertices. From Fig. 10, we also can find that the CD results obtained by the MVHFNet with the GCN have more FP regions, which are circled by red circles. Table VI provides the evaluation results of different relationship modeling tools. As can be seen in Table VI, the MVHFNet with HGL is better than that with the GCN on the three datasets. Since the relationships between features are much more complicated, the GCN cannot depict the changing relationships among features sufficiently and also produces worse detections.

4) Ablation of MVF Module: Finally, we replace the MVF module with concatenation to further validate its influence on all the datasets. Fig. 10 demonstrates that more FP regions are produced because the difference information is not extracted by the concatenation operation. Table VII shows the metric values on the three datasets. The overall performance in Table VII indicates that the direct concatenation of multiview features suffers from some performance degradation. Therefore, the MVF module can more efficiently integrate the spatial, spectral, temporal in multiview features.

TABLE VII
ABLATION STUDY OF THE MVF MODULE ON LEVIR-CD, SYSU-CD, AND CLCD DATASETS

Datasets	Setting	*P*	*R*	*F1*	IoU	OA
LEVIR-CD	w/o MVF	90.27	**93.33**	91.74	85.54	97.87
	MVHFNet	**92.03**	92.98	**92.49**	**86.72**	**98.12**
SYSU-CD	w/o MVF	92.08	84.61	87.72	79.07	93.14
	MVHFNet	**93.16**	**89.27**	**90.99**	**84.06**	**95.21**
CLCD	w/o MVF	86.31	**85.15**	**85.72**	**76.89**	95.45
	MVHFNet	**87.45**	82.92	85.01	76.05	**95.69**

The best values are labelled in bold.

TABLE VIII
ANALYSIS OF PATCH SIZE ON LEVIR-CD, SYSU-CD, AND CLCD DATASETS

Datasets	Settings	*P*	*R*	*F1*	IoU	OA
LEVIR-CD	4×4	**92.03**	**92.98**	**92.49**	**86.72**	**98.12**
	8×8	91.10	88.87	89.95	82.98	98.11
SYSU-CD	4×4	**93.16**	**89.27**	**90.99**	**84.06**	**95.21**
	8×8	92.76	85.60	88.62	80.39	93.61
CLCD	4×4	87.45	**82.92**	**85.01**	**76.05**	**95.69**
	8×8	**88.11**	81.52	84.42	75.27	95.37

The best values are labelled in bold.

TABLE IX
ANALYSIS OF HGCN LAYERS ON LEVIR-CD, SYSU-CD, AND CLCD DATASETS

Datasets	Settings	*P*	*R*	*F1*	IoU	OA
LEVIR-CD	One layer	**92.25**	93.03	92.63	86.94	**98.15**
	Two layers	92.03	**92.98**	**92.49**	**86.72**	98.12
SYSU-CD	One layer	91.39	86.54	88.70	80.48	93.49
	Two layers	**93.16**	**89.27**	**90.99**	**84.06**	95.21
CLCD	One layer	**87.99**	82.64	**85.05**	**76.07**	95.48
	Two layers	87.45	**82.92**	85.01	76.05	**95.69**

The best values are labelled in bold.

F. Analysis of Patch Size and HGCN Layers

This section discusses the impact of constructing hypergraphs using patches with different sizes and the number of hypergraph convolution network (HGCN) layers. As can be seen from Fig. 11 , there are obvious undetected regions when the patch size is set as 8×8, and only one HGCN layer is adopted. This may be because the patch size is too large, resulting in an insufficient detection region. Thus, the patch size is finally set as 4×4. Table VIII gives the metric values for different patch sizes on the three datasets. It can be seen that the best metrics are obtained when the patch size is 4×4. Table IX shows that using two layers of HGCN gives higher metrics than one layer, while more layers will increase training time and lead to oversmoothing problems. Therefore, we use two layers of HGCN finally.

G. Complexity Analysis

Table X presents model sizes and Giga floating-point operations per second (GFLOPs) of all the methods on the NVIDIA 2080Ti GPU, using image pairs with a resolution of 256×256.

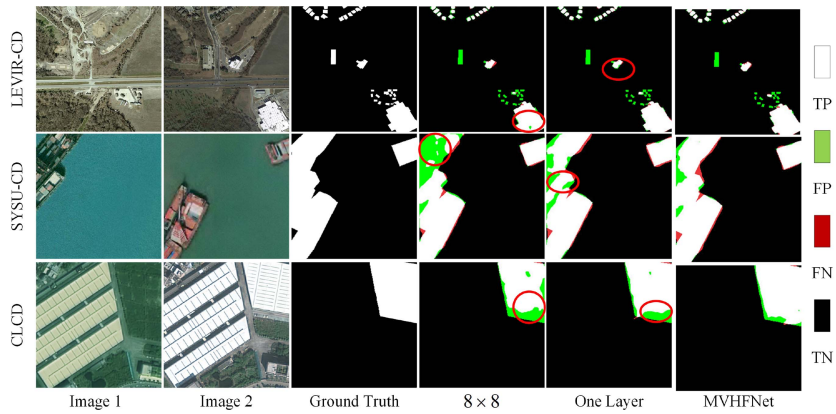

Fig. 11. Comparison of different patch sizes on LEVIR-CD, SYSU-CD, and CLCD datasets.

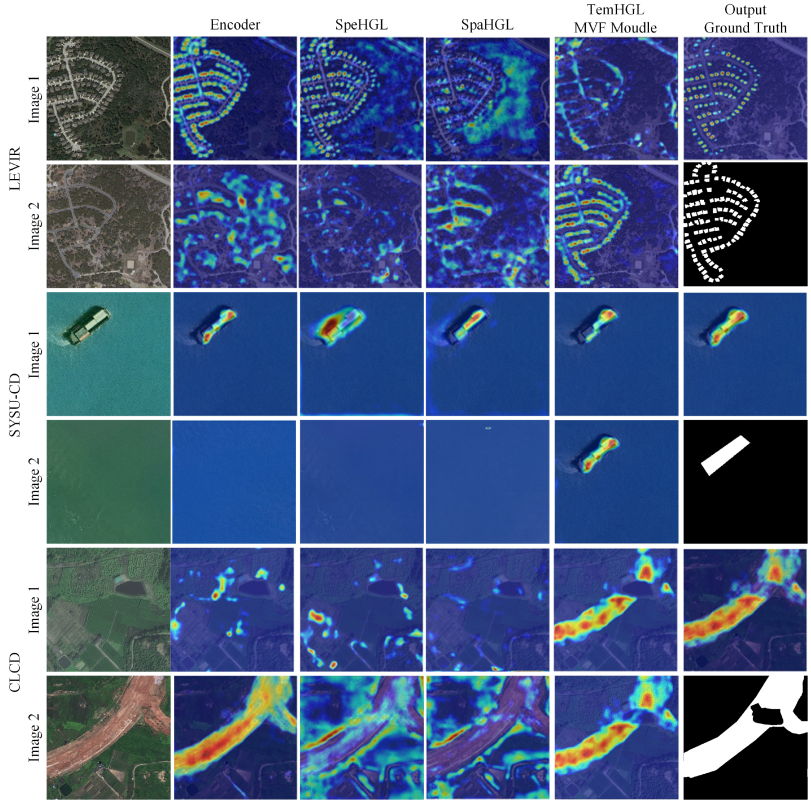

Fig. 12. Feature visualization on LEVIR-CD, SYSU-CD, and CLCD datasets.

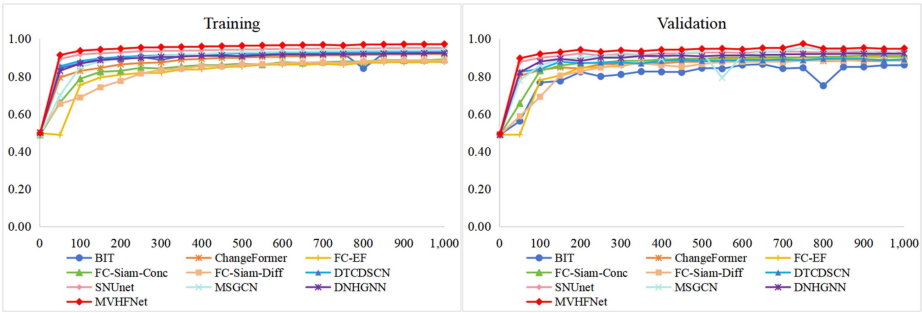

Fig. 13. *F1*-score of all the methods for each epoch on the LEVIR-CD datasets.

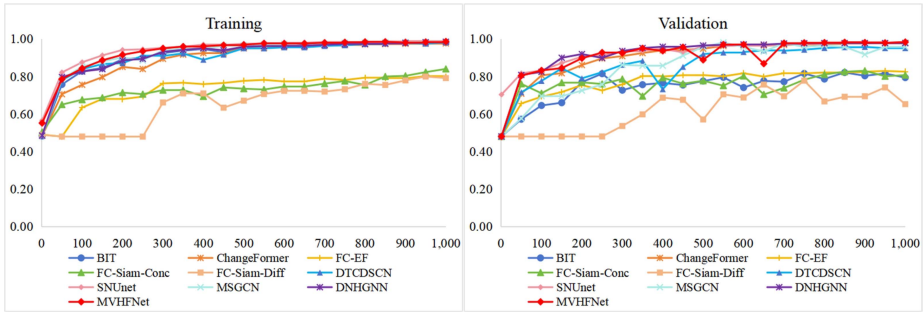

Fig. 14. *F1*-score of all the methods for each epoch on the SYSU-CD datasets.

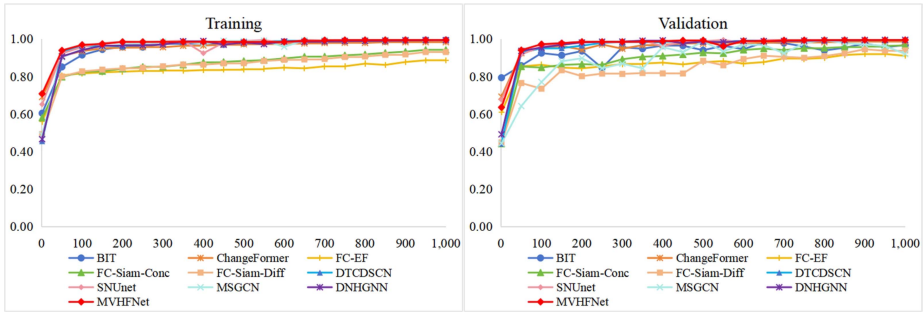

Fig. 15. *F1*-score of all the methods for each epoch on the CLCD datasets.

TABLE X
COMPARISON OF MODEL SIZES AND GFLOPS

Methods	Para.(M)	GFLOPs
FC-EF	1.351	3.572
FC-Siam-Di	1.350	4.717
FC-Siam-Conc	1.546	5.321
DTCDSCN	31.257	13.207
SNUNet	12.035	54.817
BIT	3.056	8.881
Changeformer	41.005	202.829
MSGCN	39.516	58.875
DNHGNN	31.110	55.475
MVHFNet	24.468	93.154

For model size, compared with lightweight networks (FC-EF, FC-Siam-Di, etc.), the proposed network has more parameters but achieves better detection, and compared with other methods such as DTCDSCN and DNHGNN, the MVHFNet has a relatively small model size, achieving a balance between model size and CD accuracy.

H. Feature Visualization

To better understand multiview feature learning in the MVHFNet, we used Grad-CAM to visualize the output of each module on the three datasets. Fig. 12 illustrates the outputs of encoders, three HGL branches, and MVF module. Besides, the final output is also visualized. The network is a single branch starting from the TemHGL module. In Fig. 12, the outputs of TemHGL and MVF modules are plotted in the first and second rows, respectively. It can be seen that encoders extract the difference features among multitemporal images. SpaHGL and SpeHGL capture the complex spatial and spectral relationships in images, respectively. In addition, TemHGL obtains the temporal differences in images. Moreover, all the multiview features are fused by the MVF module, and more reliable results can be found from its output. Therefore, the heat maps in Fig. 12 also demonstrate the effectiveness of multiview learning in the MVHFNet.

I. Training Convergence

Figs. 13–15 show the training and validation processes of all the methods on LEVIR-CD, SYSU-CD, and CLCD datasets, where *F1*-score is used as an indicator. From Fig. 13, we can see that all these values have a similar growth trend on this dataset, whereas there are some fluctuations on SYSU-CD and CLCD datasets for most compared methods, as shown in Figs. 14 and 15. This may be because these two datasets have more types of changes than the LEVIR-CD dataset, and thus, the CD performance is not as effective as the LEVIR-CD dataset. Compared to other methods, our proposed MVHFNet has a steady growth trend on all three datasets. When the number of epochs reaches 1000, the *F1*-score converges, and the validation performance has no more significant improvement. Therefore, the maximum number of epochs is set to 1000.

V. CONCLUSION

This article proposed an MVHFNet for CD. In this network, SpaHGL, SpeHGL, and TemHGL were designed to learn spatial, spectral, and temporal higher order information, respectively. Among them, SpaAM, SpeAM, and TemEM were embedded to extract the features of multitemporal images from three views.

To effectively fuse the higher order information from multiple views, we designed the MVF module and condensed the concatenated features for further prediction. Finally, the change map was obtained by a prediction head. The subjective and objective evaluation results on three datasets demonstrated that MVHFNet produces higher values in terms of *F1*, IoU, and OA, which benefit from the efficient high-order relationship modeling by HGL.

REFERENCES

[1] K. Zhang et al., "Panchromatic and multispectral image fusion for remote sensing and earth observation: Concepts, taxonomy, literature review, evaluation methodologies and challenges ahead," *Inf. Fusion*, vol. 937, pp. 227–242, 2023.

[2] W. Zhang, J. Li, F. Zhang, J. Sun, and K. Zhang, "Unsupervised change detection of multispectra images based on PCA and low-rank prior," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 5001605.

[3] Y. Zhou, Y. Song, S. Cui, H. Zhu, J. Sun, and W. Qin, "A novel change detection framework in urban area using multilevel matching feature and automatic sample extraction strategy," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 3967–3987, 2021.

[4] W. Zhao, A. Li, X. Nan, Z. Zhang, and G. Lei, "Postearthquake landslides mapping from Landsat-8 data for the 2015 Nepal earthquake using a pixel-based change detection method," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 10, no. 5, pp. 1758–1768, May 2017.

[5] Z. Zheng, Y. Zhong, J. Wang, A. Ma, and L. Zhang, "Building damage assessment for rapid disaster response with a deep object-based semantic change detection framework: From natural disasters to man-made disasters," *Remote Sens. Environ.*, vol. 265, 2021, Art. no. 112636.

[6] D. Peng, L. Bruzzone, Y. Zhang, H. Guan, and P. He, "SCDNET: A novel convolutional network for semantic change detection in high resolution optical remote sensing imagery," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 103, 2021, Art. no. 102465.

[7] X. Wang et al., "Double U-Net (W-Net): A change detection network with two heads for remote sensing imagery," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 122, 2023, Art. no. 103456.

[8] H. Zheng et al., "HFA-Net: High frequency attention siamese network for building change detection in VHR remote sensing images," *Pattern Recogn.*, vol. 129, 2022, Art. no. 108717.

[9] X. He, S. Zhang, B. Xue, T. Zhao, and T. Wu, "Cross-modal change detection flood extraction based on convolutional neural network," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 117, 2023, Art. no. 103197.

[10] K. Zhang, X. Zhao, F. Zhang, L. Ding, J. Sun, and L. Bruzzone, "Relation changes matter: Cross-temporal difference transformer for change detection in remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5611615.

[11] Q. Li, R. Zhong, X. Du, and Y. Du, "TransUNetCD: A hybrid transformer network for change detection in optical remote-sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5622519.

[12] X. Xu, J. Li, and Z. Chen, "TCIANet: Transformer-based context information aggregation network for remote sensing image change detection," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 1951–1971, 2023.

[13] M. Liu, Z. Chai, H. Deng, and R. Liu, "A CNN-transformer network with multiscale context aggregation for fine-grained cropland change detection," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 15, pp. 4297–4306, 2022.

[14] X. Tang et al., "An unsupervised remote sensing change detection method based on multiscale graph convolutional network and metric learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5609715.

[15] 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, 2021, Art. no. 102615.

[16] M. Jia, C. Zhang, Z. Zhao, and L. Wang, "Bipartite graph attention autoencoders for unsupervised change detection using VHR remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5626215.

[17] Y. Feng, H. You, Z. Zhang, R. Ji, and Y. Gao, "Hypergraph neural networks," in *Proc. AAAI Conf. Artif. Intell.*, 2019, pp. 3558–3565.

[18] Y. Gao, Y. Feng, S. Ji, and R. Ji, "HGNN+: General hypergraph neural networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 3, pp. 3181–3199, Mar. 2023.

[19] J. Wang, X. Yang, X. Yang, L. Jia, and S. Fang, "Unsupervised change detection between SAR images based on hypergraphs," *ISPRS J. Photogrammetry Remote Sens.*, vol. 164, pp. 61–72, 2020.

[20] J. Wu et al., "A dual neighborhood hypergraph neural network for change detection in VHR remote sensing images," *Remote Sens.*, vol. 15, 2023, Art. no. 694.

[21] R. C. Daudt, B. L. Saux, and A. Boulch, "Fully convolutional siamese networks for change detection," in *Proc. IEEE 25th Int. Conf. Image Process.*, 2018, pp. 4063–4067.

[22] Y. Feng, J. Jiang, H. Xu, and J. Zheng, "Change detection on remote sensing images using dual-branch multilevel intertemporal network," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 4401015.

[23] H. Zhang, S. Qu, and H. Li, "Dual-branch enhanced network for change detection," *Arab. J. Sci. Eng.*, vol. 47, pp. 3459–3471, 2022.

[24] Y. Quan et al., "Unified building change detection pre-training method with masked semantic annotations," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 120, 2023, Art. no. 103346.

[25] C. Ma, L. Weng, M. Xia, H. Lin, M. Qian, and Y. Zhang, "Dual-branch network for change detection of remote sensing image," *Eng. Appl. Artif. Intell.*, vol. 123, 2023, Art. no. 106324.

[26] X. Li, M. He, H. Li, and H. Shen, "A combined loss-based multiscale fully convolutional network for high-resolution remote sensing image change detection," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2021, Art. no. 8017505.

[27] X. Li, L. Yan, Y. Zhang, and N. Mo, "SDMNet: A deep-supervised dual discriminative metric network for change detection in high-resolution remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 5513905.

[28] X. Hou, Y. Bai, Y. Li, C. Shang, and Q. Shen, "High-resolution triplet network with dynamic multiscale feature for change detection on satellite images," *ISPRS J. Photogrammetry Remote Sens.*, vol. 177, pp. 103–115, 2021.

[29] Z. Lv, H. Huang, L. Gao, J. A. Benediktsson, M. Zhao, and C. Shi, "Simple multiscale UNet for change detection with heterogeneous remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 2504905.

[30] Z. Lv, P. Zhong, W. Wang, Z. You, and N. Falco, "Multi-scale attention network guided with change gradient image for land cover change detection using remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 20, 2023, Art. no. 2501805.

[31] H. Li, X. Liu, H. Li, Z. Dong, and X. Xiao, "MDFENet: A multiscale difference feature enhancement network for remote sensing change detection," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 3104–3115, 2023.

[32] J. Pan et al., "MapsNet: Multi-level feature constraint and fusion network for change detection," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 108, 2022, Art. no. 102676.

[33] W. Gao, Y. Sun, X. Han, Y. Zhang, L. Zhang, and Y. Hu, "AMIO-Net: An attention-based multiscale input–output network for building change detection in high-resolution remote sensing images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 2079–2093, 2023.

[34] W. Zhang, Q. Zhang, H. Ning, and X. Lu, "Cascaded attention-induced difference representation learning for multispectral change detection," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 121, 2023, Art. no. 103366.

[35] 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.

[36] H. Chen, C. Wu, B. Du, L. Zhang, and L. Wang, "Change detection in multisource VHR images via deep siamese convolutional multiple-layers recurrent neural network," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, pp. 2848–2864, Apr. 2020.

[37] F. Jiang, M. Gong, T. Zhan, and X. Fan, "A semisupervised GAN-based multiple change detection framework in multi-spectral images," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 7, pp. 1223–1227, Jul. 2020.

[38] C. Wu, B. Du, and L. Zhang, "Fully convolutional change detection framework with generative adversarial network for unsupervised, weakly supervised and regional supervised change detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 8, pp. 9774–9788, Aug. 2023.

[39] P. Zhu, H. Xu, and X. Luo, "MDAFormer: Multi-level difference aggregation transformer for change detection of VHR optical imagery," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 118, 2023, Art. no. 103256.

[40] X. Zhang, S. Cheng, L. Wang, and H. Li, "Asymmetric cross-attention hierarchical network based on CNN and transformer for bitemporal remote sensing images change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 2000415.

[41] W. G. C. Bandara, N. G. Nair, and V. M. Patel, "DDPM-CD: Remote sensing change detection using denoising diffusion probabilistic models," 2022, *arXiv:2206.11892*.

[42] S. Liang, Z. Hua, and J. Li, "GCN-based multi-scale dual fusion for remote sensing building change detection," *Int. J. Remote Sens.*, vol. 44, pp. 953–980, 2023.
[43] J. Qu, Y. Xu, W. Dong, Y. Li, and Q. Du, "Dual-branch difference amplification graph convolutional network for hyperspectral image change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2021, Art. no. 5519912.
[44] Y. Sun, L. Lei, D. Guan, and G. Kuang, "Iterative robust graph for unsupervised change detection of heterogeneous remote sensing images," *IEEE Trans. Image Process.*, vol. 30, pp. 6277–6291, 2021.
[45] X. Wang, K. Zhao, X. Zhao, and S. Li, "CSDBF: Dual-branch framework based on temporal–spatial joint graph attention with complement strategy for hyperspectral image change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5540118.
[46] L. Jia, T. Zhang, J. Ai, Y. Lu, and J. Fang, "A stacked stereo graph convolutional network for SAR image change detection," in *Proc. CIE Int. Conf. Radar*, 2021, pp. 3558–3565.
[47] H. Su, X. Zhang, Y. Luo, C. Zhang, X. Zhou, and P. M. Atkinson, "Nonlocal feature learning based on a variational graph auto-encoder network for small area change detection using SAR imagery," *ISPRS J. Photogrammetry Remote Sens.*, vol. 193, pp. 137–149, 2022.
[48] 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.
[49] P. Jian, K. Chen, and C. Zhang, "A hypergraph-based context-sensitive representation technique for VHR remote-sensing image change detection," *Int. J. Remote Sens.*, vol. 37, pp. 1814–1825, 2016.
[50] H. Chen and Z. Shi, "A spatial-temporal attention-based method and a new dataset for remote sensing image change detection," *Remote Sens.*, vol. 12, no. 10, 2020, Art. no. 1662.
[51] Q. Shi, M. Liu, S. Li, X. Liu, F. Wang, and L. Zhang, "A deeply supervised attention metric-based network and an open aerial image dataset for remote sensing change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2021, Art. no. 5604816.
[52] Y. Liu, C. Pang, Z. Zhan, X. Zhang, and X. Yang, "Building change detection for remote sensing images using a dual-task constrained deep siamese convolutional network model," *IEEE Geosci. Remote Sens. Lett.*, vol. 18, no. 5, pp. 811–815, May 2021.
[53] S. Fang, K. Li, J. Shao, and Z. Li, "SNUNet-CD: A densely connected siamese network for change detection of VHR images," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2021, Art. no. 8007805.
[54] H. Chen, Z. Qi, and Z. Shi, "Remote sensing image change detection with transformers," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2021, Art. no. 5920416.
[55] Q. Meng et al., "Mutual information-based disentangled neural networks for classifying unseen categories in different domains: Application to fetal ultrasound imaging," *IEEE Trans. Med. Imag.*, vol. 40, no. 2, pp. 722–734, Feb. 2021.

Xue Zhao received the B.S. degree in computer science and technology from Qilu Normal University, Jinan, China, in 2021. She is currently working toward the M.S. degree in computer science and technology with the School of Information Science and Engineering, Shandong Normal University, Jinan.

Her research interests include remote sensing image change detection and deep learning.

Kai Zhang (Member, IEEE) was born in Shanxi, China, in 1992. He received the B.S. degree in electrical engineering and automation from the North University of China, Taiyuan, China, in 2013, and the Ph.D. degree in circuits and systems from Xidian University, Xi'an, China, in 2018.

From 2021 to 2022, he was a Postdoctoral Fellow with the Remote Sensing Laboratory, Department of Information Engineering and Computer Science, University of Trento, Trento, Italy. He is currently an Associate Professor with the School of Information Science and Engineering, Shandong Normal University, Jinan, China. His research interests include multisource remote sensing image fusion, change detection, and deep learning.

Feng Zhang received the B.S. degree in electronic information engineering from Shandong Normal University, Jinan, China, in 2012, the M.S. degree in electronic information engineering from Xidian University, Xi'an, China, in 2016, and the Ph.D. degree in computer science and technology from Shandong Normal University, in 2023.

She is currently a Lecturer with the School of Information Science and Engineering, University of Jinan, Jinan. Her research interests include deep learning and image processing.

Jiande Sun received the B.S. and Ph.D. degrees in communication and information system from Shandong University, Jinan, China, in 2000 and 2005, respectively.

From 2008 to 2009, he was a Visiting Researcher with the Institute of Telecommunications System, Technical University of Berlin, Berlin, Germany. From 2010 to 2012, he was a Postdoctoral Researcher with the Institute of Digital Media, Peking University, Beijing, China, and also with the State Key Laboratory of Digital-Media Technology, Hisense Group, Qingdao, China. From 2014 to 2015, he was a DAAD Visiting Researcher with the Technical University of Berlin, Berlin, Germany, and the University of Konstanz, Konstanz, Germany. From 2015 to 2016, he was a Visiting Researcher with the School of Computer Science, Language Technology Institute, Carnegie Mellon University, Pittsburgh, PA, USA. He is currently a Professor with the School of Information Science and Engineering, Shandong Normal University. He has authored or coauthored more than 60 journal and conference papers. He is the coauthor of two books. His current research interests include multimedia content analysis, video hashing, gaze tracking, image/video watermarking, and 2-D to 3-D conversion.

Wenbo Wan received the Ph.D. degree in information and communication engineering from Shandong University, Jinan, China, in 2015, under the supervision of Prof. Ju Liu.

From June 2019 to October 2019, he was a Visiting Researcher with the Department of Computer Science, City University of Hong Kong, Hong Kong. He is currently an Associate Professor with the School of Information Science and Engineering, Shandong Normal University, Jinan. His research interests include multimedia security, multimedia quality assessment, and image/video watermarking.

Huaxiang Zhang received the Ph.D. degree in computer science and technology from Shanghai Jiao Tong University, Shanghai, China, in 2004.

From 2004 to 2005, he was an Associated Professor with the Department of Computer Science, Shandong Normal University, Jinan, China, where he is currently a Professor with the School of Information Science and Engineering, together with the School of Information Science and Electrical Engineering, Shandong Jiaotong University, Jinan. He has authored more than 200 journal and conference papers and has been granted 31 invention patents, and is supported by the program of Taishan Scholar. His current research interests include machine learning, pattern recognition, evolutionary computation, and multimedia analysis.