Local and Global Feature Learning With Kernel Scale-Adaptive Attention Network for VHR Remote Sensing Change Detection

Tao Lei, *Senior Member, IEEE*, Dinghua Xue, Hailong Ning, Shuangming Yang, *Member, IEEE*, Zhiyong Lv, and Asoke K. Nandi, *Fellow, IEEE*

***Abstract*—Change detection is an important task of identifying changed information by comparing bitemporal images over the same geographical area. Currently, many existing methods based on U-Net and attention mechanism have greatly promoted the development of change detection techniques. However, they still suffer from two main challenges. First, faced with the diversity of ground objects and the flexibility of scale changes, vanilla attention mechanisms cripple spatial flexibility in learning object details due to the same scale convolution kernels at different convolution layers. Second, the complex background and high similarity between changed information and nonchanged information makes it difficult to fuse low-level details and high-level semantic by simple skip-connection in U-Net. To address the above issues, a local and global feature learning with kernel scale-adaptive attention network (LGSAA-Net) is proposed in this article. The proposed network makes two contributions. First, a scale-adaptive attention (SAA) module has been designed to exploit the relationships between feature maps and convolutional kernel scales. The SAA module can achieve better feature discrimination than vanilla attention mechanism. Second, a multilayer perceptron based on patches embedding has been employed by skip-connection to learn the local and global pixel association, which is helpful for achieving globally deep fusion of low-level details and high-level semantics. Finally, experiments and ablation studies are conducted on three datasets of LEVIR/WHU/GZ. Experimental results demonstrate that the proposed LGSAA-Net performs favorably against comparative current approaches and provides more accurate contour and better internal compactness for changed targets, thus verifying the effectiveness and superiority of the proposed LGSAA-Net in VHR remote sensing change detection.**

***Index Terms*—Attention mechanism, change detection, multilayer perceptron, skip-connection.**

Manuscript received 27 June 2022; revised 29 July 2022; accepted 16 August 2022. Date of publication 23 August 2022; date of current version 8 September 2022. This work was supported in part by the National Natural Science Foundation of China under Grant 61871259, in part by Natural Science Basic Research Program of Shaanxi under Grant 2021JC-47, Grant 2022JQ-634, and Grant 2022JQ-018, in part by Key Research and Development Program of Shaanxi under Grant 2022GY436, 2021ZDLGY08-07, and Grant 2021GY-181, in part by Shaanxi Joint Laboratory of Artificial Intelligence under Grant 2020SS-03, and in part by Science and technology project of Xianyang city under Grant 2021ZDZX-GY-0001. *(Corresponding authors: Shuangming Yang; Hailong Ning.)*

Tao Lei is with the Shaanxi Joint Laboratory of Artificial Intelligence, Shaanxi University of Science and Technology, Xi'an 710021, China (e-mail: leitaoly@163.com).

Dinghua Xue is with the School of Electronical and Control Engineering, Shaanxi University of Science and Technology, Xi'an 710021, China (e-mail: xdinghua@sust.edu.cn).

Hailong Ning is with the School of Computer Science and Technology, Xi'an University of Posts and Telecommunications, Xi'an 710121, China (e-mail: ninghailong93@gmail.com).

Shuangming Yang is with the School of Electrical and Information Engineering, Tianjin University, Tianjin 300072, China (e-mail: yangshuangming@tju.edu.cn).

Zhiyong Lv is with the School of Computer Science and Engineering, Xi'an University of Technology, Xi'an 710048, China (e-mail: Lvzhiyong_fly@hotmail.com).

Asoke K. Nandi is with the Department of Electronic and Electrical Engineering, Brunel University London, London UB8 3PH, U.K., and also with the School of Mechanical Engineering, Xi'an Jiaotong University, Xi'an 710049, China (e-mail: asoke.nandi@brunel.ac.uk).

The available code of LGSAA-Net 35 can be found from https://github.com/SUST-reynole/LGSAA-Net

Digital Object Identifier 10.1109/JSTARS.2022.3200997

I. Introduction

CHANGE detection is the process of identifying differences in the state of an object or phenomenon by comparing two images at the same geographical area but of different time periods, which can reveal the dynamic changes in the surface and is one of the most important techniques in remote sensing interpretation [1]. As the Earth's surface is constantly evolving, real-time and accurate access to the surface changes is important for understanding of human activities, ecosystem, and their interactions. Recently, change detection based on VHR remote sensing images has been widely applied in land use [2], disaster monitoring [3], urban environmental investigation [4], etc. In change detection tasks, some factors, e.g., anthropogenic behavior, atmospheric conditions, and illumination, may lead to false detected regions [1], [4] and manual change detection is time-consuming and tedious. Under this circumstance, a large number of change detection approaches for remote sensing images have been proposed in recent years.

Existing change detection methods can be categorized roughly into traditional methods and deep learning-based methods. Furthermore, traditional change detection methods can be divided into pixel-based approaches [5], [6], [7] and object-based approaches [8], [9], [10]. The pixel-based approaches usually generate difference images by comparing the spectral or texture information of pixels and obtain results by using pixel classification. Compared to pixel-based approaches, the object-based approaches are working in units of objects and capture the image contextual information by processing homogeneous pixels of same objects. However, these methods usually depend on the hand-crafted features and show some

bottlenecks [7], [9], [10]. Specifically, it is difficult to design useful feature extraction operators for traditional methods, since remote sensing images are usually more complex than other natural images [11]. By contrast, deep learning methods [12], [13], [14], especially convolutional neural networks (CNNs), have been widely used in various fields due to their strong feature discrimination abilities [15], [16], [17]. As a result, a large number of deep learning-based approaches [17], [18], [19], [20] have been reported, although they achieve better change detection results by employing various of improved CNNs, they still face some challenges. On the one hand, the prevailing pooling operation in CNNs easily leads to a difficulty of detail feature extraction. On the other hand, a large number of parameters in CNNs may cause overfitting and some unpredictable problems [21]. Therefore, the current change detection methods based on CNNs still have much room for improvement.

The U-Net [22] is a very popular network in medical image segmentation, since it is specially designed for small samples training. Similar to medical images, remote sensing images also have difficulties in sample acquisition and data annotation [23], [24]. However, compared to medical images, remote sensing images often involve higher resolution, more complex image content, and more serious noise interference. Therefore, it is difficult to obtain good change detection results by applying the U-Net directly to remote sensing images [25]. The U-Net treats all feature maps equally, and thus, ignores the fact that different feature maps pay attention to different object regions. To address this problem, various improved U-Nets have been put forward by employing attention mechanism to improve the performance of networks. The receptive fields of image context at different convolution layers are diverse. However, the existing attention mechanisms usually adopt a fixed convolutional kernel scale at different convolutional layers, which is disadvantageous for image details representation of changed targets. Furthermore, U-Nets utilize the skip-connection to realize feature fusion of low-level details and high-level semantics. Although they can improve image feature discrimination abilities, the symmetric fusion ignores the association between shallow-layer and deep-layer features. Consequently, a lot of improved networks are proposed, such as UNet++ [26] and UNet3+ [27]. However, these networks conduct the connection using pixel-by-pixel fusion, which ignores the local and global information integration in an image.

To solve the above problems, a local and global feature learning with kernel scale-adaptive attention network (LGSAA-Net) is proposed for VHR remote sensing change detection in this article. On the one hand, we introduce a scale-adaptive convolution kernels strategy to solve the problem of the difficulty of image detail feature extraction caused by single-scale convolution kernels. On the other hand, for the large semantic gap between low-level details and high-level semantics caused by conventional skip-connection, we adopt the fusion of local and global features to alleviate this problem. The proposed LGSAA-Net achieves a good comprehensive performance in model complexity and change detection accuracy. The main contributions of this article are summarized follows.

1) To boost the feature learning effect on object details of VHR remote sensing images, a scale-adaptive attention (SAA) module is designed according to the change of feature map scales at different layers. The SAA module can establish the internal correlation between feature maps and convolution kernel scales.
2) To enhance effectively the local and global feature discrimination abilities of the proposed LGSAA-Net, a multilayer perceptron based on patches embedding (MLPPE) module is proposed. The MLPPE module uses a multilayer perceptron (MLP) to facilitate the global association learning of pixels, while employing the attention mechanism to learn the local correlation of different patches.

The rest of this article is arranged as follows. The related work is reviewed in Section II. Section III gives a detailed description of the proposed LGSAA-Net. The experimental results and analysis are shown in Section IV. Finally, Section V concludes this article.

II. Related Work

A. Attention Mechanism on Change Detection

Attention mechanism in vision perception relates to the process of selectively concentrating on parts of the most informative feature discrimination while suppressing the useless ones [28]. Previous literatures [29], [30] show that the attention mechanism can help CNNs to achieve better image classification and semantic segmentation. The most popular attention module is squeeze-and-excitation (SE) [31], which simply squeezes each feature map to model the cross-channel relationships in feature maps and efficiently build interdependencies among channels. To simplify the structure of the channel attention, an efficient channel attention network (ECA-Net) [32] adopts a 1-D convolution filter to compute channel weights. However, channel attention only considers encoding interchannel information but ignores the spatial details of feature maps. In order to capture the spatial details of feature maps and aggregate image contextual information, the gather-excite network (GENet) [33] and the pointwise spatial attention network (PSA-Net) [34] extend the attention mechanism by adopting different spatial attention or designing advanced attention blocks in spatial dimension. Moreover, from the perspective of interpretability, the hybrid model of combining channel attention and spatial attention is more conducive for improving network performance. Therefore, the bottleneck attention module (BAM) [35], the convolutional block attention module (CBAM) [36] and the global context network (GCNet) [37] refine convolutional features independently in channel- and spatial-dimension by cascading these two attentions. In particular, both BAM and CBAM exploit position information by reducing the channel dimension of the input tensor and then computing spatial attention using convolutions.

Compared to the attention modules mentioned above, the self-attention mechanism can effectively model the long-range dependencies by relating different positions of a single data sample. As a result, the self-attention [38] has obvious advantages in modeling long-range dependencies and building spatial- or channelwise attention. Due to the superiority of

self-attention, scholars have proposed many improved attention networks, including the nonlocal neural networks (NLNet) [39], the criss-cross attention (CCNet) [40], the dual attention network (DANet) [41], and the segmentation transformer (SETR) [42]. These networks aim to overcome the limitations of convolutional operators that only capture local relationships but fail in modeling long-range dependencies in vision tasks. Unlike in the presented models [39], [40], [41], the SETR [42] adopts a vision transformer (ViT) [43] encoder and two decoders that are designed based upon progressive upsampling and multilevel feature aggregation. Although the SETR exercises stronger reasoning and modeling abilities due to the excellent self-attention, the parallel computing increases the complexity of models, and direct upsampling or deconvolution is not conductive to global feature learning.

In recent years, attention mechanisms have also been widely used in change detection tasks [44], [45]. The Siamese CNN (Siam-Net) [44] incorporates the CBAM to Siamese network to extract adaptively spectrum-spatial features from bitemporal remote sensing images. To mitigate the problem of class imbalance in change detection, the dual task constrained deep Siamese convolutional network [45] constructs dual CBAMs for each bitemporal feature to emphasize change information. CBAM is also commonly used to refine bitemporal features, as Shi et al. [46] proposed a deeply supervised metric method. It utilizes CBAM to make the deep features of different phases more discriminative. In order to extend the advantage of self-attention in capturing long-range dependencies to remote sensing change detection tasks, a series of studies have appeared [47], [48], [49]. However, the above studies adopt fixed receptive fields at different layers, which easily lead to poor feature learning on the spatial details of changed targets. To tackle the problem, we propose a scale-adaptive attention (SAA) module in this article. The SAA module can establish the relationship between feature maps and convolution scales, implements the adaptive scale space operation on the basis of channel attention for change detection, and thus, achieves better feature learning.

B. Skip-Connection on Change Detection

In image semantic segmentation, feature fusion strategy is used to improve the problems of missed details, rough segmentation results, and low precision [50], [51], [52]. To achieve feature fusion, the skip-connection is one of the most important factors that decide the success of encoder–decoder networks in image semantic segmentation [53], [54]. The U-Net [22] applies multiple skip-connections to construct a contracting path and a symmetric expanding signal path. Similar to the U-Net, the SegNet [55] utilizes a small network structure and the skip-connection method to achieve better visual semantics as well as detailed contextures. Although the skip-connection can help the U-Net and the SegNet to achieve high segmentation accuracy, the symmetric fusion employed by skip-connection neglects the association between shallow- and deep-layer features.

In light of above problem, some improved models that can be considered as an extension of the U-Net based on skip-connection, such as the UNet++ [26] and the UNet3+ [27]. The UNet++ [26] uses a series of nested convolutional structure before feature fusion to capture contextual information, while the UNet3+ [27] applies full-scale skip-connection to capture fine-grained detail information and coarse-grained semantic information. However, as these networks achieve feature fusion in a pixel-by-pixel manner, it is not conducive to bridging effectively the semantic gap of feature maps between the encoding stage and the decoding stage. To alleviate this issue, some strategies have been designed and applied to the skip-path to improve network performance, such as the modified U-Net (mU-Net) [56] and the MultiResUnet [57]. They add some additional convolution operations before feature fusing, which reduces the difference between feature maps from encoder and decoder leading to better feature discrimination abilities. In addition, before concatenating the features at each resolution of the encoder with the corresponding features in the decoder, both the attention gate U-Net (Attention U-Net) [58] and the nonlocal U-Nets (nonlocal U-Net) [59] rescale the outputted features of the encoder by using an attention module. Furthermore, they utilize higher-level semantic features to guide the current features for attention selection, but this kind of strategies does not consider the local and global information integration in an image.

Due to the simplicity and superior performance of the skip-connection based on U-shaped structure, popular networks for change detection [25], [60], [61], [62] still depend on the U-shaped architecture. Based on UNet++, Peng et al. [25] emphasized the difference information learning by using skip-connections inside convolution units. Furthermore, Peng et al. [60] designed an improved UNet++ architecture to integrate low-level details and high-level semantics. In addition, an end-to-end LU-Net [61] is designed to leverage both spatiality and temporality characteristics simultaneously. Since change detection networks tend to focus on the extraction of semantic information and ignore the importance of shallow features, Fang et al. [62] proposed a densely connected U-Net. It reduces the loss of shallow location information by the network through compact transmission. It can be seen that those networks mentioned above can improve change detection accuracy by fusing low-level details and high-level semantics. However, due to the large semantic gap between high-level and low-level features, the existing skip-connection methods may result in limited abilities of feature discrimination. Therefore, to narrow the semantic gap, we further adopt a multilayer perceptron to learn the association of global pixels and the relationship of different patches to exploit more useful feature discrimination information.

III. METHODS

An overview of the proposed LGSAA-Net is shown in Fig. 1. First, the feature extraction is performed on VHR remote sensing images in the first encoding branch. Second, the raw difference images by performing subtracting on bitemporal images are fed into the second encoding branch to extract the difference information. Third, the result of each feature extraction layer in the second encoding branch is fused with the output result from the first encoding branch. Fourth, the subtraction operation is

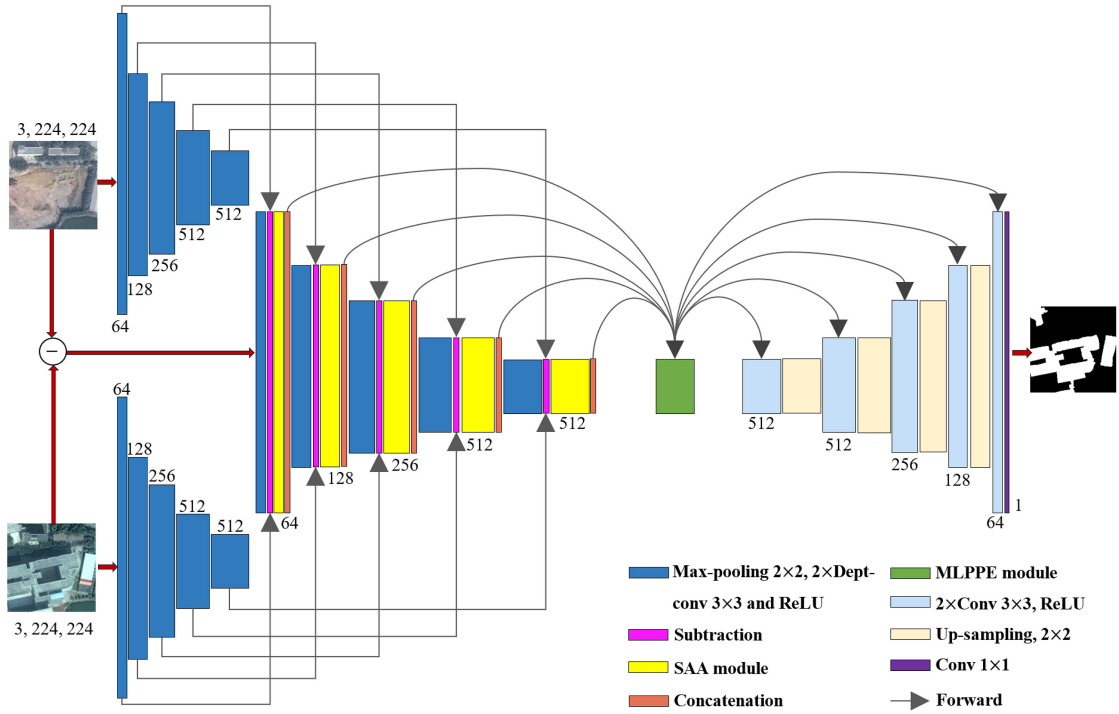


Fig. 1. Proposed LGSAA-Net for change detection. The backbone of LGSAA-Net is borrowed from the U-Net. Our architecture consists of the encoding stage and the decoding stage. In order to make the network lighter, stack depth-wise separable convolutional (Depth-conv) operators denoted by blue boxes are employed to extract features in the encoding stage. The two encoding branches convert the bitemporal images and the difference image to feature maps, and the number of channels is denoted on top of each box. For change detection, the SAA module denoted by yellow box is laid to each stack depth-conv operators, which can achieve better feature discrimination ability by establishing the internal correlation between feature maps and convolution kernel scales. In addition, the MLPPE module denoted by green box is laid to each skip-path, which facilitates the global association learning of pixels and learns the local correlation of different patches.

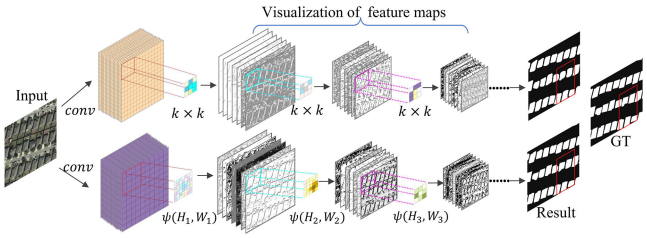


Fig. 2. Comparison of the single-scale and multiscale convolution kernels on multiresolution images. It is found that the later provides better change detection results than the former due to the employment of multiscale convolution kernels.

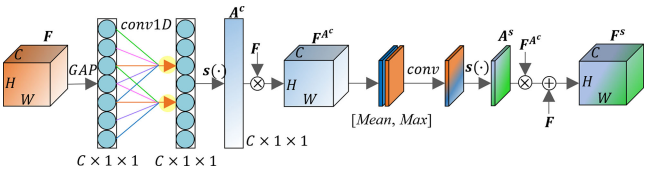


Fig. 3. SAA module.

performed on feature maps from the corresponding bitemporal paths. Finally, the fused features are fed into the next encoding layer.

In order to refine the target contour, the SAA module is proposed to establish the relationships between feature maps and convolution kernel scales, which is described in detail in Fig. 3. We also present the structure of the MLPPE, as shown in Fig. 5, which can learn local correlation of different patches and

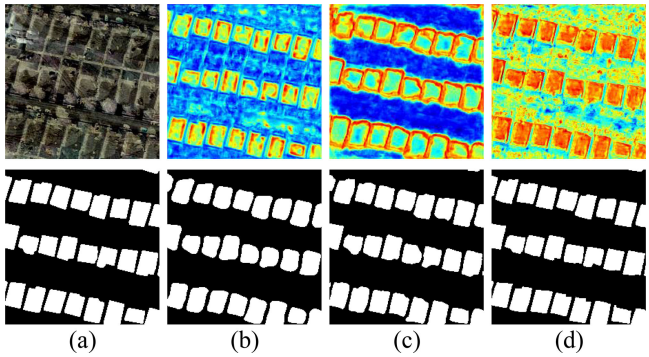


Fig. 4. Comparison of change detection using different networks. The first column: (a) Raw difference image and ground truth image. From the second to the fourth column, top is heatmaps and the bottom is change detection results from: (b) U-Net; (c) attention U-net; and (d) U-Net+multilayer perceptron.

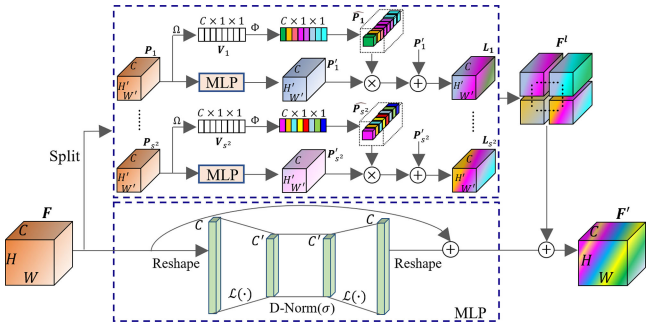


Fig. 5. MLPPE module.

facilitate the global association learning of pixels. In general, the proposed LGSAA-Net can effectively improve its feature discrimination abilities and provide excellent change detection results.

A. The SAA Module for Change Detection

Multiscale Convolution Kernels: The utilization of multiscale information is an important strategy in image segmentation applications, since multiscale convolutional kernels can learn richer features. Generally, fine-grained sampling can obtain richer detail information, while coarse-grained sampling can extract richer contextual information. The latter is in favor of getting the overall trend of an image information. In addition, the existing spatial attention networks for change detection often utilize convolution kernels with fixed size to harvest the correlation of image spatial position information, which leads to the problem of the limited performance of target contour detection. Fig. 2 shows the comparison of the single-scale and multiscale convolution kernels on multiresolution images. It is clear that the latter provides better change detection results than the former due to the employment of multiscale convolution kernels.

Design of SAA Module: In light of the above discussion, the SAA module is designed based on the scale changes of feature maps in the encoding stage. Specifically, let the output of the previous layer of the network $\boldsymbol{F} \in \mathbb{R}^{C\times H\times W}$ be the input feature maps of the module as shown in Fig. 3, then the global average pooling and convolutional operation are performed on $\boldsymbol{F}$ to obtain the A^c that refers to compressed output score of each channel. To simplify the symbol mark, we use GAP to represent the operation of global average pooling. The specific calculations are defined as follows:

$$A^c = s\left(\text{conv1}D\left(\text{GAP}\left(\boldsymbol{F}\right), k^c\right)\right) \tag{1}$$

$$T\left(\cdot\right) = \text{ROUND}|\frac{\left(\log\left(\left(C,\beta\right)\right)+\alpha\right)}{2}| \tag{2}$$

where, $A^c \in \mathbb{R}^{C\times 1\times 1}$, conv1$D$ represents 1-D convolution operation, $s(\cdot)$ stands for the sigmoid function. k^c is the size of the convolution kernel. If T is an even number, then $k^c = T(\cdot)$, otherwise $k^c = T(\cdot)+1$. In this article, we set α and β to 1 and 2 according to the empirical value in the experiment, respectively. ROUND denotes the rounding operation. The refined output is defined as

$$\boldsymbol{F}^{\boldsymbol{A^c}} = \wedge_{mc}\left(\boldsymbol{F}, A^c\right) \tag{3}$$

where $\wedge_{mc}$ denotes channelwise multiplication. This operation can adaptively adjust the value of k^c according to the number of channels, and conveniently obtain channel interactivity information.

Furthermore, the average value and maximum value in channel dimension on the feature maps $\boldsymbol{F}^{\boldsymbol{A^c}} \in \mathbb{R}^{C\times H\times W}$ are calculated, and the 2-D convolution is performed to complete the space mapping. The calculation formulae are given as follows:

$$A^s = s\left(\text{conv}\left(\text{concat}\left(\phi_{\text{ave}}\left(\boldsymbol{F}^{\boldsymbol{A^c}}\right), \phi_{\max}\left(\boldsymbol{F}^{\boldsymbol{A^c}}\right)\right), k^s\right)\right) \tag{4}$$

$$k^s = \text{ROUND}\left(\gamma \times |\log\left(H\times W, \varepsilon\right)|\right) \tag{5}$$

where $\phi_{\text{ave}}(\cdot)$ represents the channel mean operation, $\phi_{\max}(\cdot)$ denotes channel maximum operation, concat denotes the concatenate operation, and conv stands for the 2-D spatial convolution operation. k^s is the size of convolution kernels. In this article, we set ε and γ to 10 and 3 according to the empirical value in the experiment, respectively. A^s is the scale adaptive spatial attention weight. Finally, the refined output of the SAA module denoted by $\boldsymbol{F^s}$ is calculated as follows:

$$\boldsymbol{F^s} = \wedge_A\left(\boldsymbol{F}, \wedge_{ms}\left(\boldsymbol{F}^{\boldsymbol{A^c}}, A^s\right)\right) \tag{6}$$

where $\wedge_{ms}$ represents elementwise multiplication, and $\wedge_A$ denotes elementwise addition. The SAA module employs the combination of channel attention and spatial attention with the characteristics of scale adaption, establishes the relationships between the feature maps and the convolution kernel scales, and realizes the scale-adaptive spatial attention operation.

B. MLPPE Module for Change Detection

Multilayer Perceptron: The current networks [60], [61], [62] improve change detection accuracy by fusing low- and high-level features. However, these methods mainly adopt the pixel-by-pixel fusion strategy, which ignores the integration of local and global information. Thus, most networks usually employ full connection layers in middle and high-level layers, to summarize features globally and help the network effectively learning global information. In light of discussion above, we employ full connection layers in feature fusion stage to summarize features globally and help the network effectively learning global information. As shown in Fig. 4, compared with the U-Net and Attention U-Net [58], the network with multilayer perceptron [63], named U-Net + multilayer perceptron, obtains more intuitive heat maps by fusing low- and high-level features. At the same time, it provides better detection results than the U-Net and the attention U-Net, which shows that multilayer perceptron is helpful for improving change detection results. In order to further improve the network performance, the patches strategy from the original feature maps are utilized in MLPPE module.

Global Features Based on Multilayer Perceptron: In this section, we present the MLPPE module, as shown in Fig. 5. Let feature map $\boldsymbol{F}$ be an input tensor, the input feature maps are reshaped to $\boldsymbol{M}$ to facilitate the next full connection operation. The specific calculations are given as follows:

$$\boldsymbol{M} = \mathcal{L}\left(\mathcal{R}\left(\boldsymbol{F}\right)\right) \tag{7}$$

$$\boldsymbol{F^g} = \delta\left\{\widehat{\mathcal{R}}\left(\mathcal{L}\left(\sigma\left(\boldsymbol{M}\right)\right)\right)\right\} \tag{8}$$

where $\mathcal{L}(\cdot)$ denotes the linear operation between different layers, $\mathcal{R}(\cdot)$ denotes the reshape operation to realize $\boldsymbol{F} \in \mathbb{R}^{C\times H\times W} \rightarrow \boldsymbol{M} \in \mathbb{R}^{C\times HW}$, C is the number of channels, $H\times W$ is the size of the current input feature maps. $\widehat{\mathcal{R}}(\cdot)$ represents the inverse operation of $\mathcal{R}(\cdot)$ and is used to achieve $\boldsymbol{M} \in \mathbb{R}^{C\times HW} \rightarrow \boldsymbol{F^g} \in \mathbb{R}^{C\times H\times W}$. In the MLPPE, we employ two linear layers, and δ refers to the ReLU function. Furthermore, in order to avoid the problem of linear operations being sensitive to the intermediate input scale, the double-normalization method [64] is used in the MLPPE. σ stands for double-normalization operation. $\widehat{\boldsymbol{M'}}$

denotes the result of doule-normalization. Here $\sum_j \widehat{\boldsymbol{M}}'_{\boldsymbol{i,j}} = 1$, where i and j denote index of each dimension, respectively, and k is utilized to index current location information. The calculation formulas are given as follows:

$$\widehat{\boldsymbol{M}}_{\boldsymbol{i,j}} = \frac{\exp\left(\boldsymbol{M}_{\boldsymbol{i,j}}\right)}{\sum_k \exp\left(\boldsymbol{M}_{\boldsymbol{k,j}}\right)} \tag{9}$$

$$\widehat{\boldsymbol{M}}'_{\boldsymbol{i,j}} = \frac{\exp\left(\widehat{\boldsymbol{M}}_{\boldsymbol{i,j}}\right)}{\sum_k \exp\left(\widehat{\boldsymbol{M}}_{\boldsymbol{i,k}}\right)}. \tag{10}$$

Local features Based on Patches Channel Information: For VHR images containing rich spatial information and complex contexture, the local spatial details of targets play a vital role in change detection, and the modeling for channel correlation is also beneficial to improving feature discrimination abilities. Therefore, the MLPPE module captures the image attention information with spatial property by learning both patches channel association and image local spatial details, which further boosts the change detection accuracy for VHR remote sensing images. As shown in Fig. 5, $\boldsymbol{F}$ is sequentially divided into s^2 patches $\{\boldsymbol{P_1}, \boldsymbol{P_2}, \ldots, \boldsymbol{P_{s^2}}\}$. In which $\boldsymbol{P_n} \in \mathbb{R}^{C\times H'\times W'}$, where $H' = H/(s^2)$, $W' = W/(s^2)$, and $1 \leqslant n \leqslant s^2$, $n, s \in \mathrm{N}_+$. To simplify the symbol mark in Fig. 1(c), we use $\boldsymbol{\Omega}$ to synthesize the calculations of (11) and (12). Then we get the output feature information $\boldsymbol{V_n} \in \mathbb{R}^{C\times 1\times 1}$ of each corresponding patch $\boldsymbol{P_n}$ by implementing (11) and (12). The specific calculation formulas are defined as follows:

$$\boldsymbol{U_n} = W_n * \boldsymbol{P_n} \tag{11}$$

$$v_n^m = \frac{1}{H'\times W'} \sum_{n=i}^{H'} \sum_{n=i}^{W'} u_n^m(i,j) \tag{12}$$

where n indexes the nth patch of feature maps. $\boldsymbol{U_n} \in \mathbb{R}^{C\times H'\times W'}$ is the output feature map. Besides, $W_n = [w_n^1, w_n^2, \ldots, w_n^C]$, w_n^m is a mth 2-D spatial filter kernel. $\boldsymbol{V_n} = [v_n^1, v_n^2, \ldots, v_n^C]$, in which v_n^m denotes mth channel association of $\boldsymbol{V_n}$. Then, two linear layers are applied to establish the channel association of $\boldsymbol{V_n}$ and obtain features $\widehat{\boldsymbol{P}}_{\boldsymbol{n}}$. The specific calculation is defined as follows:

$$\widehat{\boldsymbol{P}}_{\boldsymbol{n}} = s\left(\mathcal{L}\left(\delta\left(\mathcal{L}\left(\boldsymbol{V_n}\right)\right)\right)\right) \tag{13}$$

similarly, to simplify the symbol mark in Fig. 5, we use $\boldsymbol{\Phi}$ to synthesize the calculation of (13), in which $\widehat{\boldsymbol{P}}_{\boldsymbol{n}} \in \mathbb{R}^{C\times 1\times 1}$ denotes the refined channel output of each patch. $\mathcal{L}(\cdot)$ denotes the linear operation between layers. Besides, $\{\widehat{\boldsymbol{P}}_{\boldsymbol{1}}, \widehat{\boldsymbol{P}}_{\boldsymbol{2}}, \ldots, \widehat{\boldsymbol{P}}_{\boldsymbol{s^2}}\}$ is the channel correlation on patch-level and the relevance of patch spatial details.

In this section, the globally contextual information of VHR remote sensing images is achieved by integrating the patch embedding result into the multilayer perceptron. Then weights obtained by patch embedding are applied to the patches corresponding to the global feature maps obtained by the multilayer perceptron. Similar to $\boldsymbol{F^s}$ in (7)–(8), we compute local patch information results $\{\boldsymbol{P'_1}, \boldsymbol{P'_2}, \ldots, \boldsymbol{P'_{s^2}}\}$ by performing (7) and (8) on $\{\boldsymbol{P_1}, \boldsymbol{P_2}, \ldots, \boldsymbol{P_{s^2}}\}$. The calculation formulas are defined as follows:

$$\boldsymbol{L_n} = \wedge_A\left(\boldsymbol{P'_n}, \wedge_{mc}\left(\boldsymbol{P'_n}, \widehat{\boldsymbol{P}}_{\boldsymbol{n}}\right)\right) \tag{14}$$

$$\{\boldsymbol{L_1}, \boldsymbol{L_2}, \ldots, \boldsymbol{L_n}\} \Rightarrow \boldsymbol{F^l} \tag{15}$$

$$\boldsymbol{F'} = \wedge_A\left(\boldsymbol{F^g}, \boldsymbol{F^l}\right) \tag{16}$$

where $\boldsymbol{L_n}$ represents the output results of nth patch feature information, $\wedge_{mc}$ denotes channelwise multiplication, $\wedge_A$ denotes elementwise addition, $\Rightarrow$ denotes the operation that can unite all patches, and $\boldsymbol{F^l}$ denotes final output with channel weights information and local patch information. Then we obtain the final refined feature fusion output $\boldsymbol{F'} \in \mathbb{R}^{C\times H\times W}$.

IV. Experiments and Analysis

In order to evaluate the proposed method, some state-of-the-art methods, including FC-EF [65], FC-di [65], FC-conc [65], FCN-PP [66], FDCNN [67], DSIFN [68], SRCD-Net [18], Trans-CD [19], are considered as comparative methods in our experiments. We complete the comparisons by the released model codes. Furthermore, we conducted the ablation studies to prove the validity of each component.

A. Experimental Setup

Datasets: In this article, three benchmark datasets, including LEVIR, WHU, and GZ, are used to assess the proposed method. All of these datasets contain raw bitemporal images, and ground truths.

LEVIR Dataset [69] is a building change detection dataset with a spatial resolution of 0.55 m. It contains 637 pairs of bitemporal images with size of 1024 × 1024. These bitemporal images are within a time span of 5 to 14 years and have significant land changes, especially the growth of buildings covered by various types of buildings, such as villas, high apartments, small garages, and large warehouse. The fully annotated LEVIR dataset contains a total of 31 333 individual changed examples. We applied overlapping and nonoverlapping manners to crop the data into image patches with size of 224 × 224, then obtained 11 083 training samples, 2880 validation samples, and 2048 testing samples.

WHU Dataset [70] is a building change detection dataset with a spatial resolution of 0.075 m. It contains one pair of bitemporal images with size of 32 507 × 15 354. We first divide the bitemporal images into four smaller images without overlapping: 32 507 × 12610/18 361 × 2744/7634 × 2744/6511 × 2744. We used the first patch as the training set, the second and third patches as the validation set, and the fourth patch as the testing set. Then, we cropped these data into image patches with size of 224 × 224, obtaining 9637 training samples, 2494 validation samples, and 1600 testing samples.

GZ Dataset [71] is acquired during 2006 and 2019 period. These bitemporal images cover the suburb area of Guangzhou City, China. In order to align the image pairs, it collects 20 pairs of bitemporal images that change with the season varying by BIGEMAP software of Google Earth. These 20 pairs of bitemporal images, which have a spatial resolution of 0.55 m

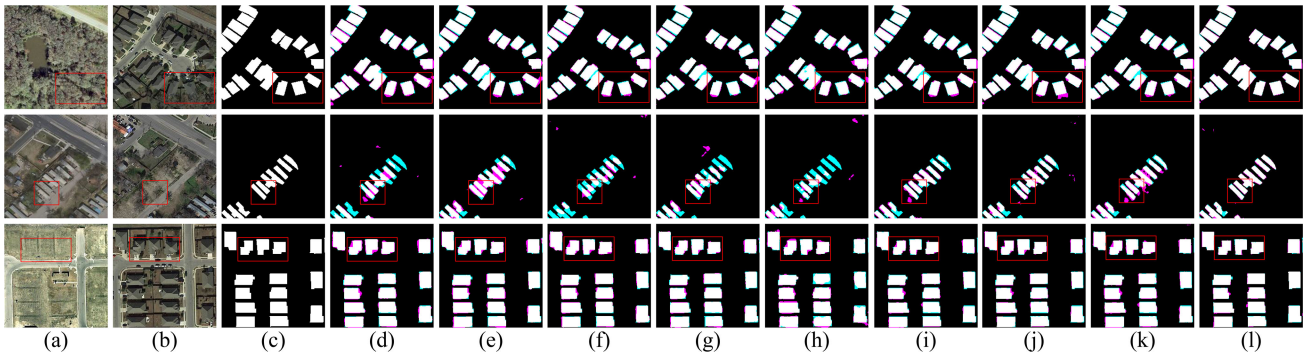


Fig. 6. Experimental results on LEVIR dataset. (a) Pretemporal images. (b) Posttemporal images. (c) Ground truths. (d) FC-EF. (e) FC-di. (f) FC-conc. (g) FCN-PP. (h) FDCNN. (i) DFISN. (j) SRCD-Net. (k) Trans-CD. (l) LGSAA-Net. Note that, the black color represents the unchanged regions, the white color represents the changed regions, the pink color denotes false-detected regions, and the cyan color denotes true-missed regions.

and a size range of 1006 × 1168 pixels to 4936 × 5224 pixels, are divided into three parts: training set (14 pairs)/validation set (3 pairs)/testing set (3 pairs). Then, we cropped these data into image patches with size of 224 × 224. Finally, we obtained 5612 training samples, 1692 validation samples, and 1456 testing samples.

Implementation Details: We implemented the proposed LGSAA-Net with PyTorch and trained it on a NVIDIA GeForce RTX 2080Ti GPU with 11 GB RAM. In this article, the parameter setting of comparative approaches follows original papers. For the proposed LGSAA-Net, we set $s = 2$ for the MLPPE module. The Adam optimizer is adopted with a learning rate of 10^{-4} as an optimization algorithm, and the batch size of the training data is set to 16.

Evaluation Metrics: To evaluate the performance of the proposed LGSAA-Net, five popular metrics have been adopted, including precision (Pre), recall (Rec), overall error (OE), overall accuracy (OA), and F1-score ($F1$). Specifically, the Pre denotes the ratio of detected areas that are really changed regions in totally detected regions. The Rec denotes the ratio of detected areas that are really changed regions compare to ground truths. The OE is usually used to evaluate the overall error ratio of object detection. OE, OA, and $F1$ are the overall evaluation indexes of the prediction change detection results. Smaller value of OE and larger values of OA and $F1$ mean that the better prediction change detection results, and vice versa. Moreover, there is a large number of testing images in each of the three datasets (e.g., LEVIR, WHU, and GZ datasets). We use $[\![\text{pre}]\!]\_\text{m}, [\![\text{rec}]\!]\_\text{m}, [\![\text{oe}]\!]\_\text{m}, [\![\text{oa}]\!]\_\text{m}, \text{f} - [\![\text{score}]\!]\_\text{m}$ to define the metrics of the m-th testing sample, in which m denotes the number of testing samples in each dataset. These metrics are defined as follows:

$$Pre = \frac{TP}{TP + FP} \tag{17}$$

$$Rec = \frac{TP}{TP + FN} \tag{18}$$

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

$$OA = \frac{TP + TN}{TP + TN + FP + FN} \tag{20}$$

$$F1 = 2 \times \frac{\text{Pre} \times \text{Rec}}{\text{Pre} + \text{Rec}} \tag{21}$$

where the TP (true positive) denotes the total number of pixels accurate-detected on really changed regions, the TN (true negative) means the total number of pixels accurate-detected on really unchanged regions, the FP (false positive) stands for the total number of pixels over-detected, and the FN (false negative) is the total number of pixels miss-detected, respectively.

B. Comparison With State-of-The-Art Methods

Comparison on LEVIR Dataset: Fig. 6 shows the change detection results on LEVIR dataset, where Fig. 6(a)(c) are the bitemporal images and the ground truths, respectively. In Fig. 6(d)(f), the first three comparison methods are based on fully convolution network and feature fusion methods. It can be seen that the change detection results provided by FC-di are better than FC-EF and FC-conc, which shows that the Siamese encoder can slightly improve the model accuracy. Also, the results provided by FC-EF are inferior to FC-di, but better than FC-conc, which indicates that FC-EF can extract better discriminative features from bitemporal images than FC-conc. In addition, as shown in Fig. 6(g) and (h), although FCN-PP and FDCNN miss some truly changed regions (cyan color) in sample_2, they achieve better change detection results in sample_1 and sample_3, since the Gaussian pyramid module of FCN-PP possesses a strong feature discrimination ability, and the multiscale and multidepth feature difference maps generated by FDCNN are beneficial for change detection. Thus, FCN-PP and FDCNN provide better change detection results than the first three comparison methods. In contrast, the missed regions (cyan color) in Fig. 6(i)–(k) are greatly reduced, and their internal compactness of objects are improved compared with the results in Fig. 6(d)–(h). Fig. 6(l) shows that the proposed LGSAA-Net achieves the best change detection results with complete boundaries and high internal compactness, since it uses patch embedding and multilayer perceptron to learn local

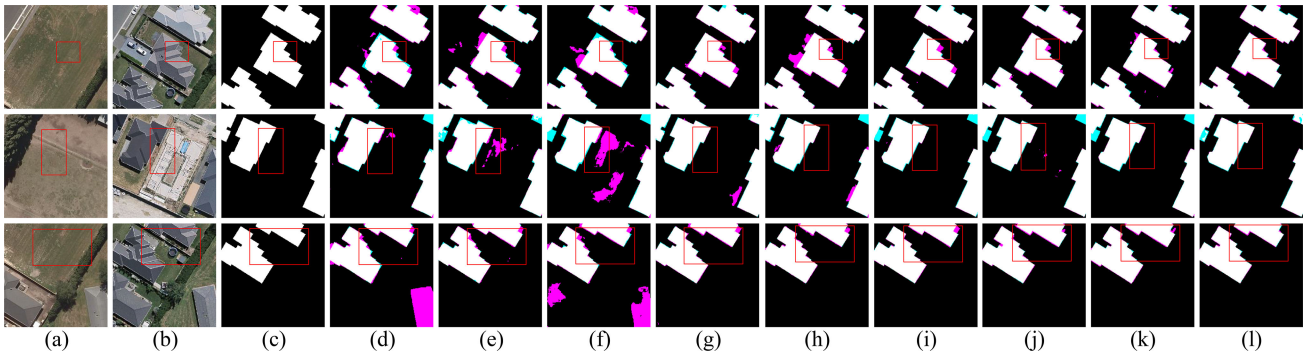


Fig. 7. Experimental results on WHU dataset. (a) Pretemporal images. (b) Posttemporal images. (c) Ground truths. (d) FC-EF. (e) FC-di. (f) FC-conc. (g) FCN-PP. (h) FDCNN. (i) DFISN. (j) SRCD-Net. (k) Trans-CD. (l) LGSAA-Net. Note that, the black color represents the unchanged regions, the white color represents the changed regions, the pink color denotes false-detected regions, and the cyan color denotes true-missed regions.

TABLE I
QUANTITATIVE EVALUATION RESULTS ON LEVIR DATASET

Methods	Evaluation metrics				
	Pre	Rec	OE	OA	$F1$
FC-EF [65]	0.8748	0.8452	0.0307	0.9693	0.8475
FC-di [65]	0.8870	0.8438	0.0270	0.9730	0.8531
FC-conc [65]	0.8668	0.7985	0.0347	0.9653	0.8211
FCN-PP [66]	0.8581	0.8901	0.0273	0.9727	0.8642
FDCNN [67]	0.9070	0.8455	0.0262	0.9738	0.8655
DFISN [68]	0.9099	0.8825	0.0231	0.9769	0.8890
SRCD-Net [18]	0.9175	0.8901	0.0209	0.9791	0.8981
Trans-CD [19]	0.9045	0.8643	0.0259	0.9741	0.8779
LGSAA-Net	**0.9290**	**0.9037**	**0.0185**	**0.9815**	**0.9116**

Best values are in bold.

TABLE II
QUANTITATIVE EVALUATION RESULTS ON WHU DATASET

Methods	Evaluation metrics				
	Pre	Rec	OE	OA	$F1$
FC-EF [65]	0.8102	0.8159	0.0301	0.9699	0.7945
FC-di [65]	0.8801	0.8250	0.0230	0.9770	0.8317
FC-conc [65]	0.6881	0.8271	0.0506	0.9494	0.7187
FCN-PP [66]	0.8910	0.8567	0.0194	0.9806	0.8539
FDCNN [67]	0.8747	0.8537	0.0208	0.9792	0.8457
DFISN [68]	0.9226	0.8691	0.0185	0.9815	0.8766
SRCD-Net [18]	0.9201	0.8819	0.0174	0.9826	0.8846
Trans-CD [19]	0.9129	0.8557	0.0169	0.9831	0.8716
LGSAA-Net	**0.9310**	**0.8947**	**0.0147**	**0.9853**	**0.9019**

Best values are in bold.

and global pixel association, and the SAA module makes the network learn the feature map information more reasonably.

The quantitative evaluation results on LEVIR dataset are summarized in Table I. It can be seen that FC-di obtains higher value of $F1$ among the first three comparative methods, since FC-di explicitly guides the network to compare the differences between the bitemporal images. The last three comparative methods show more satisfactory results. Among them, SRCD-Net obtains the highest value of $F1$ 89.81%, respectively, since the stacked attention module is in favor of capturing changed information. The third-ranked DFISN obtains $F1$ 88.90% due to the effectiveness of deep supervision for change detection. It is worth noting that the proposed LGSAA-Net achieves the lowest value of OE and the highest value of $F1$. Besides, the proposed LGSAA-Net obtains an extra 1.35% on F1 than the best result from comparative approaches due to the useful discriminative features provided by our LGSAA-Net.

Comparison on WHU Dataset: Fig. 7 shows the change detection results on WHU dataset, where Fig. 7(a)–(c) correspond to the bitemporal images and the ground truths, respectively. These changed targets mainly concentrated on buildings and suburban houses. In Fig. 7(a) and (b), the contrast of changed targets in bitemporal images is quite low, which may affect the accuracy of change detection. Fig. 7(d)–(g) contain obvious falsely changed regions (pink color). In contrast, the results in Fig. 7(h)–(k) provided by FDCNN, DSIFN, SRCD-Net, Trans-CD, and the proposed LGSAA-Net are better than the results in Fig. 7(d)–(g). Notably, the proposed LGSAA-Net can accurately detect the contour information of small changed targets more accurately, and it obtains good change detection results in Fig. 7(l) that are close to ground truths.

Table II, respectively, shows the quantitative evaluation results on WHU datasets. Compared to the LEVIR dataset, the first three comparative methods show similar performance on WHU dataset. FC-di provides higher value of $F1$ among the first three comparative methods, since FC-di considers the differences of bitemporal images in the encoding stage. Different from the results provided by FCN-PP and FDCNN on LEVIR dataset, FCN-PP outperforms FDCNN, since FCN-PP employs Gaussian pyramid to improve the ability of feature learning of models. In addition, SRCD-Net obtains the highest value of $F1$ among the last three comparative methods, DFISN and Trans-CD provide similar accuracy on $F1$. Notably, we can see that the proposed LGSAA-Net obtains the

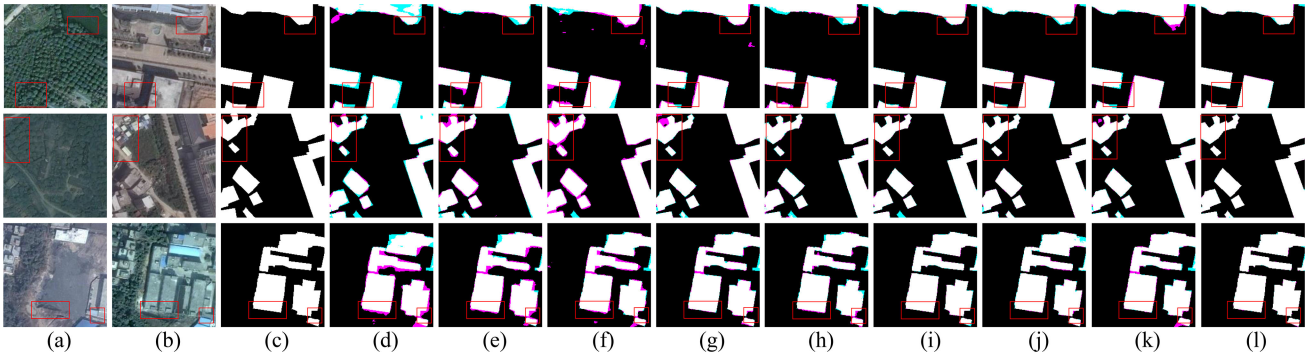


Fig. 8. Experimental results on GZ dataset. (a) Pretemporal images. (b) Posttemporal images. (c) Ground truths. (d) FC-EF. (e) FC-di. (f) FC-conc. (g) FCN-PP. (h) FDCNN. (i) DFISN. (j) SRCD-Net. (k) Trans-CD. (l) LGSAA-Net. Note that, the black color represents the unchanged regions, the white color represents the changed regions, the pink color denotes false-detected regions, and the cyan color denotes true-missed regions.

TABLE III
QUANTITATIVE EVALUATION RESULTS ON GZ DATASET

Methods	Evaluation metrics				
	Pre	Rec	OE	OA	$F1$
FC-EF [65]	0.9330	0.9682	0.0169	0.9831	0.8924
FC-di [65]	0.9320	0.9396	0.0105	0.9895	0.9318
FC-conc [65]	0.8842	0.9577	0.0148	0.9852	0.9145
FCN-PP [66]	0.9528	0.9539	0.0066	0.9934	0.9504
FDCNN [67]	0.9526	0.9471	0.0077	0.9923	0.9471
DFISN [68]	0.9720	0.9677	0.0033	0.9967	0.9686
SRCD-Net [18]	0.9776	0.9675	0.0032	0.9968	0.9708
Trans-CD [19]	0.9571	0.9567	0.0057	0.9943	0.9543
LGSAA-Net	**0.9846**	**0.9811**	**0.0014**	**0.9986**	**0.9822**

Best values are in bold.

highest value of $F1$. Furthermore, the proposed LGSAA-Net achieves a performance improvement of 1.73% on $F1$ than the best result in comparative approaches, which further illustrates the advantage of the proposed LGSAA-Net for change detection.

Comparison on GZ Dataset: Fig. 8 shows the change detection results on GZ dataset to further demonstrate the superiority and generalizability of the proposed LGSAA-Net, where Fig. 8(a)–(c) show the bitemporal images and the ground truths, respectively. Fig. 8(a) and (b) show that images from the GZ dataset contains more noise than LEVIR and WHU datasets. Therefore, some falsely changed regions (pink color) are apparent in Fig. 8(d)–(g). Compared to the first four comparative methods, the results provided by FDCNN, DSIFN, SRCD-Net, and Trans-CD are improved, as shown in Fig. 8(h)–(k). Also, it can be seen from Fig. 8(d)–(l) that the proposed LGSAA-Net provides better change detection results than comparative methods, which further verifies the advantages of the proposed LGSAA-Net for change detection.

As can be seen from Table III, the proposed method also significantly outperforms all comparative methods on GZ dataset, achieving the highest value of $F1$. Different from the results on LEVIR and WHU datasets, FC-conc performs higher accuracy than FC-EF on GZ dataset, this indicates that FC-conc on different datasets shows inconsistent performance. FC-di also achieves best segmentation accuracy among the first three comparative methods, since FC-di considers the differences of bitemporal images in the encoding stage. Also, SRCD-Net obtains higher value of $F1$ than the scores obtained by other comparative methods, since the stacked attention module in SRCD-Net is in favor of capturing changed information. Compared with SRCD-Net, the proposed LGSAA-Net obtains an extra raising 1.14% on $F1$ owing to better feature learning of the SAA module and effectively fusion of low-level details and high-level semantics of the MLPPE module. From analysis above, the proposed LGSAA-Net is effective for obtaining accurate change detection results.

C. Ablation Studies

To illustrate further the effectiveness of different modules in the proposed network, experiments about various combinations of modules are conducted on LEVIR, WHU, and GZ datasets. Fig. 9(a)–(c) are the bitemporal images and ground truths, respectively. Added modules corresponding to Fig. 9(d)–(i) are abbreviated U-Net (Base) [22], Siamese U-Net (Siam) [65], Transformer-Vit (ViT) [43], MLPPE, multibranch encoding (MB) [65], efficient channel attention (ECA) [32], the convolutional block attention module (CBAM) [36], and SAA, respectively. The ablation schemes include: U-Net based on difference images (Base+DI), Siamese U-Net based on bitemporal images (Base+Siam), Siamese U-Net based on bitemporal images and Transformer-ViT (Base+Siam+ViT), Siamese U-Net based on MLPPE (Base+Siam+MLPPE), Siamese U-Net and MLPPE with multibranch encoding (Base+Siam+MLPPE+MB), Siamese U-Net and MLPPE based on MB and ECA (Base+Siam+MLPPE+MB+ECA), Siamese U-Net and MLPPE based on MB, and CBAM (Base+Siam+MLPPE+MB+CBAM) and LGSAA-Net.

As shown in Fig. 9(c)–(f), it can be concluded that the feature extraction methods based on bitemporal images can obtain better change detection results than the methods based on difference images. The ViT module does improve the accuracy

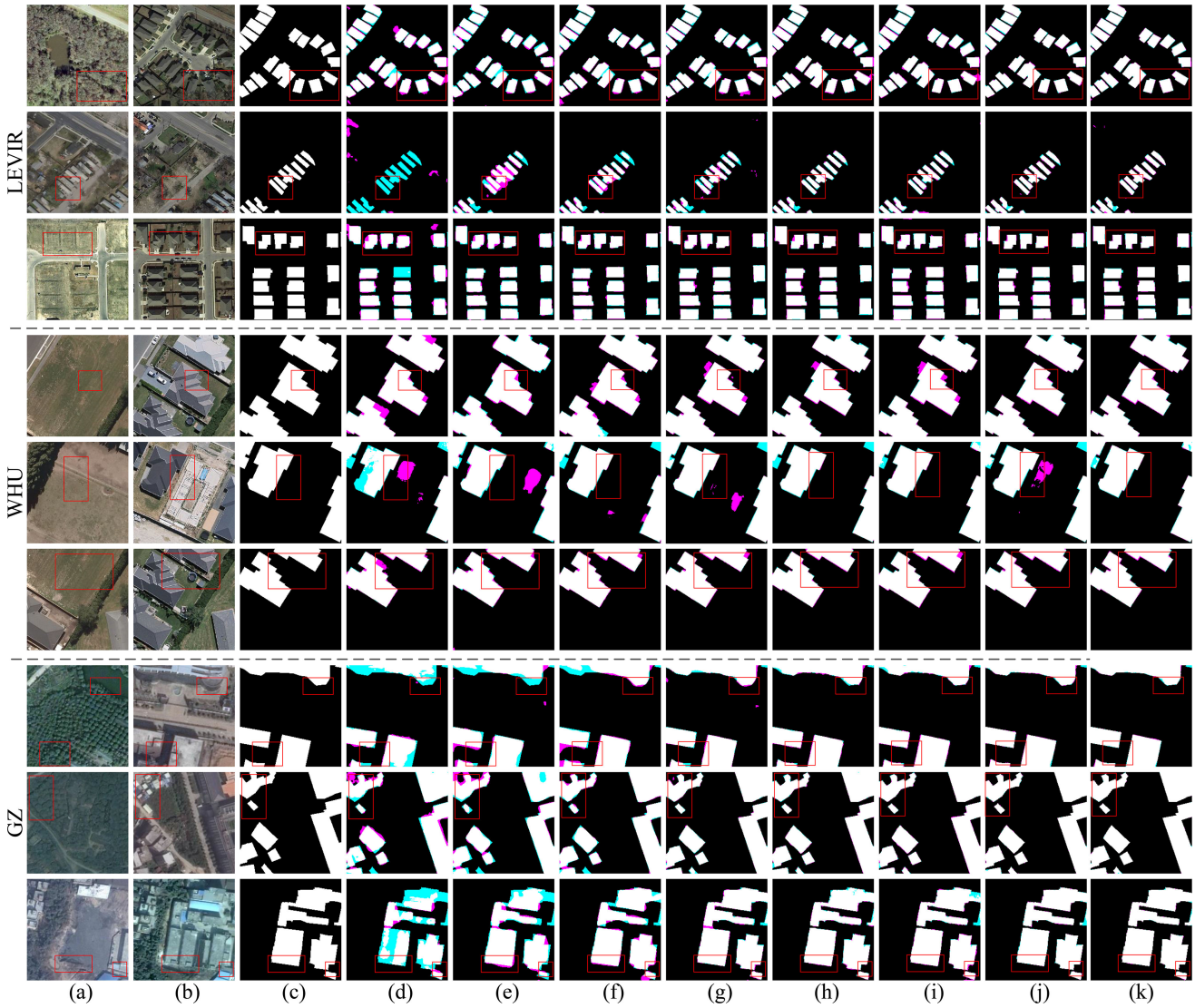


Fig. 9. Comparison of ablation experiments on LEVIR, WHU and GZ datasets: (a) pre-temporal images, (b) post-temporal images, (c) ground truths, (d) Base+DI, (e) Base+Siam, (f) Base+Siam+ViT, (g) Base+Siam+MLPPE, (h) Base+Siam+MLPPE+MB, (i) Base+Siam+MLPPE+MB+ECA, (j) Base+Siam+MLPPE+MB+CBAM (k) LGSAA-Net. Note that, the black color represents the unchanged regions, and the white color represents the changed regions, the pink color denotes false-detected regions, and the cyan color denotes true-missed regions.

of change detection, as shown in Fig. 9(f), but the sequence of image features completely replaces features maps, which ignores the contextual structure information of feature maps from original CNN and leads to false regions (pink color). In addition, it can be seen from Fig. 9(f) and (g) that the MLPPE module is beneficial to improving change detection results. On this basis, we added the multibranch encoding strategy to further enhance feature discrimination capabilities for bitemporal images and difference images, as shown in Fig. 9(h). Furthermore, compared with Fig. 9(h)–(k), it shows that the result by adding CBAM module is better than ones adding ECA module, but inferior to ones adding the SAA module, which indicates that the SAA can better respond to feature extraction of feature maps with different resolutions. In conclusion, for change detection task, the proposed LGSAA-Net can obtain clearer changed regions with more complete boundaries, and maintains a high internal compactness in truly changed regions. Table IV shows the quantitative evaluation results of our ablation experiments on LEVIR, WHU, and GZ datasets. It can be seen that the change detection results are improved with different degrees by adding these modules. Obviously, the incorporation of both the MLPPE and the SAA modules can improve the performance of networks on three datasets, which indicates that the proposed LGSAA-Net has a positive impact on change detection.

V. Discussion

In this section, the discussions about the effectiveness of the SAA and the MLPPE modules, the sensitivity experiments in the MLPPE module, as well as the model complexity are presented to demonstrate further the contributions of our studies.

A. Discussion on the Effectiveness of the SAA and the MLPPE

In order to show the feature extraction process of the deep model, we interpreted what the network learns by visualizing the heatmap of feature maps. In fact, the color of the

TABLE IV
QUANTITATIVE EVALUATION RESULTS FOR ABLATION EXPERIMENTS ON LEVIR, WHU, AND GZ DATASETS

Dateset	Base	Siam	ViT	MLPPE	MB	ECA	CBAM	SAA	Evaluation metrics				
									Pre	Rec	OE	OA	$F1$
LEVIR	✓								0.8527	0.8159	0.0348	0.9652	0.8165
	✓	✓							0.9021	0.8071	0.0308	0.9692	0.8379
	✓	✓	✓						0.8621	0.8953	0.0275	0.9725	0.8683
	✓	✓		✓					0.9091	0.8806	0.0226	0.9774	0.8857
	✓	✓		✓	✓				0.9106	0.8910	0.0215	0.9785	0.8949
	✓	✓		✓	✓	✓			0.9237	0.8917	0.0209	0.9791	0.8994
	✓	✓		✓	✓		✓		0.9281	0.8991	0.0187	0.9813	0.9079
	✓	✓		✓	✓			✓	**0.9290**	**0.9037**	**0.0185**	**0.9815**	**0.9116**
WHU	✓								0.8526	0.7698	0.0346	0.9654	0.7589
	✓	✓							0.8628	0.8262	0.0246	0.9754	0.8265
	✓	✓	✓						0.8835	0.8466	0.0205	0.9795	0.8466
	✓	✓		✓					0.8909	0.8668	0.0208	0.9792	0.8636
	✓	✓		✓	✓				0.9050	0.8744	0.0174	0.9826	0.8742
	✓	✓		✓	✓	✓			0.9085	0.8776	0.0170	0.9830	0.8798
	✓	✓		✓	✓		✓		0.9143	0.8792	0.0179	0.9821	0.8817
	✓	✓		✓	✓			✓	**0.9310**	**0.8947**	**0.0147**	**0.9853**	**0.9019**
GZ	✓								0.8825	0.8385	0.0308	0.9692	0.8454
	✓	✓							0.9005	0.8919	0.0189	0.9811	0.8909
	✓	✓	✓						0.9371	0.9225	0.0104	0.9896	0.9267
	✓	✓		✓					0.9618	0.9583	0.0050	0.9950	0.9572
	✓	✓		✓	✓				0.9697	0.9636	0.0045	0.9955	0.9646
	✓	✓		✓	✓	✓			0.9696	0.9680	0.0041	0.9959	0.9672
	✓	✓		✓	✓		✓		0.9784	0.9797	0.0023	0.9977	0.9780
	✓	✓		✓	✓			✓	**0.9864**	**0.9811**	**0.0014**	**0.9986**	**0.9822**

Best values are in bold.

heatmap reflects the correlation between the specific location information and the whole image, and various colors present the degree of contribution of network for the predicted category. In Fig. 10, the red denotes higher attention values and the blue denotes lower values, where Fig. 10(a)–(c) are the bitemporal images and ground truths, respectively. By comparing and. 10(d)–(f), it can be clearly seen that the SAA and the MLPPE modules can help the proposed network focus on the truly changed targets. Thus, the LGSAA-Net can obtain more discriminative features to guide the network outputting accurate predictions.

B. Discussion on the Sensitivity Experiments of the MLPPE

As described in Section III-B, we introduced the MLPPE module to skip-path to effectively fuse low-level details and high-level semantic features and narrow the segmentation gap. Here, the patches strategy of the MLPPE module is adopted to evaluate the change detection results, in which the scale parameter of patches s plays a decisive role in improving the model performance and the model accuracy. To explore the influence of different values on the change detection results, we conducted comparative experiments on three datasets by setting different scale parameter of patches s. As the number of network layers increases, the resolution of feature maps at different layers decreases, the minimum size of patch is set to 7×7. Therefore, we set the maximum $s = 1, 2, 4, 8, 16$ at convolutional layers of encoding stage, and the setting of s at different layers are shown in Table V.

Fig. 11 presents the visual change detection results on several samples of the three datasets. It can be seen that all values of s can detect really changed regions, except for some falsely changed regions. The change detection results are more satisfactory when $s = 2, 4$. To be more specific, on LEVIR and GZ datasets, they achieve the highest values of F1 when s =4, representing an improvement of 1.14% and 0.80% compared to $s = 1$. However, it achieves the highest values of $F1$ when $s = 2$ on WHU dataset. In addition, as the value of s continues to increase, the accuracy of the model begins to decrease, since small patches in feature maps with large resolution may reduce the correlation between patches

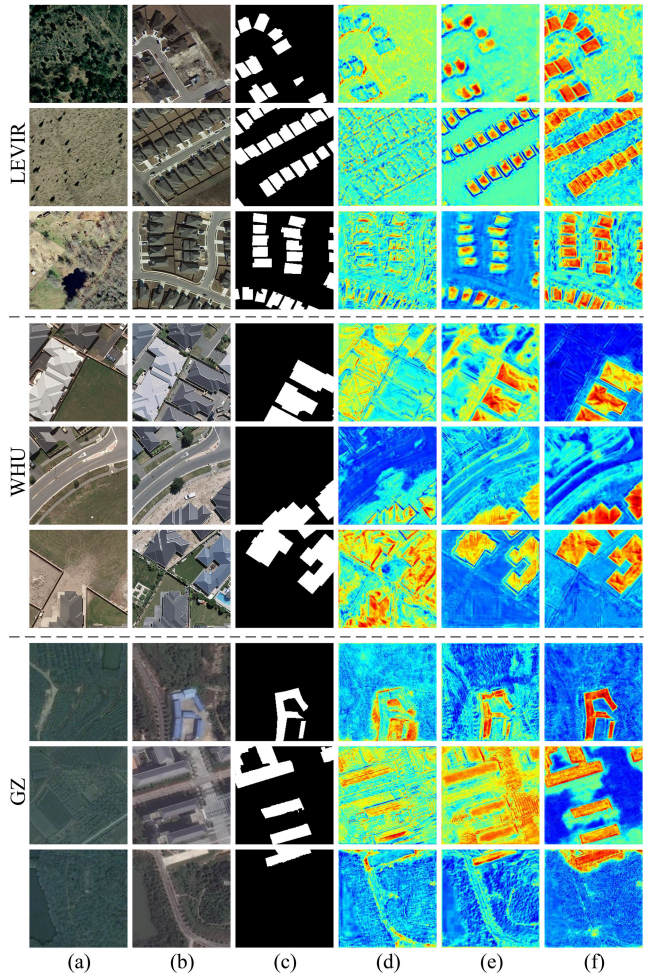


Fig. 10. Heatmap visualization of LEVIR, WHU and GZ dataset. (a) Pretemporal images. (b) Posttemporal images. (c) Ground truths. (d) Base. (e) SAA. (f) SAA+MLPPE. Red denotes higher attention values and blue denotes lower values.

with larger distances. To sum up, considering the model size and performance comprehensively, we set s to 2 in the MLPPE module.

C. Discussion on the Model Complexity

In practical applications, it is also necessary to consider factors, such as model complexity under the premise of high-precision detection results, so as to facilitate subsequent model deployment. Therefore, we evaluated the model complexity by comparing several methods with the proposed LGSAA-Net using four evaluation metrics, including floating point operations (FLOPs), parameters (Params), model size (Model), and Mean-$F1$. Mean-$F1$ denotes the average value of $F1$ on LEVIR, WHU, and GZ datasets. As shown in Fig. 12 and Table VI, the model complexity of FC-EF, FC-di, and FC-conc is relatively low, since the backbone networks of these methods are shallower than the U-Net and its variants. DFISN uses a deep supervisory strategy to achieve change detection tasks, effectively improving the change detection accuracy, but increasing model complexity. Furthermore, FCN-PP leverages the Gaussian pyramid module, so it corresponds to the larger model size, while the complexity

TABLE V
COMPARISON OF THE DIFFERENT PATCHES SCALE s IN MLPPE MODULE

s					$F1$			Model size (MB)
f1	f2	f3	f4	f5	LEVIR	WHU	GZ	
1	1	1	1	1	0.9053	0.8824	0.9769	**36.18**
2	2	2	2	2	0.9116	**0.9019**	0.9822	37.13
4	4	4	4	2	**0.9167**	0.8848	**0.9849**	39.09
8	8	8	4	2	0.9082	0.8960	0.9803	41.15
16	16	8	4	2	0.9097	0.8783	0.9784	43.26

Best values are in bold, where f1–f5 denote the first to the fifth fusion path between the encoding stage and the decoding Stage, the best values are in bold.

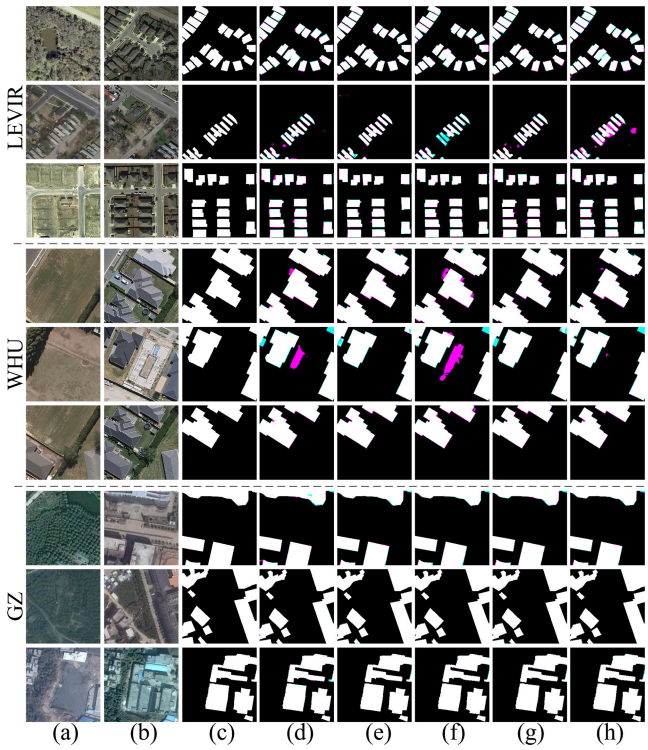


Fig. 11. (a) Pretemporal images, (b) post-temporal images, (c) ground truths, the change detection results of the maximum, (d) $s = 1$, (e) $s = 2$, (f) $s = 4$, (g) $s = 8$, and (h) $s = 16$ in the MLPPE module.

of SRCD-Net and Trans-CD is similar. SRCD-Net has a smaller model size due to small size of stacked attention module. In Trans-CD, a modified ViT module is added to U-Net, which results in a bit larger model size than U-Net. It is worth noting that FC-EF performs satisfyingly regarding FLOPs, Params, and model size. However, the Mean-$F1$ of FC-EF on three datasets is lower among these comparative methods. Finally, the proposed LGSAA-Net achieves a favorable tradeoff in model complexity and change detection accuracy relative to all comparative methods. Significantly, the proposed LGSAA-Net offers the highest Mean-$F1$ with favorable segmentation accuracy, reaching 93.19% on three datasets.

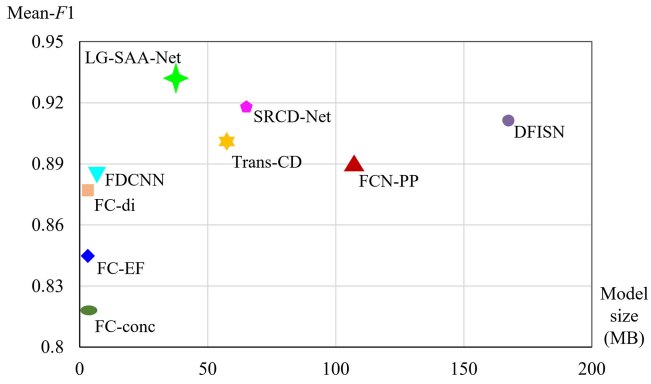


Fig. 12. Model complexity and change detection accuracy comparison of comparative methods and the proposed LGSAA-Net.

TABLE VI
QUANTITATIVE COMPARISON OF THE COMPARATIVE METHODS AND THE LGSAA-NET ON MODEL COMPLEXITY AND ACCURACY

Methods	FLOPs(GB)	Params(MB)	Model size (MB)	Mean-$F1$
FC-EF [65]	**2.63**	**0.85**	**3.34**	0.8448
FC-di [65]	3.47	0.85	3.34	0.8772
FC-conc [65]	4.07	1.07	4.09	0.8181
FCN-PP [66]	34.65	28.13	107.39	0.8895
FDCNN [67]	32.40	1.86	7.09	0.8861
DFISN [68]	112.15	43.50	166.92	0.9114
SRCD-Net [18]	27.42	16.19	64.86	0.9178
Trans-CD [19]	39.25	16.37	57.27	0.9012
LGSAA-Net	22.77	13.10	37.13	**0.9319**

Best values are in bold.

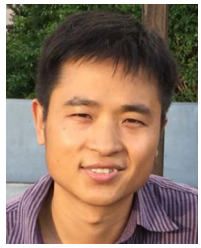
VI. CONCLUSION

In this work, we proposed the LGSAA-Net and studied change detection in bitemporal VHR remote sensing images. Different from popular change detection networks, the proposed LGSAA-Net can realize the adaptive spatial attention operation by establishing the relationships between feature maps and the convolution kernel scales. Moreover, it can effectively fuse low-level details and high-level semantics to improve feature discrimination ability by utilizing multilayer perceptron combined with patch attention mechanism. Experimental results on three change detection datasets demonstrated that the proposed LGSAA-Net can produce more accurate boundaries and high internal compactness for changed regions than state-of-the-art methods. Overall, the proposed LGSAA-Net achieves a favorable tradeoff in model complexity and change detection accuracy.

REFERENCES

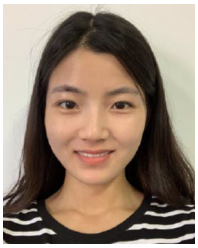
[1] A. Sebastianet al., "Temporal correlation detection using computational phase-change memory," *Nat. Commun.*, vol. 8, no. 1, 2017, Art. no. 1115.
[2] R. Saxenaet al., "Towards a polyalgorithm for land use change detection," *ISPRS J. Photogramm. Remote Sens.*, vol. 144, pp. 217–234, 2018.
[3] Y. Zheng, X. Zhang, B. Hou, and G. Liu, "Using combined difference image and k-means clustering for SAR image change detection," *IEEE Geosci. Remote Sens.*, vol. 11, no. 3, pp. 691–695, Mar. 2014.
[4] P. L. St-Charles, G. A. Bilodeau, and R. Bergevin, "SuBSENSE: A universal change detection method with local adaptive sensitivity," *IEEE Trans. Image Process.*, vol. 24, no. 1, pp. 359–373, Jan. 2015.
[5] T. Lei, D. Xue, Z. Lv, S. Li, Y. Zhang, and A. K. Nandi, "Unsupervised change detection using fast fuzzy clustering for landslide mapping from very high-resolution images," *Remote Sens.*, vol. 10, no. 9, 2018, Art. no. 1381.
[6] D. Xue, T. Lei, X. Jia, X. Wang, T. Chen, and A. K. Nandi, "Unsupervised change detection using multiscale and multiresolution gaussian-mixture-model guided by saliency enhancement," *IEEE J. Sel. Topics Appl. Earth Obs. Remote Sens.*, vol. 14, pp. 1796–1809, 2020.
[7] P. Du, X. Wang, D. Chen, S. Liu, C. Lin, and Y. Meng, "An improved change detection approach using tri-temporal logic-verified change vector analysis," *ISPRS J. Photogramm. Remote Sens.*, vol. 161, pp. 278–293, 2020.
[8] P. Lu, Y. Qin, Z. Li, A. C. Mondini, and N. Casagli, "Landslide mapping from multi-sensor data through improved change detection-based markov random field," *Remote Sens. Environ.*, vol. 231, 2019, Art. no. 111235.
[9] Z. Li, W. Shi, P. Lu, L. Yan, Q. Wang, and Z. Miao, "Landslide mapping from aerial photographs using change detection-based Markov random field," *Remote Sens. Environ.*, vol. 187, pp. 76–90, 2016.
[10] K. Tan, Y. Zhang, X. Wang, and Y. Chen, "Object-based change detection using multiple classifiers and multi-scale uncertainty analysis," *Remote Sens.*, vol. 11, no. 3, 2019, Art. no. 359.
[11] X. Zheng, X. Chen, X. Lu, and B. Sun, "Unsupervised change detection by cross-resolution difference learning," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–16, 2021.
[12] S. Wan, S. Pan, P. Zhong, X. Chang, J. Yang, and C. Gong, "Dual interactive graph convolutional networks for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–14, 2021.
[13] S. Wan, C. Gong, P. Zhong, S. Pan, G. Li, and J. Yang, "Hyperspectral image classification with context-aware dynamic graph convolutional network," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 1, pp. 597–612, Jan. 2021.
[14] S. Wan, C. Gong, P. Zhong, B. Du, L. Zhang, and J. Yang, "Multiscale dynamic graph convolutional network for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 5, pp. 3162–3177, May 2020.
[15] Y. Chen, X. Lu, and S. Wang, "Deep cross-modal image–voice retrieval in remote sensing," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 10, pp. 7049–7061, Oct. 2020.
[16] Y. Chen and X. Lu, "Deep category-level and regularized hashing with global semantic similarity learning," *IEEE Trans. Cybern.*, vol. 51, no. 12, pp. 6240–6252, Dec. 2021.
[17] Z. Lv, F. Wang, G. Cui, J. A. Benediktsson, T. Lei, and W. Sun, "Spatial-spectral attention network guided with change magnitude image for land cover change detection using remote sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–12, 2022, doi: 10.1109/TGRS.2022.3197901.
[18] M. Liu, Q. Shi, A. Marinoni, D. He, X. Liu, and L. Zhang, "Super-resolution-based change detection network with stacked attention module for images with different resolutions," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2021, Art. no. 4403718.
[19] H. Chen, Z. Qi, and Z. Shi, "Efficient transformer based method for remote sensing image change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–14, 2021, doi: 10.1109/TGRS.20 21.3095166.
[20] X. Tanget 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.
[21] H. C. Shinet al., "Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning," *IEEE Trans. Med. Imag.*, vol. 35, no. 5, pp. 1285–1298, May 2016.
[22] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput. Assist. Interv.*, 2015, pp. 234–241.
[23] H. Sun, X. Zheng, and X. Lu, "A supervised segmentation network for hyperspectral image classification," *IEEE Trans. Image Process.*, vol. 30, pp. 2810–2825, 2021.
[24] W. Chen, X. Zheng, and X. Lu, "Semisupervised spectral degradation constrained network for spectral super-resolution," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2021, Art. no. 5506205.
[25] X. Peng, R. Zhong, Z. Li, and Q. Li, "Optical remote sensing image change detection based on attention mechanism and image difference," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 9, pp. 7296–7307, Sep. 2021.
[26] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "UNet : A nested U-Net architecture for medical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Computed Assist. Intervention*, 2018, pp. 3–11.

[27] H. Huang, L. Lin, R. Tong, H. Hu, and Q. Zhang, "Unet 3: A full-scale connected UNet for medical image segmentation," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2020, pp. 1055–1059.
[28] X. Zheng, H. Sun, X. Lu, and W. Xie, "Rotation-invariant attention network for hyperspectral image classification," *IEEE Trans. Image Process.*, vol. 31, pp. 4251–4265, 2022.
[29] W. Wang and J. Shen, "Deep visual attention prediction," *IEEE Trans. Image Process.*, vol. 27, no. 5, pp. 2368–2378, May 2018.
[30] S. Fanet al., "Emotional attention: A study of image sentiment and visual attention," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7521–7531.
[31] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7132–7141.
[32] Q. Wang, B. Wu, and P. Zhu, "ECA-Net: Efficient channel attention for deep convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 11531–11539.
[33] J. Hu, L. Shen, S. Albanie, G. Sun, and A. Vedaldi, "Gather-excite: Exploiting feature context in convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2018, pp. 1–11.
[34] H. Zhaoet al., "PSANet: Point-wise spatial attention network for scene parsing," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 267–283.
[35] J. Park, S. Woo, J. Y. Lee, and I. S. Kweon, "BAM: Bottleneck attention module," in *Proc. Brit. Mach. Vis. Conf.*, 2018, pp. 1–14.
[36] S. Woo, J. Park, J. Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 3–19.
[37] Y. Cao, J. Xu, S. Lin, F. Wei, and H. Hu, "CGnet: Non-local networks meet squeeze-excitation networks and beyond," in *Proc. IEEE/CVF Int. Conf.Comput. Vis.*, 2019, pp. 1–10.
[38] A. Vaswaniet al., "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 1–11.
[39] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7794–7803.
[40] Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei, and W. Liu, "Ccnet: Criss-cross attention for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 603–612.
[41] J. Fuet al., "Dual attention network for scene segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit*, 2019, pp. 3146–3154.
[42] S. Zheng, J. Lu, H. Zhao, and X. Zhu, "Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 6881–6890.
[43] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, and X. Zhai, "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Represent.*, 2021, pp. 1–22.
[44] L. Wang, L. Wang, Q. Wang, and P. M. Atkinson, "SSA-SiamNet: Spectral-spatial-wise attention-based siamese network for hyperspectral image change detection," *IEEE Trans. Geosc. Remote Sens.*, vol. 60, pp. 1–18, 2022.
[45] Y. Liu, C. PangZ. 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.
[46] 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. Geosc. Remote Sens.*, vol. 60, 2022, Art. no. 5604816.
[47] H. Chen, Z. Qi, and Z. Shi, "Remote sensing image change detection with transformers," *IEEE Trans. Geosc. Remote Sens.*, vol. 60, 2022, Art. no. 5607514.
[48] W. G. C. Bandara and V. M. Patel, "A transformer-based Siamese network for change detection," 2022, *arXiv:2201.01293*.
[49] Y. Feng, H. Xu, J. Jiang, H. Liu, and J. Zheng, "ICIF-Net: Intra-scale cross-interaction and inter-scale feature fusion network for bitemporal remote sensing images change detection," *IEEE Trans. Geosc. Remote Sens.*, vol. 60, 2022, Art. no. 4410213.
[50] Y. Chen, D. Zhao, X. Lu, S. Xiong, and H. Wang, "Unsupervised balanced hash codes learning with multichannel feature fusion," *IEEE J. Sel. Topics Appl. Earth Obs. Remote Sens.*, vol. 15, pp. 2816–2825, 2022.
[51] Y. Chen, S. Xiong, L. Mou, and X. Zhu, "Deep quadruple-based hashing for remote sensing image-sound retrieval," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 4705814.
[52] B. Zhao, M. Gong, and X. Li, "Audio visual video summarization," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 25, pp. 1–8, 2021, doi: 10.1109/TNNLS.2021.3119969.
[53] X. Mao, C. Shen, and Y. B. Yang, "Image restoration using very deep convolutional encoder-decoder networks with symmetric skip connections," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 2810–2818.
[54] A. Karnewar and O. Wang, "MSG-GAN: Multiscale gradients for generative adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 7709–7808.
[55] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Dec. 2017.
[56] H. Seo, C. Huang, M. Bassenne, R. Xiao, and L. Xing, "Modified U-Net (mU-Net) with incorporation of object-dependent high level features for improved liver and liver-tumor segmentation in CT images," *IEEE Trans. Med. Imag..*, vol. 39, no. 5, pp. 1316–1325, May 2020.
[57] N. Ibtehaz and M. S. Rahman, "Multiresunet: Rethinking the U-net architecture for multimodal biomedical image segmentation," *Neural Netw.*, vol. 121, pp. 74–87, 2020.
[58] O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, and K. Mori, "Attention U-Net: Learning where to look for the pancreas," 2018, *arXiv:1804.03999*.
[59] Z. Wang, N. Zou, D. Shen, and S. Ji, "Non-local U-nets for biomedical image segmentation," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 6315–6322.
[60] S. Sun, L. Mu, L. Wang, and P. Liu, "L-UNet: An LSTM network for remote sensing image change detection," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2020, Art. no. 8004505.
[61] D. Peng, Y. Zhang, and H. Guan, "End-to-end change detection for high resolution satellite images using improved UNet," *Remote Sens.*, vol. 11, no. 11, 2019, Art. no. 1382.
[62] S. Fang, K. Li, J. Shao, and Z. Li, "SnuNet-CD: A densely connected Siamese network for change detection of VHR images," *IEEE Trans. Geosc. Remote Sens.*, vol. 19, 2022, Art. no. 8007805.
[63] J. Tang, C. Deng, and G. B. Huang, "Extreme learning machine for multilayer perceptron," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 4, pp. 809–821, Apr. 2016.
[64] M. Guo, J. Cai, Z. Liu, T. Jiang, R. R. Martin, and S. Hu, "PCT: Point cloud transformer," *Comput. Vis. Media.*, vol. 7, no. 2, pp. 187–199, 2021.
[65] R. C. Daudt, B. L. Saux, and A. Boulch, "Fully convolutional Siamese networks for change detection," in *Proc. IEEE Int. Conf. Image Process.*, 2018, pp. 4063–4067.
[66] T. Lei, Y. Zhang, Z. Lv, S. Li, S. Liu, and A. K. Nandi, "Landslide inventory mapping from bitemporal images using deep convolutional neural networks," *IEEE Geosc. Remote Sens. Lett.*, vol. 16, no. 6, pp. 982–986, Jun. 2019.
[67] M. Zhang and W. Shi, "A feature difference convolutional neural network-based change detection method," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 10, pp. 7232–7246, Oct. 2020.
[68] C. Zhang, P. Yue, D. Tapete, and L. Jiang, "A deeply supervised image fusion network for change detection in high resolution bi-temporal remote sensing images," *ISPRS J. Photogramm. Remote Sens.*, vol. 166, pp. 183–200, 2020.
[69] 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.
[70] S. Ji, S. Wei, and M. Lu, "Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set," *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 1, pp. 574–586, Jan. 2019.
[71] D. Peng, L. Bruzzone, Y. Zhang, H. Guan, H. Ding, and X. Huang, "SemiCDNet: A semisupervised convolutional neural network for change detection in high resolution remote-sensing images," *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 7, pp. 5891–5906, Jul. 2021.



Tao Lei (Senior Member, IEEE) received the Ph.D. degree in information and communication engineering from Northwestern Polytechnical University, Xi'an, China, in 2011.

From 2012 to 2014, he was a Postdoctoral Research Fellow with the School of Electronics and Information, Northwestern Polytechnical University, Xi'an, China. From 2015 to 2016, he was a Visiting Scholar with the Quantum Computation and Intelligent Systems group, University of Technology Sydney, Sydney, NSW, Australia. He is currently a Professor with the School of Electronic Information and Artificial Intelligence, Shaanxi University of Science and Technology. He has authored and coauthored 80+ research papers including IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE TRANSACTIONS ON FUZZY SYSTEMS, IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, etc. His research interests include image processing, pattern recognition, and machine learning.



Dinghua Xue received the M.S. degree in control science and engineering in 2019, from Shaanxi University of Science and Technology, Xi'an, China, where she is currently working toward the Ph.D. degree in light chemical industrial process systems and engineering with the School of Electrical and Control Engineering.

Her research interests include image processing and pattern recognition.

Hailong Ning received the Ph.D. degree in signal and information processing from the University of Chinese Academy of Sciences, Beijing, China, in 2021.

He is currently an Associate Professor with the School of Computer Science and Technology, Xi'an University of Posts and Telecommunications, Xi'an, China. His research interests include pattern recognition, machine learning, computer vision, and multimodal learning.

Shuangming Yang (Member, IEEE) received the M.S. and Ph.D. degrees in control science and engineering from Tianjin University, Tianjin, China, in 2016 and 2020, respectively.

He is currently an Associate Professor with the School of Electrical and Information Engineering, Tianjin University. His research interests include artificial intelligence, neuromorphic computing, neural engineering, brain-inspired computing, and machine learning.

Zhiyong Lv received the M.S. degree in geographic information system and the Ph.D. degree in cartography and geographic information system from the School of Remote Sensing and Information Engineering, Wuhan University, Wuhan, China, in 2008 and 2014, respectively.

He was an Engineer of surveying and worked with the First Institute of Photogrammetry and Remote Sensing, from 2008 to 2011. He is currently with the School of Computer Science and Engineering, Xi'An University of Technology, Xi'an, China. His research interests include multihyperspectral and high-resolution remotely sensed image processing, spatial feature extraction, neural networks, pattern recognition, deep learning, and remote sensing applications.

Asoke K. Nandi (Fellow, IEEE) received the Ph.D. degree in physics from the University of Cambridge (Trinity College), Cambridge, U.K., in 1978.

He has held academic positions with several universities, including Oxford University, Oxford, U.K., Imperial College London, London, U.K., University of Strathclyde, Glasgow, U.K., and University of Liverpool, Liverpool, U.K., as well as Finland Distinguished Professorship with Jyvaskyla University, Jyvaskyla, Finland. In 2013, he moved to Brunel University, London, U.K., to become the Chair and Head of Electronic and Computer Engineering. He is a Distinguished Visiting Professor with Xi'an Jiaotong University, Xi'an, China. In 1983, he codiscovered the three fundamental particles known as W^+, W^-, and Z^0 (by the UA1 team at CERN), providing the evidence for the unification of the electromagnetic and weak forces, for which the Nobel Committee for Physics in 1984 awarded the prize to his two team leaders for their decisive contributions. He has made many fundamental theoretical and algorithmic contributions to many aspects of signal processing and machine learning. He has much expertise in "Big and Heterogeneous Data", dealing with modeling, classification, estimation, and prediction. He has authored over 600 technical publications, including 270 journal papers as well as five books, entitled *Condition Monitoring with Vibration Signals: Compressive Sampling and Learning Algorithms for Rotating Machines* (Wiley, 2020), *Automatic Modulation Classification: Principles, Algorithms and Applications* (Wiley, 2015), *Integrative Cluster Analysis in Bioinformatics* (Wiley, 2015), *Blind Estimation Using Higher-Order Statistics* (Springer, 1999), and *Automatic Modulation Recognition of Communications Signals* (Springer, 1996). The H-index of his publications is 80 (Google Scholar) and his ERDOS number is 2. His research interests include signal processing and machine learning, with applications to communications, image segmentations, biomedical data, etc.

Prof. Nandi is a Fellow of the Royal Academy of Engineering (U.K.) as well as a Fellow of seven other institutions, including the IET. He was the recipient of the Institute of Electrical and Electronics Engineers (USA) Heinrich Hertz Award, in 2012, the Glory of Bengal Award for his outstanding achievements in scientific research, in 2010, the Water Arbitration Prize of the Institution of Mechanical Engineers (U.K.), in 1999, and the Mountbatten Premium, Division Award of the Electronics and Communications Division, of the Institution of Electrical Engineers (U.K.), in 1998. He is an IEEE Engineering in Medicine and Biology Society Distinguished Lecturer, from 2018 to 2019.