Multiscale Change Detection Network Based on Channel Attention and Fully Convolutional BiLSTM for Medium-Resolution Remote Sensing Imagery

Jialu Li, *Graduate Student Member, IEEE*, Meiqi Hu, *Graduate Student Member, IEEE*, and Chen Wu, *Member, IEEE*

***Abstract*—Remote sensing change detection (CD) is used to detect the difference in the state of objects or phenomena by observing it at different times. CD is widely used in disaster monitoring, land-use and land-cover change analysis, urban expansion detection, and other fields. Medium-resolution (MR) remote sensing imagery can be used for global and regional CD due to the real-time acquisition, extensive coverage, and historical data advantages. Therefore, medium-resolution remote sensing imagery change detection (MRCD) is a very important topic. Compared with very high resolution (VHR) imagery, MR imagery has less texture and edge information. Besides, the object has a large-scale size in VHR imagery scene while the same object will only have a small-scale size in MR imagery scene. To solve the challenge of MRCD, we propose a joint spatial–spectral–temporal network for MRCD, named Multiscale Convolution Channel Attention coupling full convolutional BiLSTM Network (MC^{2}ABNet). The MC^{2}ABNet consists of multiscale convolutional channel attention (MC^{2}A) module and fully Convolutional Bidirectional Long Short-Term Memory (ConvBiLSTM) network. In the encoder, MC^{2}A module is used to extract multiscale spatial features from multitemporal imagery at each encoding level by sharing structure, parameters, and weights. In each MC^{2}A module, the multiscale convolution extracts multiscale spatial features with different receptive fields, and then the channel attention is used to ease the information redundancy during downsampling. The ConvBiLSTM is applied to calculate the time difference features in both forward and backward directions and utilizes spatial information synergistically to smooth change noise for obtaining complete time difference features. The extensive experiments have been conducted on ONERA satellite change detection and SpaceNet7 datasets. Compared with other state-of-the-art methods, our network achieves the highest accuracy on both datasets.**

***Index Terms*—Change detection (CD), medium-resolution (MR) imagery, medium-resolution remote sensing imagery change detection (MRCD), multiscale convolutional channel attentional (MC^{2}A) module.**

Manuscript received 13 July 2023; revised 13 September 2023; accepted 27 September 2023. Date of publication 13 October 2023; date of current version 25 October 2023. This work was supported in part by the National Key Research and Development Program of China under Grant 2022YFB3903300 and in part by the National Natural Science Foundation of China under Grant T2122014 and Grant 61971317. *(Corresponding author: Chen Wu.)*

The authors are with the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan 430079, China. (e-mail: lijialu@whu.edu.cn; meiqi.hu@whu.edu.cn; chen.wu@whu.edu.cn).

Digital Object Identifier 10.1109/JSTARS.2023.3323372

I. Introduction

REMOTE sensing imagery can provide spatial and spectral information about objects on the Earth's surface; thus, the shape, size, and color information of objects can be extracted to study surface change [1]. With the rapid development of Earth observation technology, more and more remote sensing satellites are available to provide more multitemporal remote sensing imagery for change detection (CD) tasks [2], [3], such as GaoFen [4], Sentinel-2 [5], and Landsat [6]. Singh [7] defined remote sensing CD as "the process of identifying differences between imagery at different times." CD is widely used in disaster monitoring [8], land-use and land-cover (LULC) change analysis [9], and urban expansion detection [10], [11]. Medium-resolution (MR) remote sensing imagery can provide data support for the study of large-area CD task [12] due to the advantage of real-time acquisition, large coverage area, and a large amount of historical data. Therefore, Medium-Resolution remote sensing imagery Change Detection (MRCD) is a very important topic.

The most common methods for MRCD are the traditional CD methods. The traditional CD methods have been carried out for several decades and have been well developed in these decades. The traditional CD methods can be divided into three kinds [13]:

1) based on imagery arithmetical methods;
2) based on imagery transformation methods;
3) postclassification methods.

Imagery arithmetical-based methods directly compare the spectral values of multitemporal imagery and obtain the change map according to the suitable threshold value. Typical imagery arithmetical-based methods, such as imagery difference [14], imagery rationing [15], imagery regression [15], and change detection analysis [16], are used to distinguish the changed pixels from the unchanged pixels. Imagery transformation-based methods transform the spectral combinations of the imagery into a specific space to get more remarkable feature map, and then the feature map is segmented to obtain the change map according to the threshold value. The most important imagery transformation-based methods are principal component analysis (PCA) [17], multivariate alteration detection [18], and slow feature analysis [19]. For example, PCA is used to extract the data dimension that accounts for the interested change types

[20], [21]. The postclassification method is the most obvious quantitative CD method and the most commonly used CD method because it can provide detailed "from—to" change information. The method corrects and classifies imagery that come from the same space with different times, and then compare the classification maps to get the change matrix. Many scholars use postclassification methods to solve practical problems, such as LULC change analysis [22] and urban sprawl measuring [23], [24]. However, these traditional methods are difficult to extract effective high-level feature representation, resulting in lower accuracy.

In recent years, the rapid rise of big data and the improvement in computing power have promoted the development of deep learning. Deep learning has made a remarkable performance in the remote sensing imagery interpretation, such as classification, object detection, and CD [25], [26], [27]. Convolutional neural networks (CNNs) [28] have strong feature extraction ability and can extract high-level spatial–spectral features of objects. Thus, a large number of CD methods based on CNN have been proposed [29], [30], [31]. For example, Daudt et al. [32] proposed three models, FC-EF uses UNet structure for CD task, fully convolutional-Siamese-concatenation (FC-Siam-Conc) combines Siamese network with UNet structure to extract spatial features of bitemporal imagery, and fully convolutional-Siamese-difference (FC-Siam-Diff) uses the difference features between dual encoding streams for decoding. Li et al. [33] stacked two imagery as one imagery and then input it into a multiscale full convolutional network for extracting rich spatial features. Shi et al. [34] proposed a deeply supervised attention metric-based network to reduce the pseudochanges and noise caused by external factors. The above CD methods based on deep learning are designed for the VHR imagery and focus on extracting effective VHR imagery spatial features for CD tasks. However, the spatial feature of VHR imagery is different from MR imagery. Thus, the method designed for VHR imagery get terrible performance when applied to the MR imagery.

Taking building as an example, which serves as an important indicator of urban expansion and researched by many scholars [35], [36], [37], in the VHR imagery, buildings have rich texture and regular edge information [38], which is not available in the MR imagery. In addition, in the VHR imagery, building scenes with the large-scale size has a lot of detailed spatial information [39]. However, in the MR imagery, building scenes with the small-scale size has limited spatial information. More important, the MR imagery focuses on the change of large region [40]. If VHR imagery CD methods are directly applied to the MRCD, the insufficient spatial feature is easy to cause the loss of small-scale object and incomplete change boundary of large-scale object. Besides, only using spatial information, such as shape and texture, is not enough, it is important to introduce continuous time difference features for MRCD.

In order to tackle the problems mentioned above, this article proposes a Multiscale Convolution Channel Attention coupling full convolutional BiLSTM Network (MC^{2}ABNet) that joints spatial–spectral–temporal features for MRCD. The innovative multiscale convolutional channel attention (MC^{2}A) module is designed to extract the multiscale spatial feature of MR imagery. The MC^{2}A module uses the Inception V2 [41] to extract local detailed features of small-scale objects in the shallow encoding level and extract global features of large-scale object in the deep encoding level. The Inception V2 module can make the network pay more attention for large region change without losing small-scale object change in the MR imagery. In the process of downsampling, the channel number of spatial features is increasing and will lead to information redundancy. In order to reduce information redundancy, the MC^{2}A module uses channel attention to help MC^{2}ABNet focus on channels that contain important information. For the purpose of extracting complete time difference features for the MRCD task, this article employs the fully Convolutional Bidirectional Long Short-Term Memory (ConvBiLSTM) at the top of each encoding level to extract multiscale difference features. ConvBiLSTM is enable to extract difference features in both forward and backward directions by sharing the structure, parameters, and weights, and combines spatial information to smooth change noise for getting complete difference features. Besides, ConvBiLSTM is able to reduce the multiplication of the high-dimensional matrix while extracting difference features. All in all, the MC^{2}ABNet uses the Siamese network to encode the imagery of different times. In each encoding level, the MC^{2}A module is used for extracting effective multiscale spatial features, then these multitemporal spatial features are inputted into the ConvBiLSTM to calculate time difference features. Finally, these time difference features are restored to the size of the raw imagery by the decoder and then obtain the change probability map through threshold segmentation.

The main contributions of this article are conducted as follows.

1) This article presents a joint spatial–spectral–temporal network named MC^{2}ABNet for MRCD, which can extract abundant multiscale spatial features and complete time difference features of MR imagery; the MC^{2}ABNet can also be used to detect the changed area from time-series remote sensing imagery.
2) The novel MC^{2}A module is designed to extract multiscale spatial features of MR imagery by the multiscale convolution and the channel attention. The MC^{2}A can extract local detailed features of small-scale objects in the shallow encoding level and global features of large-scale objects in the deep encoding level.
3) The ConvBiLSTM network is used at the top of each encoding level to calculate the time difference features between multitemporal imagery. It can extract difference features in both forward and backward directions, and utilize spatial information synergistically to smooth change noise for obtaining complete time difference features.
4) To demonstrate the validity of the proposed MC^{2}ABNet, we have conducted extensive experiments on the MR dataset ONERA satellite change detection (OSCD) and time-series dataset SpaceNet7; MC^{2}ABNet achieves good performance on both datasets and outperforms the state-of-the-art algorithms, demonstrating the effectiveness and generalization of the proposed MC^{2}ABNet.

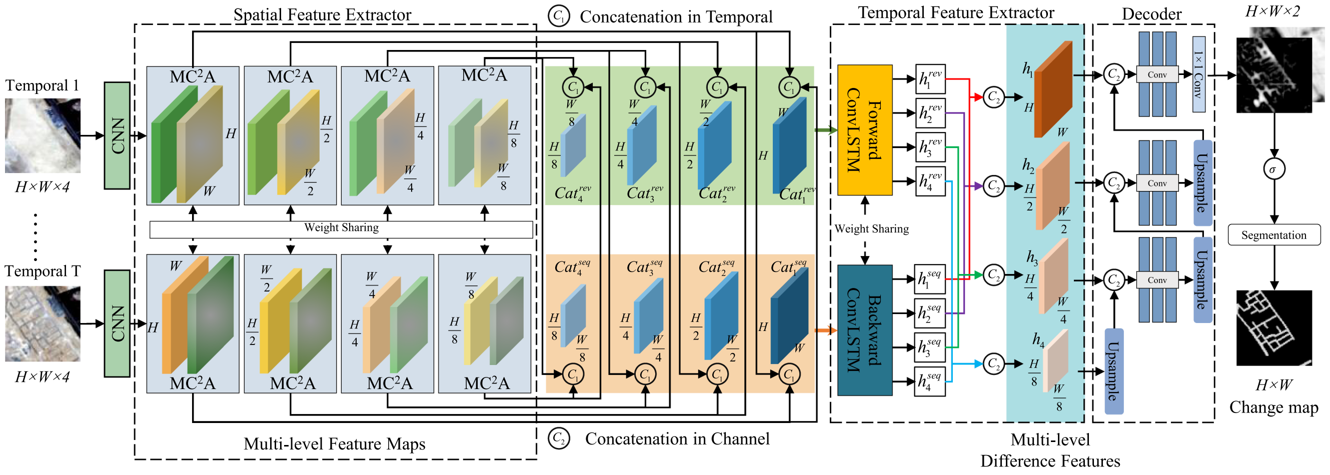

Fig. 1. Structure of the proposed MC^{2}ABNet. The spatial feature extractor—MC^{2}A module—is used to extract multiscale spatial features. The ConvBiLSTM uses the forward ConvLSTM and backward ConvLSTM to extract multilevel temporal features. These temporal features with different sizes are restored to the size of raw imagery in the decoder, finally, using the convolution layer with the kernel size of 1 × 1 to generate the final change map.

The rest of this article is organized as follows. Section II describes the proposed MC^{2}ABNet model in detail. Section III presents the implementation of the algorithms, evaluation metrics, datasets, and their experiments of the proposed network. Finally, Section IV concludes this article.

II. Methodology

A. Overview

This article proposes a joint spatial–spectral–temporal network for MRCD, named MC^{2}ABNet. The overall framework of MC^{2}ABNet is shown in Fig. 1. The network takes some multitemporal imagery separately input into T parallel encoder streams to extract spatial features of T temporal imagery in multiple encoding levels. In each encoding level of the encoding stream, the spatial extraction module MC^{2}A is used to extract accurately the multiscale spatial features of MR imagery. The multiscale convolution of MC^{2}A module can extract detailly local features through convolution with a small kernel size and extract integrally global features through convolution with a large kernel size. The channel attention of MC^{2}A module can reconstruct the optimal representation map according to the contribution of each channel and alleviate the information redundancy in the downsampling. Then, multitemporal spatial features are concatenated in both forward and backward directions in temporal dimensionality. The concatenated multitemporal spatial features are transmitted into ConvBiLSTM to obtain different features between them. The ConvBiLSTM through forward ConvLSTM and backward ConvLSTM shares structures, parameters, and weights to extract temporal difference features. The ConvBiLSTM extracts temporal difference features in both forward and backward directions, and utilizes spatial information synergistically when extracting difference features. Multilevel difference features extracted by ConvBiLSTM are restored to the size of raw imagery through the decoder. Finally, the feature map with the size of $H \times W \times C$ is mapped to the size of $H \times W \times 2$ through the convolution layer with a kernel size of 1×1 and generates the final change map through threshold segmentation.

B. Spatial Feature Extractor MC^{2}A

The spatial features extractor—MC^{2}A module—as shown in Fig. 2, is elaborated to focus on the spatial and spectral features. The InceptionV2 is used to extract multiscale spatial features with different kernel sizes. Then, the channel attention is used to explore useful and informative channels. The MC^{2}A module combines the Inception V2 with channel attention to extract the accurate spatial features of MR imagery for the CD task.

A pixel in the MR imagery fuses the information of multiple pixels in the VHR imagery. If the spatial features are extracted by the method that is designed for the VHR imagery, there will be some problems, such as missing small target detection and inaccurate change boundary. The Inception V2 module can extract the detailed local features and integrally global features through convolution layers with different kernel sizes. Thus, the Inception V2 module can extract rich multiscale spatial features for the MRCD task.

1) *Inception V2 module:* The inception V2 module uses four parallel convolution cubes with different kernel sizes to extract multiscale spatial features. In the first branch, the kernel size of convolution layer is 1×1 and padding is 0. It is mainly used to extract spatial features in the small receptive fields so that the features of one pixel can be extracted as far as possible without considering the influence of surrounding pixels. In the second branch, the kernel size of convolution layer is 3×3 and padding is 1, and the convolution layer in the second branch is used to reduce the number of parameters and calculation. In the third branch, the receptive field of convolution layer is 5×5, but the convolution layer with a kernel size of 5×5 has many parameters. In order to reduce the number of parameters, inception V2 module uses two convolution layers with a kernel size of 3×3 to replace

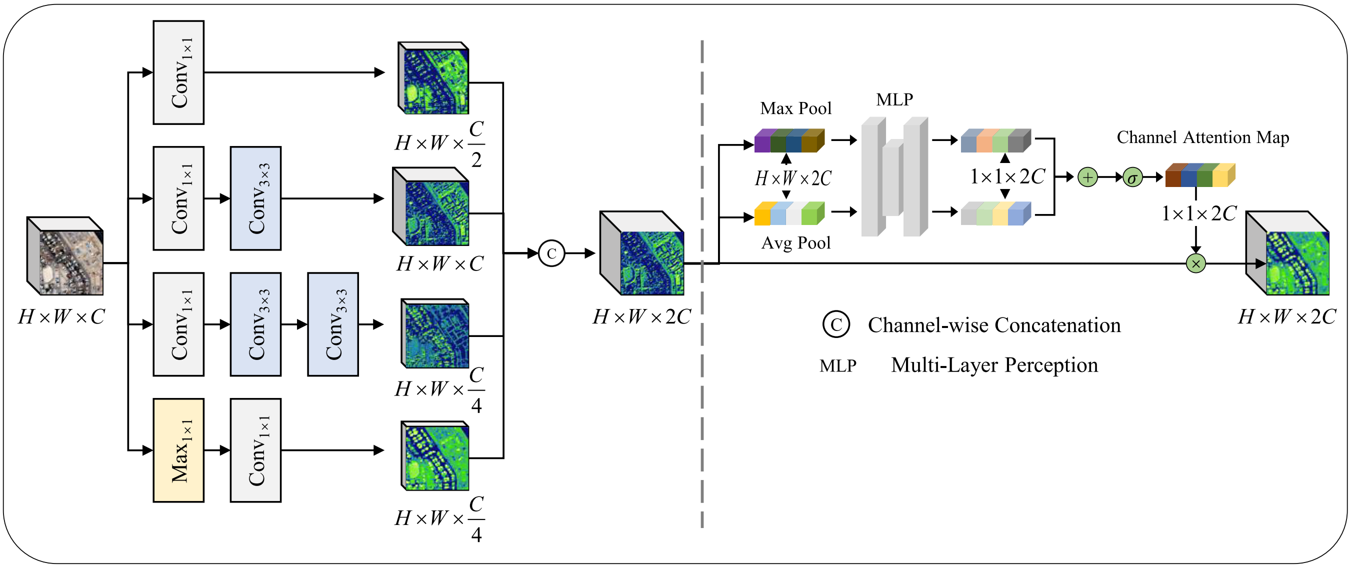

Fig. 2. Structure of the MC^{2}A module, the left half of the MC^{2}A is the inception V2 module, while the right half is the channel attention.

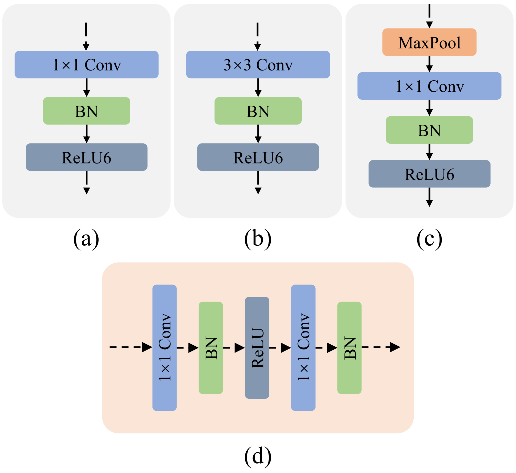

Fig. 3. Structure of the convolution cube in the Inception V2 module. (a) is the convolution cube with the kernel size of convolution layer of 1 × 1. (b) is the convolution cube with the kernel size of 3 × 3. (c) is the structure of the last branch in the Inception V2 module, which is composed of the MaxPool layer and the convolution cube, the kernel size of the MaxPool layer is 3 × 3, the padding is 1, and the stride is 1. (d) is the structure of the MLP.

one convolution layer with a kernel size of 5 × 5. In the last branch, there is a MaxPool layer with a kernel size of 3 × 3. It is mainly used to extract spatial features with strong characteristics. It should be mentioned that BatchNorm layer and ReLU activation function are added behind each convolution layer to form the convolution cube, as shown in Fig. 3(a) and (b). The structure of last branch with MaxPool layer is shown in Fig. 3(c). Finally, the multiscale spatial extracted by four branches is concatenated in the channelwise and then as the input of next step. The concatenation of multiscale spatial features not only can increase the width of the network but also can improve the adaptability of the network to different scales spatial features.

In the process of downsampling, with the increase of encoding level, the number of feature channels is also increasing. In the deep encoding level, the number of concatenated multiscale feature channels can reach up to hundreds. Since different feature channels have different contributions in the CD task, it is necessary to assign different weights to each channel by increasing the weight of important channels and suppressing the weight of unimportant channels to reduce the information redundancy of spatial features. Specifically, the channel attention automatically learns the weight of each channel and reconstructs the feature map.

2) *Channel Attention:* The structure of the channel attention that is used in the MC^{2}A module is shown in Fig. 2. Specifically, the multiscale feature with the size of $C \times H \times W$ is input into AvgPool and MaxPoolto generate two aggregated vectors with a size of $C \times 1 \times 1$. Then, the weight-sharing multilayer perception (MLP) with the channel reduction ratio *r* is used to assign weights to each channel. The structure of MLP is shown in Fig. 3(d). MLP consists of two convolution layers with a kernel size of 1×1, the first layer reduces the number of input feature channels to 1/*r* of number of original feature channels, and the second layer restores the number of feature channels to the number of original feature channels; the BatchNorm layer is placed behind the convolution layer to prevent overfitting. The first BatchNorm layer is followed by the ReLU activation function. After the second BatchNorm layer, the weights of channels obtained by two branches are added and then passed into the sigmoid activation function to produce final weights of channels. Finally, the weights are multiplied by the original features to get new feature maps. The calculation of the new feature map can be expressed by the following formula:

$$C\left(x\right) = \sigma\left(\mathrm{MLP}\left(\mathrm{AvgPool}\left(x\right)\right) + \mathrm{MLP}\left(\mathrm{MaxPool}\left(x\right)\right)\right) \tag{1}$$

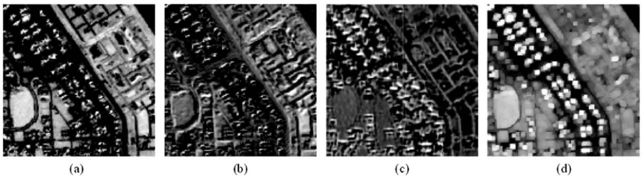

Fig. 4. Multiscale spatial features extracted by the MC^{2}A module. (a) is obtained by convolution layer with a receptive field of 1×1. (b) is obtained by convolution layer with receptive field of 3×3. (c) is obtained by convolution layer with a receptive of 5×5. (d) is obtained by the MaxPool layer with a kernel size of 3×3.

where $C(x)$ is the features reconstructed through the channel attention, $x$is the multiscale feature of size $C \times H \times W$, AvgPool and MaxPool represent the average pooling layer and the global max pooling layer, MLP is the MLP module that consists of the convolution layer, and σ is the sigmoid activation function.

Compared with the method that only uses the convolution layer with a kernel size of 3×3 to extract spatial features, this article designs a spatial feature extractor for MR imagery. Inception V2 module in MC^{2}A is used to extract rich multiscale spatial features by using four parallel branches composed of convolution layer with different kernel sizes, then uses channel attention to reconstruct the optimal feature representation according to the contribution of each channel. The multiscale spatial features extracted by the MC^{2}A module are shown in Fig. 4.

C. ConvBiLSTM

This article uses the ConvBiLSTM at the top of each encoding level to extract time difference features between multitemporal imagery. The ConvBiLSTM is composed of forward ConvLSTM and backward ConvLSTM that share the same structure, parameters, and weights. Compared with RNN, the ConvBiLSTM, such as LSTM, can solve the problem of gradient vanishing and gradient explosion. Compared with LSTM, the ConvBiLSTM can extract temporal features and utilize spatial information synergistically; thus, the ConvBiLSTM not only detects spectral difference between multitemporal imagery but also detects spatial difference between them. The ConvBiLSTM can also reduce the multiplication of high-dimensional matrices through the weights and bias matrix, which replace the convolution layer. Compared with ConvLSTM, the ConvBiLSTM can extract difference features in both forward and backward directions to obtain more accurate and reliable difference features.

The ConvBiLSTM uses forget gate, input gate, cell state, and output gate to forget unimportant information at long distance and emphasizes important information at short distance. The weights matrix and bias matrix in four gates are replaced by the convolution layer. Specifically, four gates can be defined as follows.

1) *Forget gate f_t:* It is used to control information that should be forgotten in the hidden state and cell state of the previous step

$$f_t = \sigma \left(W_f \cdot [C_{t-1}, h_{t-1}, x_t] + b_f \right). \tag{2}$$

2) *Input gate i_t:* It is used to control new information that should be added to the network

$$i_t = \sigma \left(W_i \cdot [C_{t-1}, h_{t-1}, x_t] + b_i \right). \tag{3}$$

3) *Cell state C_t:* It is used to control the information retained in the current cell state that comes from the previous cell state. Specifically, the network creates the new cell state by forgetting irrelevant features and keeping valuable features obtained from the previous cell state

$$c_t = \tanh \left(W_c \cdot [h_{t-1}, x_t] + b_c \right) \tag{4}$$

$$C_t = f_t \odot C_{t-1} + i_t \odot C_t. \tag{5}$$

4) *Output gate o_t:* It is used to control the features that will be output from the ConvBiLSTM unit

$$o_t = \sigma \left(W_o \cdot [C_t, h_{t-1}, x_t] + b_o \right) \tag{6}$$

$$h_t = o_t \ \otimes \tanh \left(C_t \right). \tag{7}$$

In the above formulae, $W_f, W_i, W_c,$ and W_o are the weight matrices; and $b_f, b_i, b_c,$ and b_o are the bias matrices of four gates $f_t, i_t, C_t,$ and o_t, which can be calculated by convolution layer with a kernel size of 3×3. The weight matrices $W_f, W_i, W_c,$ and W_o and bias matrices $b_f, b_i, b_c,$ and b_o are shared at all time steps. The forget gate f_t, the input gate i_t, and the output gate o_t can be obtained by sigmoid activation function σ. The cell state C_t can be obtained by hyperbolic tangent activation function tanh. The above two activation functions can be defined as follows:

$$\sigma \left(x \right) = 1/ \left(1 + e^{-x} \right) \tag{8}$$

$$\tanh \left(x \right) \ = \frac{\sinh \left(x \right)}{\cosh \left(x \right)} \ = \frac{e^x - e^{-x}}{e^x + e^{-x}}. \tag{9}$$

Because the forward ConvLSTM and backward ConvLSTM share the structure, parameters, and weights, this paragraph will mainly introduce the forward ConvLSTM (as shown in the first row of Fig. 5). At first, the hidden state obtained at time $t-1$ is concatenated with the imagery of time t as the ConvLSTM input of time t. Second, the ConvLSTM input of time t is inputted into four parallel convolutions with a kernel size of 3×3 and their activation function to obtain forget gate, input gate, and cell state output gate. Third, cell state of time $t-1$ is multiplied with forget gate, and input gate is multiplied with cell gate, adding the above two results to calculate the cell state of time t. Cell state is the orange rectangle in Fig. 5. Fourth, the cell state is obtained through the tanh activation function and then multiplied by the output to obtain the hidden state of time t. Hidden state is the red rectangle in Fig. 5. The key of ConvLSTM is cell state and hidden state that can transmit information between multitemporal imagery. Four gates in ConvLSTM are calculated by convolution layer; thus, ConvLSTM can reduce the multiplication of high-dimensional matrix and combine difference features with spatial features.

The structure of ConvBiLSTM is shown in Fig. 5. The first row is the forward ConvLSTM, where sequence multitemporal imagery $I \ = [I_1, I_2, \ldots, I_T]$ is input into the forward ConvBiLSTM to obtain the final forward hidden state h_T^{seq} by formula (7).

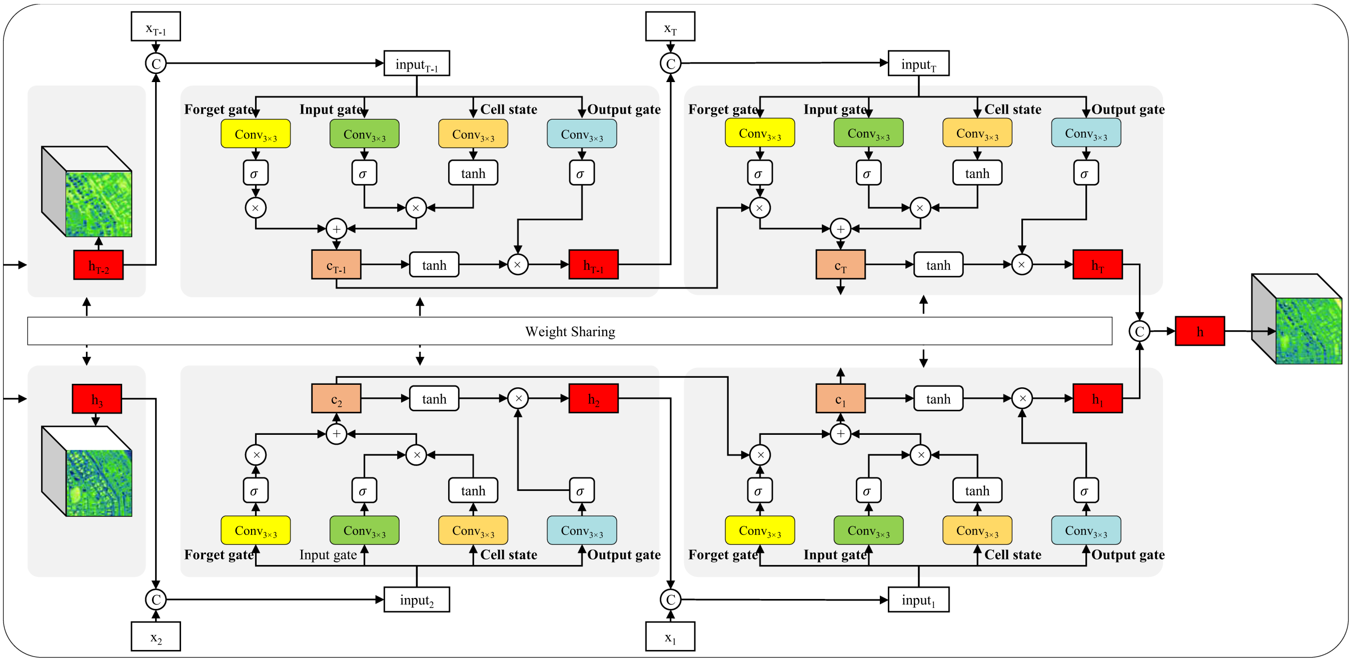

Fig. 5. Structure of the ConvBiLSTM. The first row is the forward ConvLSTM, and the second row is the backward ConvLSTM.

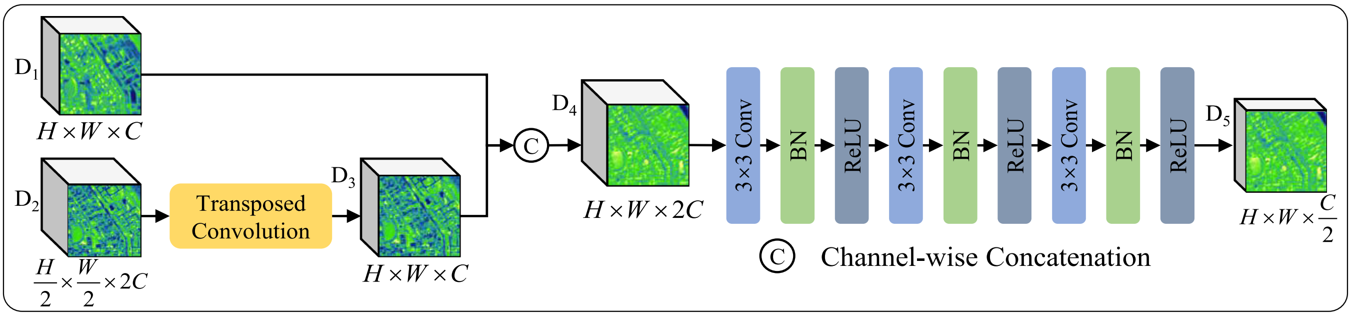

Fig. 6. Structure of the decoder.

The second row is the backward ConvLSTM, where reversed multitemporal imagery $I^{\text{rev}} = [I_T, I_{T-1}, \ldots, I_1]$ is input into the backward ConvLSTM to obtained the final backward hidden state I_0^{rev} by formula (7) as well. Then, the hidden states of forward ConvLSTM and backward ConvLSTM are concatenated in channelwise to obtain the final difference features between multitemporal imagery.

D. Decoder

Decoder is used to restore multilevel difference features to the size of the raw imagery. As shown in the Fig. 6 at first, difference features with the size of $\frac{H}{2} \times \frac{W}{2} \times C$ are inputted into the transposed convolution to obtain new difference features with the size of $H \times W \times C$. Second, new difference features are concatenated with difference features with size of $H \times W \times C$ in channelwise to obtain concatenated difference features. Third, concatenated difference features are input into three convolution layers to make the difference features map complete and smooth. Repeat the above steps until difference features are restored to the size of raw imagery.

E. Loss Function

The cross-entropy loss function is used to measure the difference between the distribution learned by the network and the ground truth. As a commonly used loss function in classification task, the cross-entropy loss function can also be used for distinguishing the change pixels from the unchanged pixels. However, the CD task has the problem of sample imbalance. Generally, unchanged pixels account for the majority of the samples, while changed pixels only account for a small part. The network tends to predict the pixel to the unchanged class so that one can obtain the small loss, but it will lead to poor performance. Thus, this article uses the weighted cross-entropy loss to optimize the model

$$L = -\sum_{n=0}^{N} R \cdot y_{s,n} \cdot \log\left(p_{s,n}\right) \tag{10}$$

where L represents the calculated weighted cross-entropy loss. $N$is the number of change classes, including changed and unchanged. $y_{s,n}$ is a binary indicator of whether ground truth n is the correct answer to the predicted s, and $p_{s,n}$ is the probability of the predicted s belonging to the ground truth n. R is the weight of each class sample. Specifically, the weight of change samples can be calculated by $R = 1 - \frac{c}{a}$, where c is the number of changed samples and a is the number of all samples.

III. Experimental Results and Analysis

To test the effectiveness of the proposed method on MRCD, extensive medium and time-series experiments have been conducted. In this section, the description of datasets is first presented. Next, the details of the experimental setting are given. Then, the analyses of the experimental result tested on medium and time-series imagery are exhibited, respectively.

A. Datasets and Implementation Details

In this article, we use two public CD datasets, the MR imagery dataset called OSCD dataset and the time-series imagery dataset called SpaceNet7, to evaluate the proposed method. It should be noticed that, for all datasets, the CD task is performed using two classes: changed and unchanged.

1) *OSCD dataset:* The first dataset is OSCD. OSCD dataset was created using the multispectral imagery taken by Sentinel-2 satellites of places with different levels of urbanization in several different countries that have experienced urban growth and changes. Labeling as changed only urban growth and changes, ignoring natural changes, such as vegetation growth. OSCD dataset depicts changes in 24 different cities of the world where urbanization was evident, such as Chongqing, China, Hongkong, China, and Paris, France. Each imagery pair with original file format tiff has 13 bands with resolutions between 10 and 60 m, with bit depth of 16. Our experimental setup is the same as the official requirements, with 14 imagery pairs for training and 10 for testing. Because the percentage of the change pixels for OSCD dataset only is 2.3%, some data augmentation techniques are used to avoid overfitting during training. The training imagery is augmented by using all possible flips and rotations multiple of 90°. In Fig. 7, this article has illustrated some imagery and ground truth of OSCD dataset.

In the training process of the OSCD dataset, the imagery has been clipped as the patch with the size of 128×128 and the stride is 64. There are about 1500 patches containing both changed and unchanged pixels, about 1200 of which are used for training the network and about 300 are used for validation purposes. For the MC^{2}ABNet, the Adam optimizer was adopted with the initial learning rate of 0.001 on OSCD dataset, a batch size of 8 sample pairs was utilized to accelerate the convergence of the model, the L2 regularization coefficient is 0.001, and the reduction ratio *r* in the channel attention is 8. It should be noted that in the experiment, this article utilized four bands (red band, blue band, green band, and near infrared with a resolution of 10 m) of the Sentinel-2 satellite for comparison and ablation experiments.

2) *SpaceNet7 dataset:* The second dataset is SpaceNet7. SpaceNet7 dataset was recently released for multitemporal building detection in the NeuraIPS 2020 challenges. This dataset contains 100 locations distributed around the globe and comprises over 40 000 km^2, where buildings have been changed, but only 60 data cubes provide building labels. Each data cube contains about 24 imagery at different times acquired in 2018, 2019, and 2020. Each imagery of data cubes has four spectra: red band, blue band, green band, and near-infrared band; the original file format is tiff, the size of each imagery is approximately 1024×1024, the resolution is 4 m, and the bit depth is 32. Our experimental setup is based on 60 data cubes that provide ground truth of which 48 data cubes are used for training and 12 data cubes for verification and testing. Since the SpaceNet7 dataset is a semantical segmentation dataset, there are some data cubes that have very few changes. When randomly dividing the training and testing data, do not divide data cubes that have very little change into testing. Because the label provided by SpaceNet7 dataset is the buildings segmentation for each imagery, the building segmentation of the first imagery and the last imagery is used to obtain the building change map during the period. In Fig. 8, this article has illustrated some imagery and ground truth of the SpaceNet7 dataset.

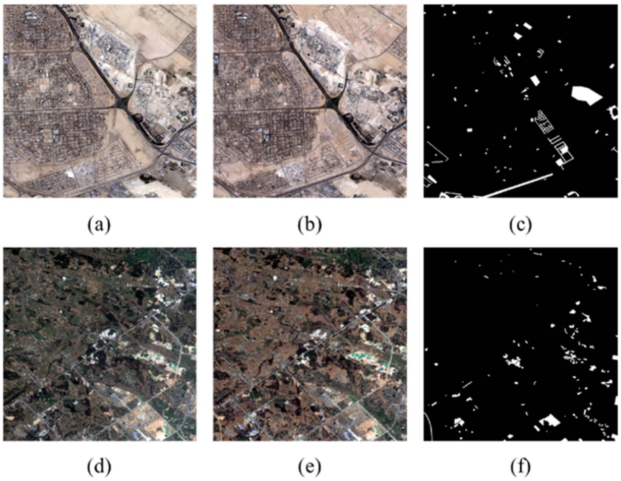

Fig. 7. Imagery of OSCD dataset, the first and second columns are visible light imagery obtained at different times, and the third column is the ground truth of the corresponding imagery pair. (a) 01/2016. (b) 03/2018. (c) Ground truth. (d) 12/2016. (e) 03/2018. (f) Ground truth.

In the training process of the SpaceNet7 dataset, the imagery has been clipped as the patch with a size of 256×256 and the stride of 128. There are about 3000 patches containing both changed and unchanged pixels, about 2400 patches are used for training the model and about 600 patches for verifying and testing the model. For the MC^{2}ABNet, the Adam optimizer was also used during the optimization, the initial learning rate is 0.0003, the batch size is 8 for bitemporal imagery and 2 for time-series imagery, and the L2 regularization is also used for avoiding overfitting. In the experiment, all bands are used for bitemporal imagery and time-series imagery comparison experiments. It should be noticed that all our experiments are conducted on the GeForce RTX 3090Ti GPU to train the model.

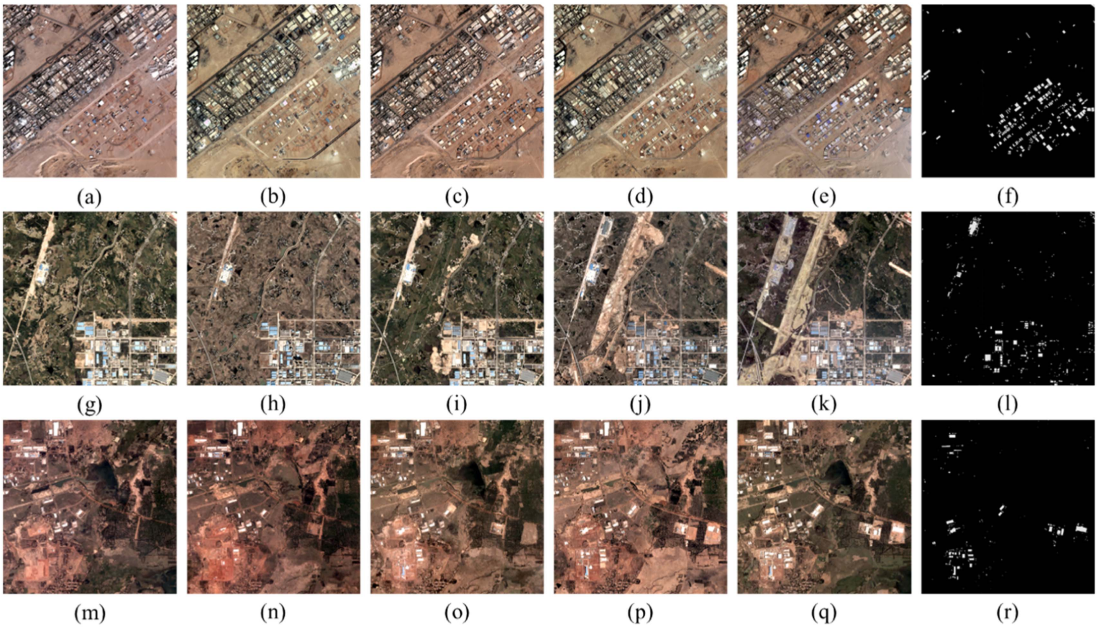

Fig. 8. Time-series imagery of SpaceNet7 dataset. Each imagery cube has many imagery from different times, and this article has shown five imagery of them. From the first column to the fifth column are the time-series imagery, and the last column is the ground truth. (a) 01/2018. (b) 07/2018. (c) 01/2019. (d) 07/2019. (e) 01/2020. (f) Ground truth. (g) 07/2017. (h) 02/2018. (i) 07/2018. (j) 03/2019. (k) 09/2019. (l) Ground truth. (m) 01/2018. (n) 07/2018. (o) 01/2019. (p) 07/2019. (q) 01/2020. (r) Ground truth.

B. Evaluation Metrics

In order to evaluate the performance of the proposed network, this article calculates six evaluation metrics to compare the predicted results obtained by the proposed MC^{2}ABNet and other state-of-the-art CD networks. These evaluation metrics are overall accuracy (OA), precision (Pre), recall (Rec), $F1$-score ($F1$), Kappa coefficient (Kappa), and Mean Intersection over Union (MIoU). In the CD task, Pre is the proportion of the number of correctly predicted positive pixels to the total predicted positive pixels, and higher Pre represents the lower false detection rate. Rec is the proportion of the number of correctly predicted positive pixels to the total labeled positive pixels, and higher Rec means the lower missed detection rate. Because Pre and Rec are contradictory, the harmonic average of two indicators is used for comprehensive evaluation. OA, $F1$, Kappa, and MIoU are comprehensive evaluation indices and they can judge the quality of model, with the higher value, the better the performance of model. Note that Kappa is more important for the assessment of CD task because of the imbalance between changed and unchanged samples.

C. Comparison Method

1) *FC-Siam-Conc [32]:* Fc-Siam-Conc was proposed on the basis of FC-EF and combined the Siamese structure with the UNet network. The Siamese encoding stream with sharing weights is used to extract the features of bitemporal imagery. In the decoding stage, the bitemporal features are concatenated to the decoder as a skip connection to fuse the different scale features.
2) *FC-Siam-diff [32]:* FC-Siam-diff was proposed at the same time as FC-Siam-Conc. It is also a network that combines Siamese structure with the UNet network. The difference is that, in the decoding stage, the difference between bitemporal features is input to the decoder as a skip connection to fuse the features of different scales.
3) *UNet-like architecture (L-UNet) [12]:* L-UNet was proposed for urban CD and combined the fully convolutional network and recurrent networks (LSTM). In the decoding stage, LSTM is used to extract the temporal relationship between multitemporal imagery and input into the network as a skip connection to fuse the features of different scales.
4) *SNUNet-CD [29]:* It is a densely connected Siamese network used to alleviate the loss of deep localization information through the compact information transmission between encoder and decoder, and between decoder and decoder.
5) *Difference enhancement and spatial–spectral nonlocal network (DESSN) [42]:* DESSN designs a difference module that can learn the difference representation between foreground and background, and a spatial–spectral nonlocal module that can strengthen edge integrity and internal tightness of changed objects by learning long-range correlation.

D. Experimental Results and Analysis

In this section, we evaluate the performance of the proposed MC^{2}ABNet and other state-of-the-art CD approaches, as introduced in Section III-C. To benchmark their performance, we report results on OSCD and SpaceNet7 datasets. Because there

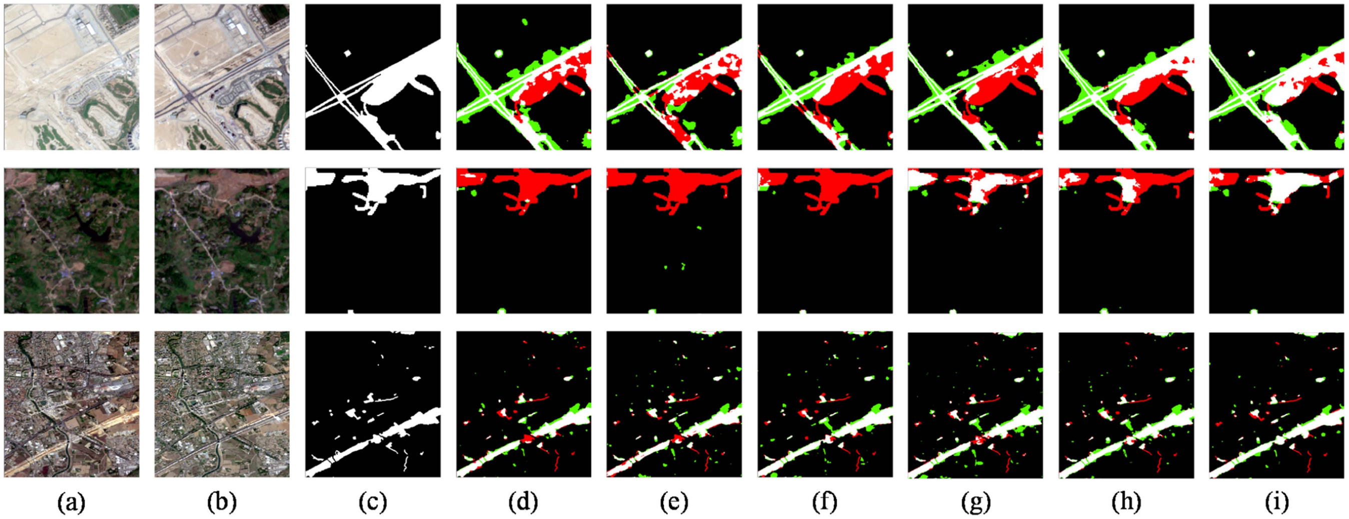

Fig. 9. Bitemporal imagery, ground truth, and binary change maps obtained by the proposed MC^{2}ABNet and comparison methods on OSCD dataset. (a) t_1 imagery. (b) t_2 imagery. (c) Ground truth. (d) FC-Siam-Conc. (e) FC-Siam-diff. (f) L-UNet. (g) SNUNet. (h) DESSN. (i) MC^{2}ABNet.

TABLE I
Accuracy Assessment on CD Results Obtained by Different Methods on OSCD Dataset

Method	Precision	Recall	OA	Kappa	F1	MIoU
FC-Siam-Conc	0.5534	0.5781	0.9541	0.5413	0.5655	0.6734
FC-Siam-diff	0.5600	0.5846	0.9548	0.5482	0.5721	0.6770
L-UNet	<u>0.5877</u>	0.5593	<u>0.9569</u>	0.5504	0.5731	0.6787
SNUNet	0.5499	**0.6289**	0.9542	0.5627	0.5868	0.6839
DESSN	0.5684	<u>0.6136</u>	0.9560	<u>0.5669</u>	<u>0.5901</u>	<u>0.6865</u>
MC^{2}ABNet	**0.6174**	0.6023	**0.9602**	**0.5887**	**0.6097**	**0.6987**

are time-series imagery in each imagery cube of SpaceNet7 dataset, we set up two sets of experiments for SpaceNet7 dataset: bitemporal and time-series experiments.

The bitemporal imagery, the ground truth, and the binary change map obtained by MC^{2}ABNet and other comparison methods on OSCD dataset are shown in Fig. 9. In the first example, we can see that the results of FC-Siam-Conc are unsatisfactory, and there are many changed pixels that are detected as unchanged pixels and many unchanged pixels are falsely detected as changed pixels. FC-Siam-diff reduced false detection by considering the difference between bitemporal imagery. But there are still many changed pixels detected as unchanged pixels. The binary change map obtained by L-UNet is visually compact, and the falsely detected pixels are tightly joined together. But there are still many changed pixels detected as unchanged pixels. Because SNUNet is a densely connected network that takes the shallow features into account, the result obtained by SNUNet is better than L-UNet. Compared with the previous methods, the binary change map obtained by DESSN greatly reduced the number of omitted changed pixels. The proposed MC^{2}ABNet takes the advantage of the multiscale convolution to extract local detail features in shallow encoding levels and global features in deep encoding levels. Then, the ConvBiLSTM is used to capture the time correlation between bitemporal imagery in both forward and backward directions. The result shows that the proposed method achieves the best performance with fewer false and omitted pixels and correctly detected changed areas.

In the second example, the area is changed from grassland to bare land. The FC-Siam-Conc, FC-Siam-diff, and L-UNet have omitted these grass changes, while SNUNet and DESSN obtained better binary change maps with correctly detected some change pixels. The MC^{2}ABNet presented the best performance and correctly detected most changed pixels. In the third example, compared with other methods, our method can balance precision and recall better and obtain the best accuracy. It is worth mentioning that, in the first and third examples, there are some pixels that have been changed from other kinds to roads and all methods can detect this kind of change. It means that MC^{2}ABNet has good performance in road detection.

The accuracy assessments of CD results on the OSCD dataset based on six evaluation criteria, as described in Section III-B, are displayed in Table I. The maximum value is marked in bold, and the second-best value is underlined. The quantitative results are similar to the qualitative results, FC-Siam-Conc and FC-Siam-diff have lower accuracy with OA of 0.9541 and 0.9548, Kappa of 0.5413 and 5482, F1 of 0.5655 and 0.5721, and MIoU of 0.6734 and 0.6770. L-UNet captures the time correction between bitemporal imagery. SNUNet combines the shallow features and deep features through dense connection. DESSN strengthens the edge integrity and the internal tightness of changed objects. So, L-UNet, SNUNet, and DESSN have better accuracy than FC-Siam-Conc and FC-Siam-Diff. MC^{2}ABNet achieves the best result with OA of 0.9602, Kappa of 0.5887, F1 of 0.6097, and MIoU of 0.6987. Compared with FC-Siam-Conc, the proposed MC^{2}ABNet increased OA, Kappa, F1, and MIoU by 0.61%, 4.74%, 4.42%, and 2.53%, respectively. It means that the proposed MC^{2}ABNet can effectively extract the spatial feature by novel MC^{2}A module and capture time correction between bitemporal in both forward and backward directions by ConvBiLSTM.

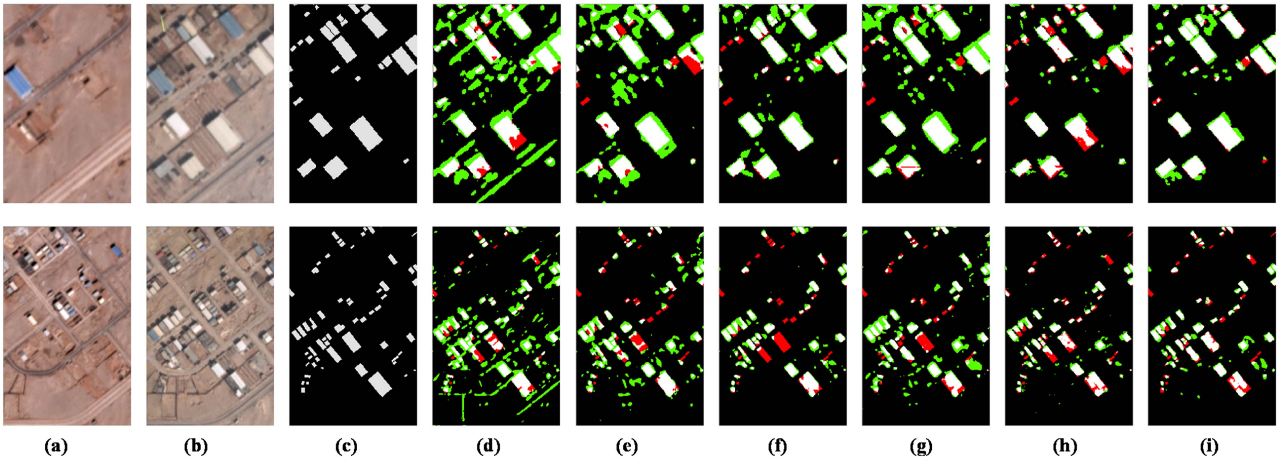

Fig. 10. Bitemporal imagery, ground truth, and binary change maps obtained by the proposed MC^{2}ABNet and comparison methods on SpaceNet7 dataset with bitemporal imagery. (a) t_1 imagery. (b) t_2 imagery. (c) Ground truth. (d) FC-Siam-Conc. (e) FC-Siam-diff. (f) L-UNet. (g) DESSN. (h) SNUNet. (i) MC2ABNet.

Because SpaceNet7 is a dataset of time-series imagery, we design two sets of experiments for the SpaceNet7 dataset. The first set of experiments selects the first and the last imagery of time-series imagery to detect changed pixels, and the second set of experiments selects time-series imagery to verify the improving accuracy ability of time-series imagery. Fig. 10 shows CD results obtained by the proposed MC^{2}ABNet and comparison methods on the SpaceNet7 dataset with bitemporal imagery. Similar to the qualitative results on OSCD dataset, in the first example, the misclassification and noise in the binary change maps obtained by FC-Siam-Conc and FC-Siam-Diff are obvious. There are many changed pixels detected as unchanged pixels and some unchanged pixels falsely detected as changed pixels. But for FC-Siam-Diff, the false detection has been eased. For L-UNet and DESSN, due to the superiority of framework, their results are better than those obtained by FC-Siam-Conc and FC-Siam-Diff. These two results have fewer false detection pixels than the above two results. The results obtained by SNUNet are much clearer visually, and it has fewer false detection pixels than previously mentioned methods but it has many missed detected pixels as much as FC-Siam-Conc and FC-Siam-Diff. Once again, the proposed MC^{2}ABNet generates better qualitative results. Because SpaceNet7 is a dataset about building segmentation and CD, the changed pixels focus on the building object. Compared with other methods, the binary change map obtained by the proposed MC^{2}ABNet has more regular boundary range for change areas and the interior of the changed object is tighter. There are almost no missed pixels and fewer false pixels than the other methods.

The building object in the second example is smaller than in the first example. The results obtained by FC-Siam-Conc are unsatisfactory and there are many falsely detected pixels. Compared with FC-Siam-Conc, the results obtained by FC-Siam-Diff, L-UNet, DESSN, and SNUNet are much clearer. The binary change map obtained by the proposed MC^{2}ABNet once again performs the best, especially on small objects.

TABLE II
ACCURACY ASSESSMENT ON CD RESULTS OBTAINED BY DIFFERENT METHODS ON SPACENET7 DATASET OF BI-TEMPORAL IMAGERY

Method	Precision	Recall	OA	Kappa	F1	MIoU
FC-Siam-conc	0.4777	**0.7084**	0.9814	0.5614	0.5706	0.6902
FC-Siam-diff	0.5459	0.6091	0.9843	0.5678	0.5758	0.6942
L-UNet	0.6113	0.5952	0.9863	0.5961	0.6031	0.7090
DESSN	0.5689	0.6905	0.9855	0.6165	0.6238	0.7193
SNUNet	0.6665	0.6082	0.9878	0.6298	0.6360	0.7270
MC^{2}ABNet	**0.6792**	0.6122	**0.9882**	**0.6380**	**0.6440**	**0.7315**

The accuracy assessments of CD results on SpaceNet7 dataset of bitemporal imagery based on six evaluation criteria are displayed in Table II. Starting from FC-Siam-Conc, it gets the best recall score, but precision is only 0.4777, so accuracy is not very good with OA of 0.9814 and Kappa of 0.5614. FC-Siam-Diff uses differences between imagery to balance precision and recall, so it performs better than FC-Siam-Conc. L-UNet uses the Siamese network to extract the spatial features of bitemporal imagery and capture the temporal correlation between them. Kappa is 0.5961, which increased by 2.83% compared with the FC-Siam-diff. The accuracy of DESSN and SNUNet is next to the proposed MC^{2}ABNet due to accurate and rich features. It should be noticed that the accuracy of results obtained by DESSN is better than SNUNet on OSCD dataset, but the accuracy of results obtained by SNUNet is better than DESSN on SpaceNet7 dataset with bitemporal imagery. Because SpaceNet7 dataset has a large number of small building objects, SNUNet performs better than DESSN by taking account of local details through dense connections. Finally, the proposed MC^{2}ABNet achieves the highest accuracy with OA of 0.9822, Kappa of 0.6380, $F1$ of 0.6440, and MIoU of 0.7315, which proves the effectiveness of the proposed network in CD task once again.

Because SpaceNet7 dataset has time-series imagery in each imagery cube, L-UNet and proposed MC^{2}ABNet can be used for time-series imagery CD task, so we designed a set of experiments to verify the effectiveness of time-series imagery in the CD task by comparing bitemporal and time-series imagery. In the time-series imagery experiments, we evenly selected ten imagery,

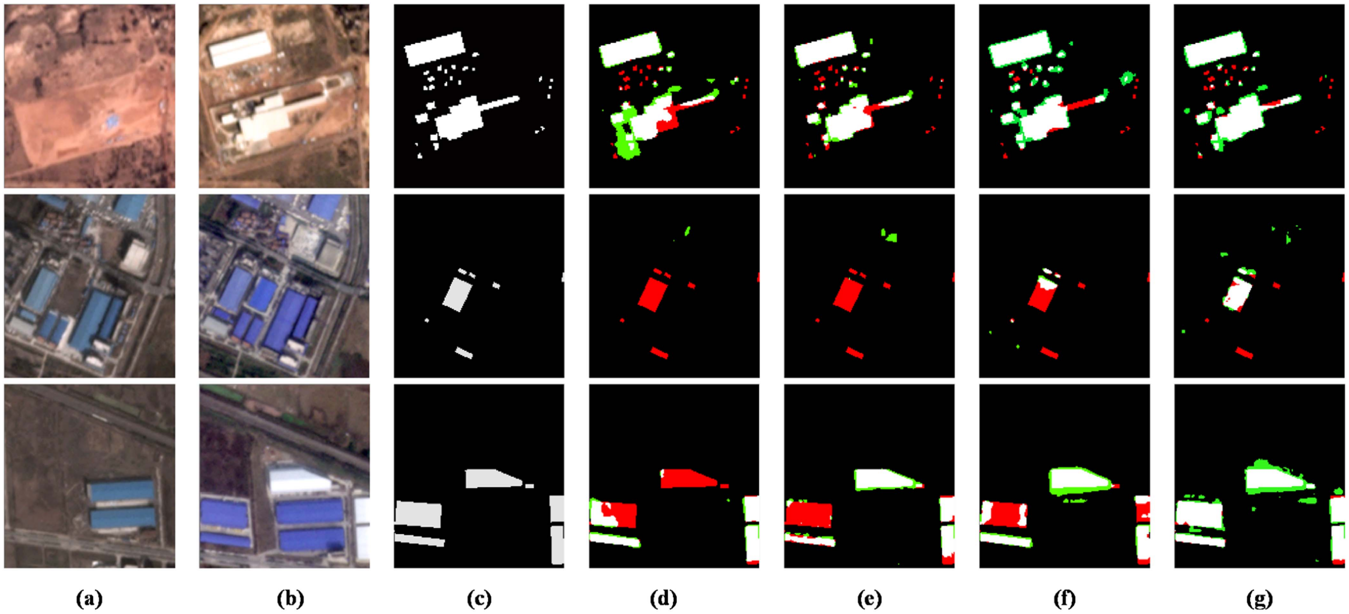

Fig. 11. First and the last imagery of time-series imagery, ground truth, and binary change maps obtained by the proposed multi-MC^{2}ABNet and comparison methods on SpaceNet7 dataset with time-series imagery. (a) t_1 imagery. (b) t_T imagery. (c) Ground truth. (d) L-UNet. (e) MC^{2}ABNet. (f) Multi-L-UNet. (g) Multi-MC^{2}ABNet.

TABLE III
ACCURACY ASSESSMENT ON CD RESULTS OBTAINED BY DIFFERENT METHODS ON SPACENET7 DATASET OF TIME-SERIES IMAGERY

Method	Precision	Recall	OA	Kappa	F1	MIoU
L-UNet	0.6113	0.5952	0.9863	0.5961	0.6031	0.7090
MC^{2}ABNet	<u>0.6792</u>	0.6122	<u>0.9882</u>	<u>0.6380</u>	<u>0.6440</u>	<u>0.7315</u>
multi-L-UNet	0.6112	**0.6403**	0.9865	0.6186	0.6254	0.7207
multi-MC^{2}ABNet	**0.6850**	<u>0.6275</u>	**0.9884**	**0.6491**	**0.6550**	**0.7376**

including the first imagery and the last imagery, to train the model.

Fig. 11 shows CD results obtained by the proposed multi-MC^{2}ABNet and comparison methods on SpaceNet7 dataset with time-series imagery. In the first example, L-UNet has many falsely detected pixels, but it has been eased in the proposed MC^{2}ABNet. For the multi-L-UNet, there are fewer false detection pixels on large objects but more false detection pixels on small objects. Compared with L-UNet, multi-L-UNet has fewer missed pixels. Compared with other methods, multi-MC^{2}ABNet performs the best and detects building boundaries with little noise. A more obvious comparison can be seen in the second and third examples. The proposed multi-MC^{2}ABNet accurately detects changed objects that comparison methods fail to detect, verifying the superiority of time-series imagery in change task.

The accuracy assessments of CD results on SpaceNet7 dataset with time-series imagery based on six evaluation criteria are displayed in Table III. Starting with L-UNet, it has the lowest accuracy with OA of 0.9863, Kappa of 0.5961, $F1$ of 0.6031, and MIoU of 0.7090. By extracting the time correlation between the features of time-series imagery, multi-L-UNet achieves a higher accuracy with OA of 0.9865 and Kappa of 0.6186. The proposed MC^{2}ABNet with the OA of 0.9882 and Kappa of 0.6380 is higher than the corresponding OA and Kappa of multi-L-UNet, which indicates that the improvement of accuracy mainly depends on the structure of the model. Finally, the accuracy of multi-MC^{2}ABNet is the highest with OA of 0.9884, Kappa of 0.6491, $F1$ of 0.6550, and MIoU of 0.7376. Compared with MC^{2}ABNet, they are improved by 0.02%, 1.11%, 1.10%, and 0.61%, respectively, which proved the validity of time-series imagery in CD task. It should be noticed that multi-MC^{2}ABNet takes four times as long as MC^{2}ABNet to train the network.

E. Ablation Experiment

On the basis of metric learning, MC^{2}ABNet integrates the MC^{2}A module and ConvBiLSTM for accurate CD. Therefore, we design ablation experiments on MC^{2}ABNet to verify the validity of the MC^{2}A module and ConvBiLSTM. In the following experiment, the “B” represents the base model, which combines the UNet network with the Siamese network for the CD task, and the structure of “B” is similar to the structure of FC-Siam-Conc. “M” represents the multiscale convolution, “A” represents the channel attention, “C” represents the ConvBiLSTM, and “M+A” represents the MC^{2}A module. “$\surd$” represents adding this module to the base model.

The accuracy assessments of ablation experiments on OSCD dataset based on five evaluation criteria are displayed in Table IV. As can be seen from Table IV, the combination of the MC^{2}A module and ConvBiLSTM can improve the performance of the model. More specifically, the results obtained by the base model have the lowest accuracy with OA of 0.9514, Kappa of 0.5383, and $F1$ of 0.5639, and the quantitative results are similar to those of FC-Siam-Conc in Table I. After adding multiscale convolution, OA, Kappa, and $F1$ increased by 0.38%, 2.32%, and 2.13%, respectively, on OSCD dataset. It shows that multiscale convolution does extract more efficient spatial features

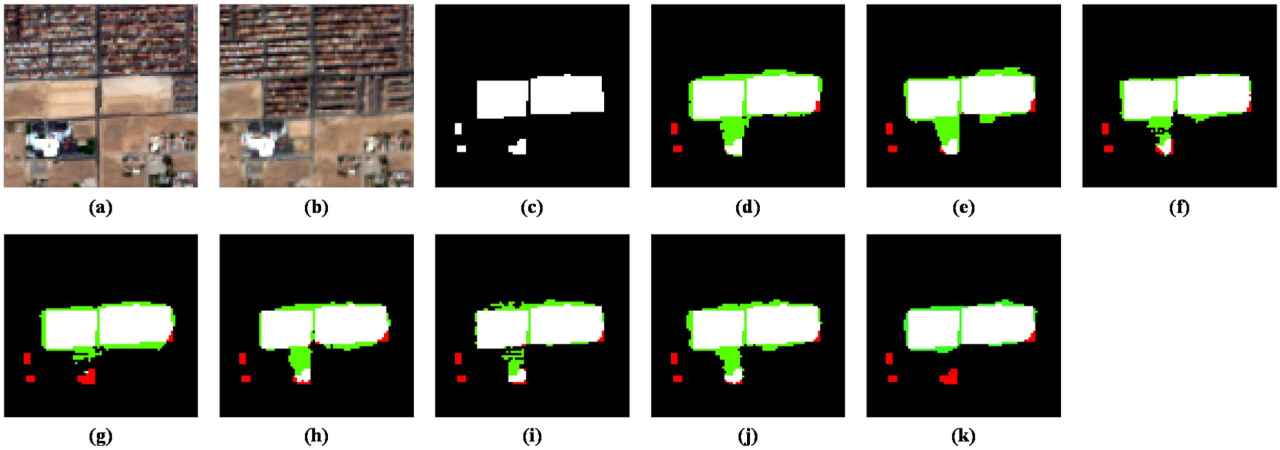

Fig. 12. Bitemporal imagery, ground truth, and binary change maps obtained by the ablation experiments on OSCD dataset. (a) t_1 imagery. (b) t_2 imagery. (c) Ground truth. (d) B. (e) B+M. (f) B+A. (g) B+C. (h) B+M+A. (i) B+M+C. (j) B+A+C. (k) MC^{2}ABNet.

TABLE IV
ACCURACY ASSESSMENT ON CD RESULTS OBTAINED BY ABLATION EXPERIMENT ON OSCD DATASET

Method	M	A	C	Precision	Recall	OA	Kappa	F1
B				0.5262	0.6073	0.9514	0.5383	0.5639
B+M	√			0.5615	0.6109	0.9552	0.5615	0.5852
B+A		√		0.5339	0.6454	0.9525	0.5594	0.5843
B+C			√	0.6015	0.5897	0.9586	0.5737	0.5955
B+M+A	√	√		0.5814	0.6080	0.9571	0.5718	0.5944
B+M+C	√		√	0.5562	**0.6539**	0.9551	0.5775	0.6011
B+A+C		√	√	0.5706	0.6362	0.9565	0.5786	0.6016
MC^{2}ABNet	√	√	√	**0.6174**	0.6023	**0.9602**	**0.5887**	**0.6097**

for CD task. After adding channel attention, OA, Kappa, and $F1$ increased by 0.11%, 2.11%, and 2.04%, respectively. It shows that channel attention does reduce the information redundancy during downsampling. On the basis of the base model, add the MC^{2}A module, composed of multiscale and channel attention. OA, Kappa, and $F1$ increased by 0.57%, 3.35%, and 3.05%, respectively, in the dataset. The results show that the MC^{2}A module can improve the performance of the model. After adding the ConvBiLSTM, OA, Kappa, and $F1$ increased by 0.72%, 3.54%, and 3.16%, respectively. This indicates that ConvBiLSTM does extract richer temporal difference features and made substantial contributions for the CD task. We also designed experiments combined inception V2 with ConvBiLSTM and channel attention with ConvBiLSTM to prove the contribution of each module. Notably, MC^{2}ABNet integrated the MC^{2}A module and ConvBiLSTM obtained the accuracy with OA of 0.9602, Kappa of 0.5887, and $F1$ of 0.6097 on OSCD dataset. Compared with the base model, MC^{2}ABNet increased OA, Kappa, and $F1$ by 0.88%, 5.04%, and 4.58%, respectively. The great improvement of MC^{2}ABNet further demonstrates not only the effectiveness of the MC^{2}A module and ConvBiLSTM but also the gain results of their combination.

Fig. 12 shows CD results obtained by the ablation experiments on OSCD dataset. Compared with the base model, the binary change map obtained by MC^{2}ABNet is much clearer visually and has less false and missed alarms. Therefore, MC^{2}ABNet combined with the MC^{2}A module with ConvBiLSTM can greatly improve the edge integrity and the internal tightness of change areas, which further verifies that MC^{2}A module can extract effective spatial features and ConvBiLSTM can extract richer temporal difference features for CD task. Thus, the MC^{2}ABNet can obtain the accurate change maps for MRCD.

IV. CONCLUSION

In this article, a new joint spatial–spectral–temporal CD network applicable for MR imagery and time-series imagery is proposed, which is called MC^{2}ABNet. MC^{2}ABNet consists of MC^{2}A module and ConvBiLSTM. The MC^{2}A module is used to extract multiscale spatial features of MR imagery. The MC^{2}A module can extract local detailed features of small-scale objects in the shallow encoding level and extract global features of large-scale object in the deep encoding level. The channel attention of MC^{2}A increases the weight of important channels and decreases the weight of unimportant channels for reducing the information redundancy of spatial features during downsampling. The ConvBiLSTM calculates difference in both forward and backward directions for obtaining more accurate temporal difference features and replaces the fully connected layer with the convolution layer to utilize spatial information synergistically, which smooth the noise of final CD map and obtain complete difference features. Qualitative and quantitative results in MR OSCD dataset and time-series SpaceNet7 dataset demonstrate that MC^{2}ABNet outperforms the widely used CD methods. Especially compared with another five prediction methods based on deep learning, the proposed MC^{2}ABNet obtained better overall accuracy, Kappa coefficient, $F1$ score, and MIoU. The results demonstrate the effectiveness of joint spatial–spectral–temporal MC^{2}ABNet in MRCD.

MR imagery has the advantage of real-time acquisition and extensive coverage. Thus, there are many time-series imagery that can be used for CD task. But time-series imagery contains three changes, including intra-annual, interannual, and abrupt changes. In the following study, we will focus on detecting the accurate abrupt change in time-series MR imagery.

REFERENCES

[1] A. Goswami et al., "Change detection in remote sensing image data comparing algebraic and machine learning methods," *Electronics*, vol. 11, no. 3, Jan. 2022, Art. no. 431, doi: 10.3390/electronics11030431.

[2] C. Han, C. Wu, H. Guo, M. Hu, and H. Chen, "HANet: A hierarchical attention network for change detection with bitemporal very-high-resolution remote sensing images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 16, pp. 3867–3878, Apr. 2023, doi: 10.1109/JSTARS.2023.3264802.

[3] M. Hu, C. Wu, B. Du, and L. Zhang, "Binary change guided hyperspectral multiclass change detection," *IEEE Trans. Image Process.*, vol. 32, pp. 791–806, Jan. 2023, doi: 10.1109/TIP.2022.3233187.

[4] J. Luo, Q. Chu, C. Sun, Y. Wang, and D. Sun, "Staple crop mapping with Chinese Gaofen-1 and Gaofen-6 satellite images: A case study in Yanshou County, Heilongjiang Province, China," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2021, pp. 6769–6772, doi: 10.1109/IGARSS47720.2021.9553921.

[5] M. K. Vanderhoof, L. Alexander, J. Christensen, K. Solvik, P. Nieuwlandt, and M. Sagehorn, "High-frequency time series comparison of Sentinel-1 and Sentinel-2 satellites for mapping open and vegetated water across the United States (2017–2021)," *Remote Sens. Environ.*, vol. 288, Apr. 2023, Art. no. 113498, doi: 10.1016/j.rse.2023.113498.

[6] X. Xu et al., "Long-term analysis of the urban heat island effect using multisource Landsat images considering inter-class differences in land surface temperature products," *Sci. Total Environ.*, vol. 858, Feb. 2023, Art. no. 159777, doi: 10.1016/j.scitotenv.2022.159777.

[7] A. Singh, "Review article digital change detection techniques using remotely-sensed data," *Int. J. Remote Sens.*, vol. 10, no. 6, pp. 989–1003, Jun. 1989, doi: 10.1080/01431168908903939.

[8] A. Koltunov and S. L. Ustin, "Early fire detection using non-linear multitemporal prediction of thermal imagery," *Remote Sens. Environ.*, vol. 110, no. 1, pp. 18–28, Sep. 2007, doi: 10.1016/j.rse.2007.02.010.

[9] G. Xian, C. Homer, and J. Fry, "Updating the 2001 national land cover database land cover classification to 2006 by using Landsat imagery change detection methods," *Remote Sens. Environ.*, vol. 113, no. 6, pp. 1133–1147, Jun. 2009, doi: 10.1016/j.rse.2009.02.004.

[10] H. Luo, C. Liu, C. Wu, and X. Guo, "Urban change detection based on Dempster–Shafer theory for multitemporal very high-resolution imagery," *Remote Sens.*, vol. 10, no. 7, Jun. 2018, Art. no. 980, doi: 10.3390/rs10070980.

[11] M. E. Zelinski, J. Henderson, and M. Smith, "Use of Landsat 5 for change detection at 1998 Indian and Pakistani nuclear test sites," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 7, no. 8, pp. 3453–3460, Aug. 2014, doi: 10.1109/JSTARS.2013.2294322.

[12] M. Papadomanolaki, M. Vakalopoulou, and K. Karantzalos, "A deep multitask learning framework coupling semantic segmentation and fully convolutional LSTM networks for urban change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 9, pp. 7651–7668, Sep. 2021, doi: 10.1109/TGRS.2021.3055584.

[13] 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, doi: 10.1016/j.isprsjprs.2013.03.006.

[14] R. D. Jackson, "Spectral indices in N-space," *Remote Sens. Environ.*, vol. 13, no. 5, pp. 409–421, Nov. 1983, doi: 10.1016/0034-4257(83)90010-X.

[15] W. J. Todd, "Urban and regional land use change detected by using Landsat data," *J. Res. U.S. Geol. Surv.*, vol. 5, no. 5, pp. 529–534, 1977.

[16] L. Bruzzone and D. F. Prieto, "Automatic analysis of the difference image for unsupervised change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 38, no. 3, pp. 1171–1182, May 2000, doi: 10.1109/36.843009.

[17] L. I. Kuncheva and W. J. Faithfull, "PCA feature extraction for change detection in multidimensional unlabeled data," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 25, no. 1, pp. 69–80, Jan. 2014, doi: 10.1109/TNNLS.2013.2248094.

[18] A. A. Nielsen, K. Conradsen, and J. J. Simpson, "Multivariate alteration detection (MAD) and MAF postprocessing in multispectral, bitemporal image data: New approaches to change detection studies," *Remote Sens. Environ.*, vol. 64, no. 1, pp. 1–19, Apr. 1998, doi: 10.1016/S0034-4257(97)00162-4.

[19] 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, doi: 10.1109/TGRS.2013.2266673.

[20] 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, Oct. 2009, doi: 10.1109/LGRS.2009.2025059.

[21] J. S. Deng, K. Wang, Y. H. Deng, and G. J. Qi, "PCA-based land-use change detection and analysis using multitemporal and multisensor satellite data," *Int. J. Remote Sens.*, vol. 29, no. 16, pp. 4823–4838, Aug. 2008, doi: 10.1080/01431160801950162.

[22] A. Ghosh, N. S. Mishra, and S. Ghosh, "Fuzzy clustering algorithms for unsupervised change detection in remote sensing images," *Inf. Sci.*, vol. 181, no. 4, pp. 699–715, Feb. 2011, doi: 10.1016/j.ins.2010.10.016.

[23] M. Bouziani, K. Goïta, and D.-C. He, "Automatic change detection of buildings in urban environment from very high spatial resolution images using existing geodatabase and prior knowledge," *ISPRS J. Photogramm. Remote Sens.*, vol. 65, no. 1, pp. 143–153, Jan. 2010, doi: 10.1016/j.isprsjprs.2009.10.002.

[24] W. Ji, J. Ma, R. W. Twibell, and K. Underhill, "Characterizing urban sprawl using multi-stage remote sensing images and landscape metrics," *Comput. Environ. Urban Syst.*, vol. 30, no. 6, pp. 861–879, Nov. 2006, doi: 10.1016/j.compenvurbsys.2005.09.002.

[25] E. Maggiori, Y. Tarabalka, G. Charpiat, and P. Alliez, "Fully convolutional neural networks for remote sensing image classification," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2016, pp. 5071–5074, doi: 10.1109/IGARSS.2016.7730322.

[26] Q. Shi et al., "Domain adaption for fine-grained urban village extraction from satellite images," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 8, pp. 1430–1434, Aug. 2020, doi: 10.1109/LGRS.2019.2947473.

[27] M. Hu, C. Wu, L. Zhang, and B. Du, "Hyperspectral anomaly change detection based on autoencoder," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 3750–3762, Mar. 2021, doi: 10.1109/JSTARS.2021.3066508.

[28] D. Marmanis, M. Datcu, T. Esch, and U. Stilla, "Deep learning Earth observation classification using ImageNet pretrained networks," *IEEE Geosci. Remote Sens. Lett.*, vol. 13, no. 1, pp. 105–109, Jan. 2016, doi: 10.1109/LGRS.2015.2499239.

[29] 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, 2022, Art. no. 8007805, doi: 10.1109/LGRS.2021.3056416.

[30] 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, no. 4, pp. 2848–2864, Apr. 2020, doi: 10.1109/TGRS.2019.2956756.

[31] D. Wang, X. Chen, M. Jiang, S. Du, B. Xu, and J. Wang, "ADS-Net: An attention-based deeply supervised network for remote sensing image change detection," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 101, Sep. 2021, Art. no. 102348, doi: 10.1016/j.jag.2021.102348.

[32] R. C. Daudt, B. Le Saux, and A. Boulch, "Fully convolutional Siamese networks for change detection," in *Proc. IEEE 25th Int. Conf. Image Process.*, 2018, pp. 4063–4067, doi: 10.1109/ICIP.2018.8451652.

[33] 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, 2022, Art. no. 8017505, doi: 10.1109/LGRS.2021.3098774.

[34] 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, 2022, Art. no. 5604816, doi: 10.1109/TGRS.2021.3085870.

[35] S. Wei, T. Zhang, S. Ji, M. Luo, and J. Gong, "BuildMapper: A fully learnable framework for vectorized building contour extraction," *ISPRS J. Photogramm. Remote Sens.*, vol. 197, pp. 87–104, Mar. 2023, doi: 10.1016/j.isprsjprs.2023.01.015.

[36] S. Chen, Y. Ogawa, C. Zhao, and Y. Sekimoto, "Large-scale individual building extraction from open-source satellite imagery via super-resolution-based instance segmentation approach," *ISPRS J. Photogramm. Remote Sens.*, vol. 195, pp. 129–152, Jan. 2023, doi: 10.1016/j.isprsjprs.2022.11.006.

[37] B. Xu, J. Xu, N. Xue, and G.-S. Xia, "HiSup: Accurate polygonal mapping of buildings in satellite imagery with hierarchical supervision," *ISPRS J. Photogramm. Remote Sens.*, vol. 198, pp. 284–296, Apr. 2023, doi: 10.1016/j.isprsjprs.2023.03.006.

[38] H. Guo, B. Du, L. Zhang, and X. Su, "A coarse-to-fine boundary refinement network for building footprint extraction from remote sensing imagery," *ISPRS J. Photogramm. Remote Sens.*, vol. 183, pp. 240–252, Jan. 2022, doi: 10.1016/j.isprsjprs.2021.11.005.

[39] F. Fang et al., “Spatial context-aware method for urban land use classification using street view images,” *ISPRS J. Photogramm. Remote Sens.*, vol. 192, pp. 1–12, Oct. 2022, doi: 10.1016/j.isprsjprs.2022.07.020.

[40] B. Chai and P. Li, “An ensemble method for monitoring land cover changes in urban areas using dense Landsat time series data,” *ISPRS J. Photogramm. Remote Sens.*, vol. 195, pp. 29–42, Jan. 2023, doi: 10.1016/j.isprsjprs.2022.11.002.

[41] C. Szegedy et al., “Going deeper with convolutions,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 1–9, doi: 10.1109/CVPR.2015.7298594.

[42] T. Lei et al., “Difference enhancement and spatial–spectral nonlocal network for change detection in VHR remote sensing images,” *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4507013, doi: 10.1109/TGRS.2021.3134691.

Jialu Li (Graduate Student Member, IEEE) received the B.S. degree in surveying and mapping engineering from the Taiyuan University of Technology, Taiyuan, China, in 2019, and the M.E. degree in resource and environment in 2023 from the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan, China, where she is currently working toward the Ph.D. degree in photogrammetry and remote sensing.

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

Meiqi Hu (Graduate Student Member, IEEE) received the B.S. degree in surveying and mapping engineering from the School of Geoscience and Info-Physics, Central South University, Changsha, China, in 2019. She is currently working toward the Ph.D. degree with the State Key Laboratory of Information Engineering in Surveying, Mapping, and Remote sensing, Wuhan University, Wuhan, China.

Her research interests include deep learning, multitemporal remote sensing image change detection and unmixing. More information can be found by https://meiqihu.github.io/.

Chen Wu (Member, IEEE) received the B.S. degree in surveying and mapping engineering from Southeast University, Nanjing, China, in 2010, and the Ph.D. degree in photogrammetry and remote sensing from the State Key Lab of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan, China, in 2015.

He is currently a Professor with the State Key Laboratory of Information Engineering in Surveying, Mapping and Remote Sensing, Wuhan University, Wuhan, China. His research interests include multitemporal remote sensing image change detection and analysis in multispectral and hyperspectral images.