Difference Enhancement and Spatial–Spectral Nonlocal Network for Change Detection in VHR Remote Sensing Images

Tao Lei, *Senior Member, IEEE*, Jie Wang, Hailong Ning, Xingwu Wang, Dinghua Xue, Qi Wang, *Senior Member, IEEE*, and Asoke K. Nandi, *Life Fellow, IEEE*

***Abstract*— The popular Siamese convolutional neural networks (CNNs) for remote sensing (RS) image change detection (CD) often suffer from two problems. First, they either ignore the original information of bitemporal images or insufficiently utilize the difference information between bitemporal images, which leads to the low tightness of the changed objects. Second, Siamese CNNs always employ dual-branch encoders for CD, which increases computational cost. To address the above issues, this article proposes a network based on difference enhancement and spatial–spectral nonlocal (DESSN) for CD in very-high-resolution (VHR) images. This article makes threefold contributions. First, we design a difference enhancement (DE) module that can effectively learn the difference representation between foreground and background to reduce the impact of irrelevant changes on the detection results. Second, we present a spatial–spectral nonlocal (SSN) module that is different from vanilla nonlocal because multiscale spatial global features are incorporated to model the large-scale variation of objects during CD. The module can be used to strengthen the edge integrity and internal tightness of changed objects. Third, the asymmetric double convolution with Ghost (ADCG) module is exploited instead of standard convolution. The ADCG can not only refine the edge information of the changed objects, since horizontal and vertical convolutional kernels have good contour preservation advantages, but also greatly reduce the computational complexity of the proposed model. The experiments on two public VHR CD datasets demonstrate that the proposed network can provide higher detection accuracy and requires smaller memory usage than state-of-the-art networks.**

***Index Terms*— Change detection (CD), difference enhancement (DE) module, Siamese convolutional neural networks (CNNs), spatial–spectral nonlocal (SSN) module.**

Manuscript received August 23, 2021; revised October 13, 2021 and November 18, 2021; accepted December 4, 2021. Date of publication December 10, 2021; date of current version March 1, 2022. This work was supported in part by the Natural Science Basic Research Program of Shaanxi under Program 2021JC-47, in part by the National Natural Science Foundation of China under Grant 61871259 and Grant 61861024; in part by the National Natural Science Foundation of China-Royal Society, U.K., under Grant 61811530325 (IECnNSFCn170396); in part by the Key Research and Development Program of Shaanxi under Program 2021ZDLGY08-07; in part by the Shaanxi Joint Laboratory of Artificial Intelligence under Program 2020SS-03; and in part by the Special Construction Fund for Key Disciplines of Shaanxi Provincial Higher Education. *(Corresponding author: Hailong Ning.)*

Tao Lei and Xingwu Wang are with the Shaanxi Joint Laboratory of Artificial Intelligence and the School of Electronic Information and Artificial Intelligence, Shaanxi University of Science and Technology, Xi'an 710021, China (e-mail: leitao@sust.edu.cn; wangxwu1949@163.com).

Jie Wang and Dinghua Xue are with the School of Electronical and Control Engineering, Shaanxi University of Science and Technology, Xi'an 710021, China (e-mail: wcjsust@163.com; 903438920@qq.com).

Hailong Ning is with the School of Computer Science and Technology, Xi'an University of Posts and Telecommunications, Xi'an 710121, China, also with the Shaanxi Key Laboratory of Network Data Analysis and Intelligent Processing, Xi'an University of Posts and Telecommunications, Xi'an 710121, China, and also with the Xi'an Key Laboratory of Big Data and Intelligent Computing, Xi'an University of Posts and Telecommunications, Xi'an 710121, China (e-mail: ninghailong93@gmail.com).

Qi Wang is with the School of Computer Science and the School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an 710072, China (e-mail: crabwq@gmail.com).

Asoke K. Nandi is with the Department of Electronic and Electrical Engineering, Brunel University London, Uxbridge 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).

Digital Object Identifier 10.1109/TGRS.2021.3134691

I. Introduction

CHANGE detection (CD) in remote sensing (RS) images aims to identify the differences between two images from different periods but the same area. It is an important branch of RS image analysis [1] and has been widely applied in urban expansion [2], land exploration [3], disaster assessment [4], environmental monitoring [5], and so on. As the limitation of imaging technique, the early RS images have low resolution, and each pixel of images usually includes several different objects, such as trees, roads, and grass. With the continuous development of optical sensor equipment, the resolution of RS images has been greatly improved [6]. In recent years, high-resolution (HR), especially very-high-resolution (VHR), RS images have become more pervasive, which makes RS image CD more challenging. Under this circumstance, more and more researchers are devoted to solving the problem of CD for VHR RS images.

Before using deep learning (DL) for CD, transformation-based and image algebra approaches, such as the principal component analysis (PCA) [7], the independent component analysis (ICA) [8], Gabor [9], multivariate alteration detection (MAD) [10], and change vector analysis (CVA) [11], are often employed to achieve CD. The main idea of these methods is first to obtain difference images (DIs), then performs threshold- or clustering-based pixel classification on DIs to extract change features or obtain the change image by maximizing the difference. Since these methods only extract the spectral information of images while ignoring the context relationship, they are only suitable for low- and medium-resolution images [12]. For HR and VHR images, as the texture of ground objects is more abundant and the heterogeneity within a class is enhanced [13], many methods are designed to group pixels into objects, and then, these objects are considered as units

to determine whether they are changed or not according to their color, shape, and neighborhood information [14], [15]. Although these methods improve the CD effect of VHR RS images to some extent, they are sensitive to noise and provide low detection accuracy with high computational cost since they only employ handcrafted features and require a complex classifier.

In recent years, with the rapid development of DL technology [16], [17], some deep neural networks, such as deep belief network (DBN) [18], autoencoder (AE) [19], and deep convolutional neural networks (CNNs) [20], are used for RS image CD since they can automatically learn the abstract and multilevel features of complex ground objects and demonstrate much robustness to noise. Although these models show better detection performance than traditional methods mentioned above, they remain to suffer from some problems, such as propagation errors and high computational complexity. In order to overcome these shortcomings and learn guided information from the labeled images, the fully CNNs (FCNs) [21] are designed to improve image segmentation [22]–[24]. Although the FCNs do not require full connection, most of them easily cause the information loss of bitemporal images since they mainly adopt a single-branch network. For this, researchers applied the dual-branch network introduced by Siamese [25] to RS image CD [26], which greatly improves the detection accuracy due to the utilization of richer image features from bitemporal images. Subsequently, the idea has been widely adopted and has become the baseline network in CD tasks [27]–[34].

Based on the previous analysis, although many DL models have been presented and used for RS image CD, they remain to suffer from some problems. First, the existing methods cannot effectively construct the relationship between bitemporal images, resulting in the adverse impact of irrelevant changes on the detection results. Second, the integrity of the boundary and internal tightness of changed objects are not fully considered, leading to edge information loss in the predicted change maps. Third, the existing dual-branch networks expand the size of models, increasing the computational cost and easily causing the problem of overfitting.

To address the above problems, this article proposes a novel network based on difference enhancement and spatial–spectral nonlocal (DESSN) for VHR RS image CD. The proposed method is based on Siamese [25] and U-Net [35], which is lightweight and composed of three main modules: the difference enhancement (DE) module, the spatial–spectral nonlocal (SSN) module, and the asymmetric double convolution with Ghost (ADCG) module.

To construct the relationship between bitemporal images, the DE module is designed to make the network focus more attention on changed areas, thereby suppressing irrelevant changes caused by noise and better distinguishing the objects from the background. Although the attention mechanism has been widely used in the design of CNNs, it is rare in Siamese CNNs for the task of CD. As a DI only contains changed and unchanged areas, it is easy to distinguish them. However, in practice, it is difficult since a DI usually contains some uncertain areas due to some disturbance from noise and light. The attention mechanism is very useful for finding significant features that can distinguish objects and background. Thus, we design the DE module with an attention mechanism to explore intrinsic features that can identify really changed areas.

To strengthen the integrity of the boundary and internal tightness of changed objects, the SSN module is designed and used during the stage of feature fusion. On the one hand, the proposed SSN is different from the regular spatial–spectral feature fusion since the former can provide long-range correlation. On the other hand, it is also different from the regular nonlocal module since it integrates multiscale spatial information into the SSN that achieves better feature representation for classifying and locating changed objects.

To reduce the number of network parameters, the ADCG module is designed to replace the vanilla convolution. Although the asymmetric convolution can reduce the number of parameters to some extent, it may lead to the slight degeneration of network performance. Adding asymmetric convolution to vanilla convolution can enhance feature learning ability, but the ability to reduce parameters is limited. The presented ADCG module combines the advantages of both asymmetric convolution and Ghost. It is not only able to reduce the number of parameters due to the utilization of fewer convolutional kernels but also improve feature representation due to the effect of asymmetric convolution that can provide additional information for changed objects.

To sum up, the main contributions of this article include the following.

1) A DE module is designed to reduce the impact of irrelevant changes on the detection results by effectively learning the difference representation between foreground and background.
2) An SSN module is introduced to strengthen the edge integrity and internal tightness of changed objects by learning long-range correlation.
3) An ADCG module is proposed to refine boundary information of changed objects and reduce the number of the network parameters more effectively.

The rest of this article is organized as follows. The related works are reviewed in Section II. A detailed description of the proposed method is provided in Section III. The experimental results and discussion of key issues are reported in Section IV. The summary and conclusion are drawn in Section V.

II. Related Work

In this section, we present the related work that mainly includes the RS image CD methods and attention neural networks.

A. RS Image Change Detection Methods

In the early stage of CD research, since RS images are difficult to be collected and their resolution is low, most studies adopted unsupervised methods to deal with CD. Traditional unsupervised methods generally follow three main steps: preprocessing, generating DIs, detecting, and analyzing the obtained DIs to get the CD results. The last step is the core part of the CD. Detection and analysis are usually based

on threshold or clustering algorithms to identify changes. The threshold-based algorithms are more sensitive to selected thresholds that are subjective, with empirical errors and poor robustness [36], [37]. As a result, clustering-based algorithms tended to be more popular for the task [38], [12].

Traditional clustering-based algorithms, such as c-means clustering and fuzzy c-means clustering (FCM) algorithm, tend to ignore the image spatial information and are sensitive to noise defects [12]. In consequence, Gong *et al*. [39] proposed a fuzzy clustering algorithm based on the Markov random field (MRF) energy function. The algorithm improves the detection accuracy to a certain extent but increases the computational complexity. In order to improve CD accuracy and efficiency, Ghosh *et al*. [12] proposed a fast FCM unsupervised CD algorithm for VHR RS images. Furthermore, Gong *et al*. [39] proposed a multiscale and multiresolution Gaussian-mixture model guided by saliency-enhancement. The above methods all rely on the generation of DIs that inevitably ignore the meaningful information from original bitemporal images and are sensitive to noise, making it difficult to distinguish effectively the foreground from the background. In addition, all clustering-based algorithms need to set a different number of clustering for different datasets leading to poor robustness.

With the improvement of the RS imaging technique, it is difficult for unsupervised methods to suppress the impact of irrelevant changes on detection results [42]. Fortunately, benefiting from recent advances in computer vision [43]–[47] and the emergence of abundant labeled data, many supervised DL-based methods are reported and used for RS image CD [21]–[23], [48], [49]. Nevertheless, these methods directly learn image features from DIs, which ignores some meaningful information contained in original bitemporal images. It is clear that these methods do not consider the specific task requirement of CD and only employ general deep network models to solve the task leading to the limited accuracy of CD. In fact, RS image CD is different from other classification tasks since its input consists of a pair of bitemporal images: the pretemporal image and the posttemporal image. Inspired by this, Bromley *et al*. [50] first extracted features of each temporal image and then combined and compared the extracted features in the subsequent network layer to generate a change map, which was more conducive to realize high-precision CD. Based on the insight, the Siamese network [25] is proposed and used for CD tasks.

The Siamese network is a neural network framework consisting of two branches with sharing weights, which was first proposed to solve the problem of image matching. Specifically, it maps the respective input to a new space to form a new representation, then obtains the difference between two images by measuring distance (such as the Euclidean distance and the Mahalanobis distance), and, finally, outputs the similarity of the two input images by calculating loss. The two-channel network [51] inspired by the Siamese network directly learns relationships by extracting features on the channel fusion maps of two inputs. Daudt *et al*. [26] first applied the Siamese network and the two-channel network to CD tasks. They designed three networks training in an end-to-end manner and compared it with some popular methods to demonstrate the effectiveness of the Siamese network. Since then, Siamese networks have been widely utilized as part of feature extraction for CD [27], [28]. In order to improve the detection accuracy further, some methods introduce long-short term memory (LSTM) networks or recurrent neural networks (RNNs) on this basis to explore spatial–temporal relationships [29]–[31], [34], and some methods introduce attention mechanisms to exploit the importance of difference feature maps and spatial positions for improving the CD effect [32], [33], [52], [53].

The Siamese network-based methods mentioned above either use the DIs for skip-connection or directly perform the change analysis on DIs. Since the process of feature extraction of each branch is independent, the difference information is neglected, resulting in blurred changed areas and serious adhesion. In addition, these methods often require much memory usage due to a large number of parameters, which leads to the difficulty of deploying networks on mobile devices.

B. Nonlocal Neural Networks

It is a fact that the receptive field of small convolution kernels, such as 3×3 and 5×5, is too narrow. Although the global information of images can be obtained by stacking more convolution layers, this manner may lead to training difficulty and more complex models. The attention mechanism can capture the overall relationship effectively and focus limited energy on important positions from a global perspective [54], [55], which can reduce resource consumption and obtain more useful information [56]. Exploring the global position and feature relationship of pixels in CD can better and fast determine the changed and unchanged attributes.

Unlike previous spatial attention operations, the squeeze-and-excitation network (SENet) [57] pays more attention to the relationship between channels. The squeeze-and-excitation (SE) module first performs the squeeze operation on feature maps to obtain the channel-level global features. Then, the excitation operation is performed on the global features to learn the relationship among features and obtain the weights of different channels. Finally, the original feature maps and the channel weights are multiplied to obtain the final weighted features. The convolutional block attention module (CBAM) [58] further develops SENet by combining channel attention and spatial attention in a sequential manner. It introduces the branch of max-pooling based on the SE module and then combines the two branches with elementwise summation before performing sigmoid activation. Spatial attention utilizes average-pooling and max-pooling along the channel axis on feature maps and then concatenates the two pooling results before convolution and sigmoid activation operations. Zhang *et al*. [32] applied the idea of CBAM in RS image CD. They fused the spatial attention and the channel attention in series to reconstruct the DIs, thereby enhancing the internal tightness of changed objects.

The main idea of attention mentioned above is biased toward the distribution of weights, while the nonlocal module [59] that is an application of self-attention [60] not only focuses on the weight distribution but also strengthens the connection of context. It can solve the problem of long-distance

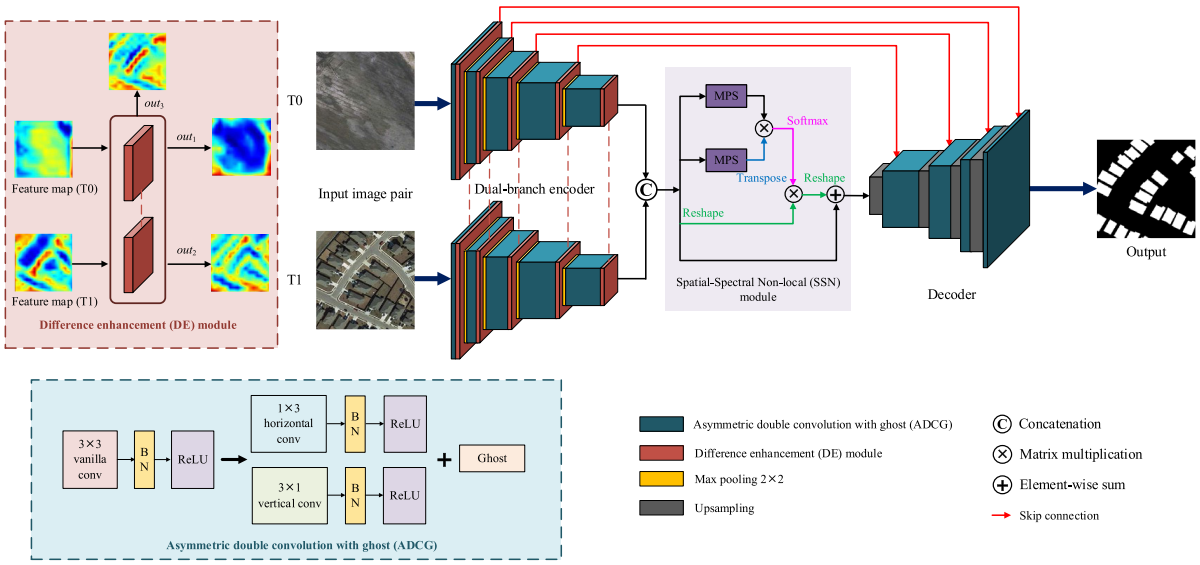


Fig. 1. Framework of the proposed network. The ADCG module aims to refine the extracted features and reduce computations. The DE module is purposed to make full use of differences to emphasize the changed features. The SSN module introduces a multiscale parallel sampling (MPS) module to strengthen the edge integrity and internal tightness of changed objects. out_1, out_2, and out_3 indicate the three outputs of the DE module. The details of the three modules are described in Section III.

information transmission and improve long-distance dependence. Specifically, the input is divided into three branches from top to bottom by three different 1×1 convolutions. Then, matrix multiplication is performed between the first two branches, and a softmax layer is applied to calculate the spatial attention map. Finally, the third branch and the spatial attention map are multiplied to obtain a weighted map. Chen and Shi [33] applied this idea for RS image CD. They utilized the self-attention module to calculate the attention weights between any two pixels at different times and positions to generate more discriminative features and, thus, achieved better detection results.

To better model semantic interdependencies in both spatial and channel dimensions, the dual-attention network (DANet) [61] combines the ideas of the nonlocal module and CBAM. It utilizes a parallel method to perform spatial-wise self-attention and channelwise self-attention on the deep feature maps. Compared with CBAM, the nonlocal module directly calculates the autocorrelation of features, which avoids manual design of pooling layer, multilayer perceptron, and other complicated operations in CBAM. The nonlocal module can quickly obtain the global relationship, but the computational cost and memory consumption are high. In addition, its efficiency is low when the input image has a high resolution. Although the operations of channel dimensionality reduction and pooling have been utilized to solve these problems, it is still hard to find a balance between model compression and high precision.

III. Methods

A. Overview

In this section, we propose a network based on Siamese and U-Net to enhance change representation and squeeze model for CD. The framework of the proposed method is shown in Fig. 1, which consists of a dual-branch encoder and decoder. The network involves three main modules: the ADCG module, the DE module, and the SSN module. A pair of bitemporal images are fed into the weight-sharing dual-branch network composed of ADCG to perform feature extraction, respectively. After each layer of convolution operation, the DE module is used to enhance the features of changed objects. The SSN module aims to reduce the redundancy of the dual-branch fusion features and establish long-term correlation. The result of the SSN module is utilized for upsampling. Finally, the dimensionality reduction and normalization operations are performed to output the final change map, in which each element is binarized to 0 or 1 according to a predefined threshold. The binarization process is defined as

$$Y_{i,j} = \begin{cases} 0, & 0 \le P_{i,j} \le T \\ 1, & T < P_{i,j} \le 1 \end{cases} \tag{1}$$

where $Y_{i,j}$ and $P_{i,j}$ stand for the value of the ith row and the jth column before and after binarization, respectively. T denotes the predefined threshold, which is set to 0.5 in our experiments.

B. Difference Enhancement Module

In general, the DIs are directly computed by performing subtractions between posttemporal images and pretemporal images and then used for feature learning. However, there exists severe noise in the directly obtained DIs, which is not good for detecting changed objects. In order to reduce the noise and learn high-quality DIs, this article proposes the DE module drawing lessons from SA-Gate [62]. Different from the SA-Gate, the proposed DE module aims at learning the difference information rather than the complementary information for CD by fully taking advantage of bitemporal images. The architecture of the DE module is shown in Fig. 2. It is introduced to filter out the irrelevant changes and concentrate on the really changed objects. In the beginning, the original

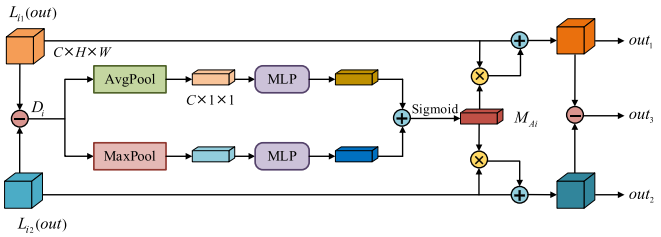


Fig. 2. Structure of the DE module, where "−" implies the difference operation, "+" implies the elementwise summation, and "×" implies the multiplication operation.

bitemporal images are fed into the encoder for learning the semantic features as the inputs of the DE module, where the encoder adopts five feature extraction layers. Let $L_{i1}(\text{out})$ and $L_{i2}(\text{out})$ represent the output bitemporal feature maps of the ith layer from the encoder, where $i = 1, 2, 3, 4$, and 5 indexes the feature extraction layers. First, the difference feature map (D) is learned by performing subtractions between the bitemporal feature maps

$$D_i = |L_{i1}(\text{out}) - L_{i2}(\text{out})| \tag{2}$$

where $|\cdot|$ denotes the absolute value operation to ensure that the obtained difference feature is meaningful.

Second, to exploit effectively the interchannel relationship of the obtained difference features, a channel attention operation is conducted [58]. Specifically, the spatial dimension of the difference feature is squeezed by the average-pooling and max-pooling simultaneously to generate different spatial context descriptors. Then, the spatial context descriptors are forwarded to a shared multilayer perceptron network and successively merged using an elementwise summation operation. Formulaically, the channel attention operation is represented as

$$M_{Ai} = \sigma(\text{MLP}(\text{AvgPool}(D_i)) + \text{MLP}(\text{MaxPool}(D_i))) \tag{3}$$

where M_{Ai} denotes the obtained attention map after channel attention operation, MLP represents the multilayer perceptron network, AvgPool and MaxPool denote the average-pooling and max-pooling operation, respectively, and σ stands for the Sigmoid function.

Finally, to obtain the enhanced map out_n, the previous features of different phases are multiplied with the attention map, respectively. The equation of generating the enhanced feature is defined as

$$\text{out}_n = M_{Ai} \times L_{\text{in}}(\text{out}) + L_{\text{in}}(\text{out}) \tag{4}$$

where $n = 1, 2$ indexes the dual-branch path and $L_{\text{in}}(\text{out})$ represents the output of the ith feature extraction layer on the nth branch of the encoder.

After obtaining the enhanced features out_1 and out_2, the enhanced difference features out_3 generated by subtracting out_1 and out_2 are considered as the third output for skip-connection to provide richer detail information so as to resolve the erroneous attention. It is worth noting that we also introduce the idea of residuals here, adding the enhanced feature map to the original feature map, so that the final representation reflecting the remote context can be obtained and the gradient can be prevented from vanishing. To the best of our knowledge, this article is the first work leveraging the attention mechanism on the DIs to capture more detailed position information of the changed objects and mapping the attention weights of DIs back to the bitemporal feature maps. It has a guiding effect on the feature extraction of the next layer. To validate the effectiveness of the proposed DE module, the comparative experiments are performed in Section IV.

C. Spatial–Spectral Nonlocal Module

Due to the limitation of the imaging technique, the details of objects are unclear in VHR RS images. However, detailed information often plays important role in the CD of RS images [33]. To this end, the nonlocal mechanism [59] is introduced into CD to improve the detection accuracy by capturing the long-range correlation of pixels [33]. Nevertheless, it still suffers from many difficulties due to the large-scale variation of the changed objects. Specifically, the edge information is lost in the difference maps since the integrity of the boundary and internal tightness of the changed objects are insufficiently considered. To address this issue, the concatenation of feature maps for upsampling provides an insight to retain all useful features. In this way, the features are mostly similar except for the changed areas after concatenation, which results in a lot of feature redundancy [63]. As a result, this article proposes the SSN module with MPS module to suppress redundant information and, meanwhile, strengthen the edge integrity and internal tightness of changed objects, which can further reflect the spectral information well and improve the CD of large-scale variation.

The architecture of the SSN module is shown in Fig. 3. In the SSN module, the input is divided into three branches Q, K, and V, where $\{Q, K, V\} \in \mathbb{R}^{C\times H\times W}$. The feature maps are sampled through the MPS module. Since the purpose of the SSN module is to obtain the global relationship among features, we choose the average-pooling that can represent the global relationship, rather than max-pooling that only emphasizes local features of images. The sampling process is described as

$$M_{Sn}(x) = \text{AvgPool}_n(x) \tag{5}$$

where $x \in \mathbb{R}^{C\times H\times W}$ represents the image to be sampled, n indicates the scale of pooling, and $M_{Sn} \in \mathbb{R}^{C\times (H/n)\times (W/n)}$ stands for the sampled map.

First, the MPS module employs four scales ($n = 16, 8, 4, 2$) for parallel sampling on branches Q and K and obtains four feature maps of different scales denoting the global spatial information, respectively. Then, each obtained feature map is reshaped to $\mathbb{R}^{C\times Z}$, where $Z = (H/n)\times(W/n)$ is the number of pixels. MPS$(\cdot)$ is the function of MPS module. The result from MPS is the concatenation of these four reshaped maps in the second dimension, which is defined as

$$\text{MPS}(x) = R(M_{S16}(x)); R(M_{S8}(x)); R(M_{S4}(x)); R(M_{S2}(x)) \tag{6}$$

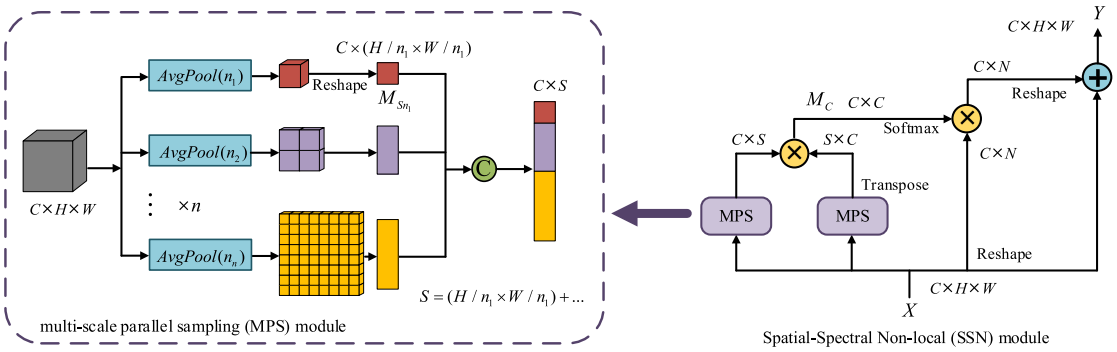


Fig. 3. Architecture of the SSN module. (Left) MPS is the multiscale parallel sampling module described in detail.

where $R(\cdot)$ denotes the operation of reshape and ";" means the concatenate operation. The size of the output from MPS module is $C \times S$, where $S = \sum_{n\in\{16,\ 8,\ 4,\ 2\}} (H/n) \times (W/n)$ is the sum of pixels of all sampled maps.

Second, to acquire the interrelationship of the obtained global features, the two results of MPS acting on Q and K are multiplied

$$M_C = \mathrm{MPS}(Q) \times \mathrm{MPS}(K)^T \tag{7}$$

where the result from $\mathrm{MPS}(K)$ is transposed and $M_C \in \mathbb{R}^{C\times C}$ is the channel attention matrix.

Finally, to obtain the enhanced feature map Y, the third branch V reshaped to $\mathbb{R}^{C\times N}$ is multiplied with M_C. The equation of generating the enhanced feature map is defined as

$$Y = R[\sigma(M_C) \times V] + X \tag{8}$$

where $R(\cdot)$ is used to reshape $\mathbb{R}^{C\times N}$ to $\mathbb{R}^{C\times H\times W}$, σ represents the softmax function, X is the original input of SSN, and Y is the final output.

The MPS module incorporates the global feature information from spatial and spectral dimensions. Unlike the previous parallel or series ways of spatial and spectral fusions, it embeds multiscale spatial information into channel attention, which is not only more conducive to the establishment of relationships to improve the performance of the network but also greatly reduces the numbers of parameters. Compared to directly using single-channel attention, it reduces the computations. Suppose that the size of the feature map is Channel $\times$ Height $\times$ Width ($C \times H \times W$), and the time complexity involved in the matrix multiplication of the original channel attention is $\mathcal{O}(C^2N)$ ($N = H \times W$, represents the number of all pixels in each channel). The time complexity involved in SSN is $\mathcal{O}(C^2S)$, which is only the S/N times of original channel attention, where S is smaller than N.

To validate the effectiveness of the proposed SSN module, the comparative experiment can be seen in our experiments.

D. Asymmetric Double Convolution With Ghost Module

Most methods in the field of computer vision achieve high precision based on the deep and large models [32], [33]. However, it is inadvisable to simply pursue high precision while ignoring the computational cost, especially for RS image CD with multibranch networks. As a result, the model compression is necessary to be introduced into RS image CD. To this end, many methods have emerged, such as asymmetric convolution [64], depthwise separable convolution [65], pruning [66], and Ghost module [67]. Inspired by the good contour preservation advantage of horizontal and vertical convolutions and strong robustness to the data distorted by rotation or flipping [68], asymmetric convolutions is introduced for CD in this work.

Considering the computations and performance, this article proposes the asymmetric double convolution (ADC) on the basis of the double convolution (DC, two cascaded 3×3 convolutions) in U-Net [35]. We replace the second 3×3 convolution in DC with 1×3 and 3×1 convolutions in parallel instead of the cascade. It should be noted that, if the two 3×3 convolutions of each group are replaced with 1×3 and 3×1 convolutions, the importance of other parts will be ignored, and the closeness between features will be reduced [68]. As a result, we only replace the second one to refine the boundary of the result from the first square convolution. The reason for parallel replacement is that factorizing a $n \times n$ convolution into a $1 \times n$ convolution followed by a $n \times 1$ convolution does not work well on early layers in practice [69]. It is clear that the proposed ADC can reduce the number of parameters due to $(1 \times n + n \times 1) \leq n \times n$, $n \geq 3$. In our experiments, all the DCs in the original U-Net will be replaced with ADCs.

A well-trained deep neural network usually can generate rich feature maps to ensure a comprehensive understanding of the input data. Han *et al.* [67] pointed out that there are many redundancies in these rich feature maps. Accordingly, they proposed the Ghost module that aims to generate more feature maps through cheap operations. Specifically, a series of linear transformations are applied to generate many Ghost features maps that can dig out the required information from original features at a small cost. In order to realize further the compression of the model, we combine the idea of the Ghost module and ADC to present a novel module named ADCG. The structure of ADCG is shown in Fig. 4. Assuming that each channel is linearly mapped M times, the final parameters and calculation amount are only $1/M$ compared to the vanilla convolution.

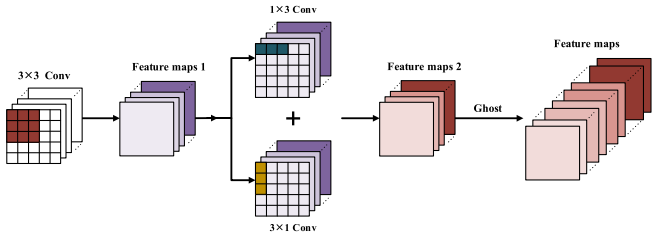


Fig. 4. Architecture of the ADCG module. The same color in feature maps indicates similar features.

We have illustrated in detail the proposed three modules: DE, SSN, and ADCG. They are embedded into the proposed framework, as shown in Fig. 1; they can help our network to achieve better feature representation for CD in VHR RS images and consumes lower memory usage and computational cost. We will demonstrate the effectiveness of the efficiency of the proposed network in our experiments.

IV. Experiments

The experiments are conducted on two public VHR CD datasets, LEVIR-CD[1] [33] and CDD[2] [70]. Eight state-of-the-art networks are considered as comparative methods to show the superiority of the proposed network. In addition, the efficiency comparison is completed to verify the advantage of our method in compressing a model while obtaining high performance. Furthermore, ablation studies are performed to verify the effectiveness of each module. Finally, the discussions about the effectiveness of the DE module and the SSN module, and the optimal selection of MPS module scales in the SSN module are presented.

A. Experimental Setup

1) Datasets: Two popular datasets are employed in our experiments, including LEVIR-CD [33] and CDD [70] datasets. The LEVIR-CD dataset contains 637 VHR Google Earth image pairs with a resolution of 0.5 m and a size of 1024 × 1024 pixels. The collection time span is five to 14 years, covering various types of buildings, such as villas, high-rise apartments, small garages, and large warehouses. The fully annotated LEVIR-CD contains a total of 31 333 individual changed building instances. In our experiments, we use 70% of the data as the training set, 10% as the validation set, and 20% as the testing set. They are cropped into 256 × 256 image pairs by random cropping. In order to enhance the diversity of the data and prevent overfitting, we perform the necessary data enhancement operations, including a rotation at a random angle and random flipping. The CDD dataset includes RS images with seasonal changes in the same region obtained by Google Earth, which are marked with the changes of buildings and vehicles, and the spatial resolution of the obtained images ranges from 0.03 to 1 m. A total of 16 000 image pairs each with a size of 256 × 256 are obtained through random cropping and data enhancement, where 10 000 pairs are used for training, 3000 pairs are adopted for verifying, and the remaining 3000 pairs are applied for testing in [70]. For a fair comparison, all the methods adopt the same data partitioning setting.

2) Evaluation Metrics: Three commonly used evaluation metrics are adopted for making an overall evaluation of the experimental results, including precision (Pre), recall (Rec), and F1-score (F1). The Pre shows how many pixels classified as true are actually true, which can be defined as

$$\text{Pre} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{9}$$

where false negative (FN) indicates the number of pixels that are incorrectly classified as unchanged; false positive (FP) indicates the number of pixels that are incorrectly classified as changed; true negative (TN) indicates the number of pixels that are correctly classified as unchanged; and true positive (TP) is the number of pixels that are correctly classified as changed. The Rec indicates how many true pixels are correctly classified as true, which can be calculated as

$$\text{Rec} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \tag{10}$$

F1 is known as the harmonic mean between the precision and recall values, and it is defined as

$$\text{F1} = 2 \times \frac{\text{Pre} \times \text{Rec}}{\text{Pre} + \text{Rec}}. \tag{11}$$

3) Loss Function: RS image CD can essentially be regarded as pixel-level classification, which is divided into two types: changed (represented by 1) and unchanged (represented by 0) pixels. The final output of the network is a binary image. To train the network, the BCELoss is adopted in this work

$$\mathcal{L} = -\frac{1}{N}\sum_{i} w_i \left[y_i \times \log x_i + (1 - y_i) \times (1 - x_i) \right] \tag{12}$$

where N denotes the number of all training samples, y_i indicates the label of the ith sample, x_i stands for the predicted value of the ith sample, and w_i denotes the weight added to the loss of the ith sample.

4) Implementation Details: We implemented the proposed method with the Pytorch framework and trained it on a GeForce RTX 3090Ti GPU with 24-GB VRAM. The number of training epoch is set to 200. The Adam gradient descend with momentum is used to optimize the model [71]. During training, the learning rate is set to 0.0001, and the batch size is set to 20. We set the hyperparameters of Ghost to be the same as the original paper [67]; the multiple of compression was set to 2, and the kernel of linear operation was set to 3.

B. Comparison Experiments

1) Comparison Methods: To demonstrate the superiority of the proposed method for RS image CD, eight state-of-the-art methods are adopted for comparison, including Siam-UNet, FC-EF [26], FC-Siam-conc [26], FC-Siam-diff [26], W-Net [72], FCN-PP [24], DSIFN [32], and FDCNN [73]. The Siam-UNet is modified based on Siamese [25] and U-Net [35] and is employed as the backbone of the proposed network. The FC-EF [26] adopts the pipeline

[1]https://justchenhao.github.io/LEVIR/

[2]https://drive.google.com/file/d/1GX656JqqOyBi_Ef0w65kDGVto-nHrNs9

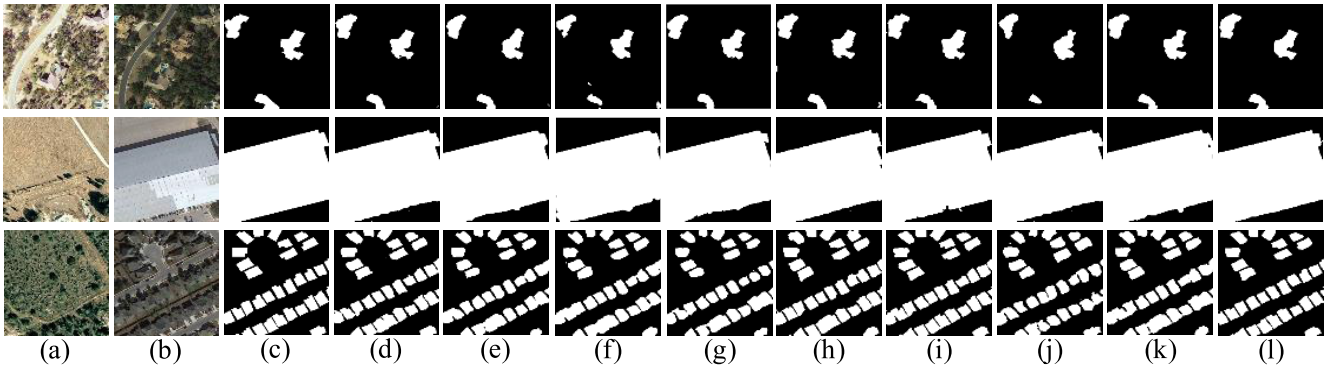


Fig. 5. Visual comparison of results of three different situations: sparse changes (row 1), large changes (row 2), and dense changes (row 3). (a) Pretemporal image. (b) Posttemporal image. (c) Label. (d) U-Net. (e) FC-EF. (f) FC-Siam-conc. (g) FC-Siam-diff. (h) W-Net. (i) DSIFN. (j) FCN-PP. (k) FDCNN. (l) Ours.

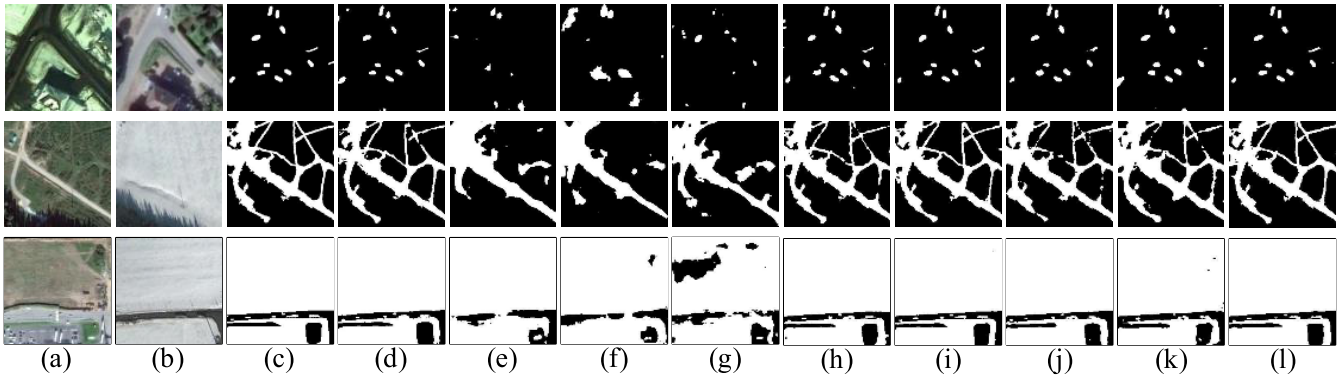


Fig. 6. Visual comparison of results of three different situations: sparse small changes (row 1), complex changes (row 2), and large changes (row 3). (a) Pretemporal image. (b) Posttemporal image. (c) Label. (d) U-Net. (e) FC-EF. (f) FC-Siam-conc. (g) FC-Siam-diff. (h) W-Net. (i) DSIFN. (j) FCN-PP. (k) FDCNN. (l) Ours.

that the bitemporal images are fused in the early stage and then fed into a U-shaped architecture with skip-connection. The FC-Siam-conc [26] is a U-shaped dual branch architecture, which concatenates the feature maps of each layer in each branch for skip-connection. The FC-Siam-diff [26] adopts the network architecture similar to FC-Siam-conc, where the difference is that FC-Siam-diff utilizes the difference of feature maps instead of the concatenation in the skip-connection. The W-Net [72] changes the pooling to convolution with a stride of 2 to avoid too much information loss and merges features from dual networks for skip-connection. To overcome the limitations of traditional global pooling, the FCN-PP [24] applies pyramid pooling in the network to expand the receptive field. The DSIFN [32] introduces a depth supervision difference recognition network for CD and uses attention on multilevel fusion features to reconstruct the change maps. The FDCNN [73] aims to generate multiscale and multidepth feature difference maps that are useful for improving CD results.

2) Comparison on the LEVIR-CD Dataset: The visual analysis of experimental results on the LEVIR-CD dataset is shown in Fig.5. The quantitative analysis is presented in Table I, where the best values are in bold. There are error detection areas in the situation of sparse CDs, such as the first row in Fig. 5(d)–(f) and (h). Due to the influence of the sunlight angle and the degree of tree coverage, the results of these methods in some areas are all relatively ambiguous, but, in general, our network is quite advantageous in boundary accuracy. The detection of large buildings mainly depends on whether the boundary is smooth and whether there are missing and wrong detection areas. These problems to varying degrees are exhibited in the second row of Fig. 5(d)–(k), and our network shows the best performance. In Fig.5 (row 3), when the changed buildings are concentrated, there will be adhesions in (d)–(k) and missing detections in (d)–(f), (j), and (k). Our network achieves the best performance in the integrity and firmness of changed objects. To illustrate further the superiority of our method, the quantitative evaluation results are reported in Table I. We can see that our network obtains the best result, which is 3.53% higher than the highest result in comparative methods. This demonstrates that our network can effectively improve the performance of VHR RS image CD by refining the edges and strengthening the integrity and internal compactness of changed objects.

3) Comparison on the CDD Dataset: The visual analysis of experimental results on the CDD dataset is shown in Fig. 6, and the quantitative analysis is shown in Table II. Since the changed objects are small and sparse, detection results are more susceptible to noise leading to some error detections. Compared with other networks, our network [see Fig. 6 (row 1) (l)] provides fewer areas of misdetection and omission. Images in the second row of Fig. 6 show very complex changes, which involves the changes of rugged paths. If the feature

TABLE I
QUANTITATIVE ANALYSIS OF LEVIR-CD DATASET CD RESULTS. THE BEST VALUES ARE IN BOLD

Methods	Pre (%)	Rec (%)	F1 (%)
Siam-UNet	84.75	89.60	87.11
FC-EF	82.27	86.18	84.18
FC-Siam-conc	76.17	86.90	81.18
FC-Siam-diff	84.44	86.38	85.40
W-Net	81.01	87.02	83.91
FCN-PP	82.09	84.48	83.27
DSIFN	86.00	89.73	87.83
FDCNN	83.87	87.56	85.68
ours	**90.99**	**91.73**	**91.36**

TABLE II
QUANTITATIVE ANALYSIS OF CDD DATASET CD RESULTS. THE BEST VALUES ARE IN BOLD

Methods	Pre (%)	Rec (%)	F1 (%)
Siam-UNet	88.31	85.30	86.78
FC-EF	72.68	58.56	64.86
FC-Siam-conc	64.17	47.70	54.72
FC-Siam-diff	76.63	57.09	65.43
W-Net	88.01	86.26	87.13
FCN-PP	88.14	84.22	86.14
DSIFN	90.72	86.50	88.56
FDCNN	87.90	86.99	87.44
ours	**95.04**	**88.77**	**91.80**

representation and classification are not so good, there will be obvious missing detections in Fig. 6(e)–(g) and detection area discontinuity in Fig. 6(h), (j), and (k). Compared with other networks, our network deals with the edge of such irregular changes more closely to the effect of the label. Images in the third row of Fig. 6 show the detection results of large changed objects. These objects may have different reflectivity to sunlight in different parts, leading to incomplete detection results, as shown in Fig. 6(e)–(g). For the detection in the lower right corner, our network also achieves a more complete and accurate detection. To verify further the superiority of our network, the quantitative evaluation results are reported in Table II, and the best values are in bold. Our network gets the highest F1, 3.24% higher than the best comparison method, because it makes full use of the long-range dependencies of context information.

4) Comparison of Efficiency: The purposes of this article are to achieve high-precision detection and a small model size. Therefore, we analyze our network and the comparison networks from multiple perspectives, including the floating-point operations (FLOPs), the number of parameters (Params), the storage usage of models (storage usage), and F1-score (F1). The specific results are given in Table III. It can be seen that the FLOPs, Params, and storage usage of our proposed network are lower than half of the backbone, and F1 from our network is the highest, which can verify that our model compression is effective. The FC-EF, FC-Siam-conc, FC-Siam-diff, and FDCNN adopt fewer feature extraction layers, and the number of channels in the last layer is small, leading to small models. However, when confronted with more complex detection problems, their feature extraction ability is weak, resulting in low accuracy and poor robustness. Although the other networks can obtain relatively good accuracy, the calculation and model size are very large. Compared with these networks, our network not only greatly reduces the number of parameters and model size but also achieves the best detection performance.

TABLE III
COMPARISON OF THE EFFICIENCY OF DIFFERENT NETWORKS ON THE LEVIR-CD DATASET. THE BEST VALUES ARE IN BOLD

Methods	FLOPs(GB)	Params(MB)	Storage usage(MB)	F1 (%)
Siam-UNet	80.65	39.70	151.51	87.11
FC-EF	2.63	0.85	3.34	84.18
FC-Siam-conc	4.07	1.07	4.08	81.18
FC-Siam-diff	3.47	0.85	3.33	85.40
W-Net	95.80	31.57	120.48	83.91
FCN-PP	34.65	28.13	107.39	83.27
DSIFN	112.15	43.50	116.92	87.83
FDCNN	32.40	1.86	7.09	85.68
ours	36.75	19.35	73.95	**91.36**

C. Ablation Study

To verify the effectiveness of different modules in our proposed network, we conducted a series of experiments on the LEVIR-CD dataset, including the usage of different combinations of modules. The experimental results are shown in Table IV.

As shown in Table IV, the backbone Siam-UNet introduces the idea of Siamese on the basis of U-Net. The ADC utilizes two different strip convolutions to refine the contour of changed objects provided by the vanilla convolution, which makes the F1 increased by 4.5%, while also reduces the number of parameters and computational cost. Ghost further solves the consumption of parameters and calculation of similar feature maps without decreasing performance. The DE module is added in the process of feature extraction on the basis of ADC and Ghost to enhance difference features that further increase F1 by 1.66%. The SSN module is added after the last layer of feature extraction on the basis of ADC and Ghost to strengthen the integrity and internal tightness of changed objects, which further increases the F1 by 1.31%. The result of the combination of the four modules is 4.25% higher than the backbone, which fully demonstrates the effectiveness of the proposed network.

D. Discussion

1) Discussion on the Effectiveness of DE Module and SSN Module: The DE module is used to avoid the adverse impact of fake changes caused by noise, sunlight, and so on, and some visualized results are shown in Fig. 7 to verify further its effectiveness. Specifically, the inputs and outputs of the last DE module on the LEVIR-CD validation set are shown in Fig. 7, where red indicates the area with higher attention, and blue indicates lower attention. The reason that we choose the last DE module for visualization is that the last layer

TABLE IV

QUANTITATIVE ANALYSIS OF ABLATION EXPERIMENTS ON THE LEVIR-CD DATASET. THE BEST VALUES ARE IN BOLD

Methods	ADC	Ghost	DE	SSN	Pre(%)	Rec(%)	F1(%)
Siam-UNet+					84.75	89.60	87.11
Siam-UNet+	✓				85.23	90.02	87.56
Siam-UNet+	✓	✓			85.54	89.82	87.63
Siam-UNet+	✓	✓	✓		88.74	89.85	89.29
Siam-UNet+	✓	✓		✓	87.80	90.11	88.94
Siam-UNet+	✓	✓	✓	✓	**90.99**	**91.73**	**91.36**
Siam-UNet+SAM+	✓	✓			85.78	89.90	87.79
Siam-UNet+CAM+	✓	✓			86.23	89.91	88.03

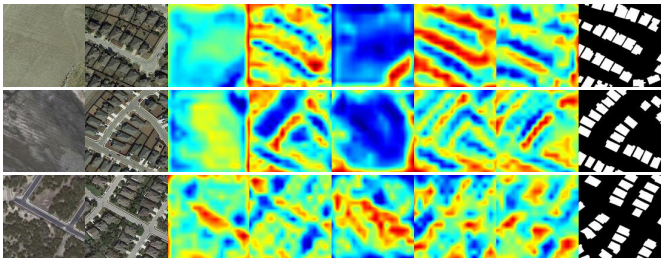


Fig. 7. Visualization of attention maps of the last DE module. (a) Pretemporal image. (b) Posttemporal image. (c) and (d) Two inputs of the DE module. (e)–(g) Three outputs of the DE module. (h) Label.

contains rich semantic information and appears to present a better performance. Posttemporal images of the input and output of the DE module are represented in Fig. 7(d) and (e), respectively. We can see that the enhanced image pays more attention to buildings, and the attention areas of the difference map (g) are more similar to real changes, which benefits the decoder to provide a better position and improve the contour accuracy of changed objects.

In fact, as the channel fusion of the two feature maps easily leads to extremely redundant features, we, thus, adopt a nonlocal channel attention mechanism to suppress the redundancy. In addition, spatial attention can obtain the relationship among pixels. When the relationships of some pixels are similar, they can be regarded as one category and can be classified more easily and clearly. Therefore, we adopt the method of combining spatial and spectral information. In view of the disadvantage of the high computational cost in the previous combination methods, we propose the SSN module by introducing spatial information into channel attention, which greatly reduces the computational cost and achieves higher accuracy. To demonstrate the effectiveness of the module, the qualitative result is shown in Fig. 8. We can find that the effect of the SSN module is quite obvious. Although there is also a flaw with this module that it also pays attention to the roads, this problem can be solved by combining the DE module simultaneously because the skip-connection performed on DIs can provide more detailed information to resolve the erroneous attention. In this way, the final combined model can get a satisfactory result. Besides, we also compare the performance of this module with only the original spatial attention module (SAM) and channel attention module (CAM); the results

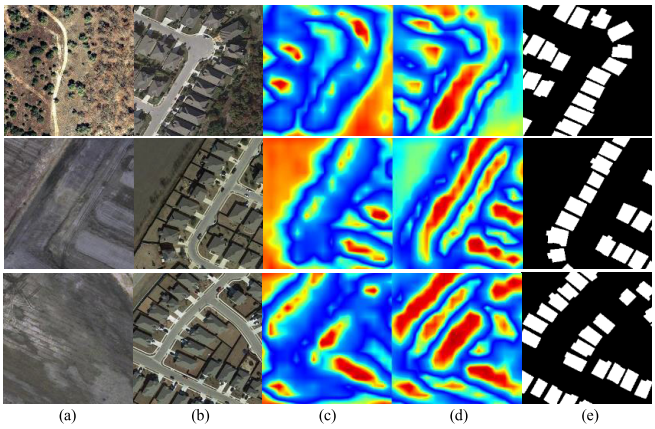


Fig. 8. Visualization of attention maps of the SSN module. (a) Pretemporal image. (b) Posttemporal image. (c) Input of the SSN module. (d) Output of the SSN module. (e) Label.

TABLE V

COMPARISONS BETWEEN DIFFERENT COMBINATIONS OF POOLING SCALES IN MPS ON THE LEVIR-CD TEST SET. THE BEST VALUES ARE IN BOLD

Methods	Pooling scale				Pre(%)	Rec(%)	F1(%)
	16	8	4	2			
SSN+	✓				84.61	90.21	87.32
SSN+		✓			85.91	89.88	87.85
SSN+			✓		86.34	89.62	87.95
SSN+				✓	86.00	89.93	87.92
SSN+	✓	✓			85.07	89.31	87.14
SSN+	✓	✓	✓		86.04	90.01	87.98
SSN+	✓	✓	✓	✓	**87.80**	**90.11**	**88.94**

are shown in the last two rows of Table IV. When the SSN module is adopted, the accuracy is 1.25% higher than SAM and 1.01% higher than CAM, which further demonstrates the effectiveness of the SSN module.

2) Discussion on Optimal Selection of MPS Module Scales in SSN Module: The purpose of the nonlocal module is to capture the relationship among features. Considering the huge amount of calculation by conducting direct matrix multiplication among feature maps for obtaining the autocorrelation matrix, the downsampling using average pooling is conducted to obtain fewer representative pixels. Then, the feature maps after downsampling are leveraged to calculate the autocorrelation matrix. However, it is a problem to determine an appropriate sampling scale. In response to this problem, we conducted experiments on different scales and their combinations. The experimental results are shown in Table V. It can be found that, when the value of a single scale is too large or small, i.e., 16 and 2, the SSN+ cannot achieve the best performance. However, when the value is 4, the SSN+ achieves the best performance. Furthermore, we can see that the information fusion of difference scales can improve the performance of SSN+ since the combination of local and global information leads to better feature representation. In general, the pooling scale with (16, 8, 4, 2) produces the best performance in our experiment.

V. Conclusion

In this article, we have proposed a DESSN network for CD in VHR RS images. The proposed DESSN network addresses the main problems in popular Siamese CNNs for the RS image CD by introducing three main modules: the DE module, the SSN module, and the ADCG module. Specifically, the DE module can reduce the impact of irrelevant changes on the detection results. The SSN module can reduce the redundancy of features after fusion and strengthen the compactness between changed pixels. The ADCG module can refine the edges of changed objects and greatly reduce the number of parameters. The experiments on two popular CD datasets, including LEVIR-CD and CDD, have demonstrated the validity of these modules. Moreover, the experimental results show that the proposed DESSN network is superior to popularly state-of-the-art networks in terms of detection accuracy and efficiency.

References

[1] D. Lu, P. Mausel, E. Brondizio, and E. Moran, "Change detection techniques," *Int. J. Remote Sens.*, vol. 25, no. 12, pp. 2365–2401, Apr. 2004.

[2] J. Xiao *et al.*, "Evaluating urban expansion and land use change in Shijiazhuang, China, by using GIS and remote sensing," *Landscape Urban Planning*, vol. 75, nos. 1–2, pp. 69–80, Feb. 2006.

[3] 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, 2009.

[4] Y. Fan, Q. Wen, W. Wang, P. Wang, L. Li, and P. Zhang, "Quantifying disaster physical damage using remote sensing data—A technical work flow and case study of the 2014 Ludian earthquake in China," *Int. J. Disaster Risk Sci.*, vol. 8, pp. 471–488, Sep. 2017.

[5] C. A. Mucher, K. T. Steinnocher, F. P. Kressler, and C. Heunks, "Land cover characterization and change detection for environmental monitoring of pan-Europe," *Int. J. Remote Sens.*, vol. 21, nos. 6–7, pp. 1159–1181, 2000.

[6] 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.

[7] J. S. Deng, K. Wang, Y. H. Deng, and G. 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, 2008.

[8] S. Marchesi and L. Bruzzone, "ICA and kernel ICA for change detection in multispectral remote sensing images," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Jul. 2009, pp. 980–983.

[9] Z. Li, W. Shi, H. Zhang, and M. Hao, "Change detection based on Gabor wavelet features for very high resolution remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 5, pp. 783–787, May 2017.

[10] 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, 1998.

[11] 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.

[12] 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, 2011.

[13] L. Bruzzone and F. Bovolo, "A novel framework for the design of change-detection systems for very-high-resolution remote sensing images," *Proc. IEEE*, vol. 101, no. 3, pp. 609–630, Mar. 2013.

[14] Z. Lei, T. Fang, H. Huo, and D. Li, "Bi-temporal texton forest for land cover transition detection on remotely sensed imagery," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 2, pp. 1227–1237, Feb. 2014.

[15] K. Tan, J. Xiao, A. Plaza, X. Wang, X. Liang, and P. Du, "Automatic change detection in high-resolution remote sensing images by using a multiple classifier system and spectral-spatial features," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 9, no. 8, pp. 3439–3451, Aug. 2016.

[16] X. Li, B. Zhao, and X. Lu, "MAM-RNN: Multi-level attention model based RNN for video captioning," in *Proc. 26th Int. Joint Conf. Artif. Intell.*, Aug. 2017, pp. 2208–2214.

[17] B. Zhao, H. Li, X. Lu, and X. Li, "Reconstructive sequence-graph network for video summarization," *IEEE Trans. Pattern Anal. Mach. Intell.*, early access, Apr. 9, 2021, doi: 10.1109/TPAMI.2021.3072117.

[18] A. Mohamed, G. E. Dahl, and G. Hinton, "Acoustic modeling using deep belief networks," *IEEE Trans. Audio, Speech, Language Process.*, vol. 20, no. 1, pp. 14–22, Jan. 2012.

[19] J. Zabalza *et al.*, "Novel segmented stacked autoencoder for effective dimensionality reduction and feature extraction in hyperspectral imaging," *Neurocomputing*, vol. 185, pp. 1–10, Apr. 2016.

[20] Y. LeCun *et al.*, "Backpropagation applied to handwritten zip code recognition," *Neural Comput.*, vol. 1, no. 4, pp. 541–551, 1989.

[21] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3431–3440.

[22] P. F. Alcantarilla, S. Stent, G. Ros, R. Arroyo, and R. Gherardi, "Street-view change detection with deconvolutional networks," *Auto. Robots*, vol. 42, no. 7, pp. 1301–1322, Oct. 2018.

[23] D. Peng and H. Guan, "Unsupervised change detection method based on saliency analysis and convolutional neural network," *J. Appl. Remote Sens.*, vol. 13, no. 2, 2019, Art. no. 024512.

[24] 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 Geosci. Remote Sens. Lett.*, vol. 16, no. 6, pp. 982–986, Jun. 2019.

[25] S. Chopra, R. Hadsell, and Y. LeCun, "Learning a similarity metric discriminatively, with application to face verification," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 1, Jun. 2005, pp. 539–546.

[26] R. Caye Daudt, B. Le Saux, and A. Boulch, "Fully convolutional Siamese networks for change detection," in *Proc. 25th IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2018, pp. 4063–4067.

[27] R. Liu, M. Kuffer, and C. Persello, "The temporal dynamics of slums employing a CNN-based change detection approach," *Remote Sens.*, vol. 11, no. 23, p. 2844, Nov. 2019.

[28] R. Caye Daudt, B. Le Saux, A. Boulch, and Y. Gousseau, "Multitask learning for large-scale semantic change detection," *Comput. Vis. Image Understand.*, vol. 187, Oct. 2019, Art. no. 102783.

[29] H. Lyu and H. Lu, "Learning a transferable change detection method by recurrent neural network," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, Jul. 2016, pp. 5157–5160.

[30] M. RuBwurm and M. Korner, "Temporal vegetation modelling using long short-term memory networks for crop identification from medium-resolution multi-spectral satellite images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jul. 2017, pp. 11–19.

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

[32] C. Zhang *et al.*, "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, Aug. 2020.

[33] 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, p. 1662, May 2020.

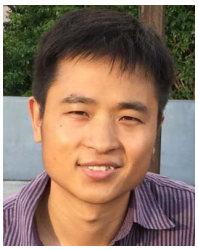
[34] H. Chen, C. Wu, B. Du, L. Zhang, and L. Wang, "Change detection in multisource VHR images via deep Siamese convolutional multiple-layers recurrent neural network," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 4, pp. 2848–2864, Apr. 2020.

[35] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, 2015, pp. 234–241.

[36] L. Bruzzone and D. F. Prieto, "A minimum-cost thresholding technique for unsupervised change detection," *Int. J. Remote Sens.*, vol. 21, no. 18, pp. 3539–3544, 2000.

[37] F. Bovolo and L. Bruzzone, "An adaptive thresholding approach to multiple-change detection in multispectral images," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Jul. 2011, pp. 233–236.

[38] S. Krinidis and V. Chatzis, "A robust fuzzy local information C-means clustering algorithm," *IEEE Trans. Image Process.*, vol. 19, no. 5, pp. 1328–1337, May 2010.
[39] M. Gong, L. Su, M. Jia, and W. Chen, "Fuzzy clustering with a modified MRF energy function for change detection in synthetic aperture radar images," *IEEE Trans. Fuzzy Syst.*, vol. 22, no. 1, pp. 98–109, Feb. 2014.
[40] 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, p. 1381, Aug. 2018.
[41] D. Xue *et al.*, "Unsupervised change detection using multiscale and multiresolution gaussian-mixturemodel guided by saliency enhancement," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 1796–1809, Dec. 2020, doi: 10.1109/JSTARS.2020.3046838.
[42] X. Zheng, X. Chen, X. Lu, and B. Sun, "Unsupervised change detection by cross-resolution difference learning," *IEEE Trans. Geosci. Remote Sens.*, early access, May 31, 2021, doi: 10.1109/TGRS.2021.3079907.
[43] 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.
[44] B. Zhao, X. Li, and X. Lu, "CAM-RNN: Co-attention model based RNN for video captioning," *IEEE Trans. Image Process.*, vol. 28, no. 11, pp. 5552–5565, Nov. 2019.
[45] H. Ning, B. Zhao, and Y. Yuan, "Semantics-consistent representation learning for remote sensing image–voice retrieval," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2021, Art. no. 4700614.
[46] X. Lu, W. Zhang, and X. Li, "A coarse-to-fine semi-supervised change detection for multispectral images," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 6, pp. 3587–3599, Jun. 2018.
[47] Y. Zhong, A. Ma, Y. S. Ong, Z. Zhu, and L. Zhang, "Computational intelligence in optical remote sensing image processing," *Appl. Soft Comput.*, vol. 64, pp. 75–93, Mar. 2018.
[48] M. Gong, J. Zhao, J. Liu, Q. Miao, and L. Jiao, "Change detection in synthetic aperture radar images based on deep neural networks," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 1, pp. 125–138, Jan. 2015.
[49] Y. Zhan, K. Fu, M. Yan, X. Sun, H. Wang, and X. Qiu, "Change detection based on deep Siamese convolutional network for optical aerial images," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 10, pp. 1845–1849, Aug. 2017.
[50] J. Bromley, I. Guyon, Y. LeCun, E. Säckinger, and R. Shah, "Signature verification using a 'Siamese' time delay neural network," in *Proc. Adv. Neural Inf. Process. Syst.*, 1994, pp. 737–744.
[51] S. Zagoruyko and N. Komodakis, "Learning to compare image patches via convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 4353–4361.
[52] J. Chen *et al.*, "DASNet: Dual attentive fully convolutional siamese networks for change detection in high-resolution satellite images," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 1194–1206, Nov. 2020, doi: 10.1109/JSTARS.2020.3037893.
[53] H. Chen, C. Wu, B. Du, and L. Zhang, "Deep Siamese multi-scale convolutional network for change detection in multi-temporal VHR images," in *Proc. 10th Int. Workshop Anal. Multitemporal Remote Sens. Images (MultiTemp)*, Aug. 2019, pp. 1–4.
[54] H. Sun, X. Zheng, X. Lu, and S. Wu, "Spectral-spatial attention network for hyperspectral image classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 5, pp. 3232–3245, May 2020.
[55] 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.
[56] A. Vaswani *et al.*, "Attention is all you need," 2017, *arXiv:1706.03762*.
[57] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.
[58] S. Woo, J. Park, J. Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, vol. 2018, pp. 3–19.
[59] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2018, pp. 7794–7803.
[60] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena, "Self-attention generative adversarial networks," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 7354–7363.
[61] J. Fu *et al.*, "Dual attention network for scene segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3146–3154.
[62] C. Qian, H. Li, and G. Zeng, "Bi-directional cross-modality feature propagation with separation-and-aggregation gate for RGB-D semantic segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2020, pp. 561–577.
[63] H. Sun, S. Li, X. Zheng, and X. Lu, "Remote sensing scene classification by gated bidirectional network," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 1, pp. 82–96, Jan. 2020.
[64] E. L. Denton, W. Zaremba, J. Bruna, Y. LeCun, and R. Fergus, "Exploiting linear structure within convolutional networks for efficient evaluation," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 1269–1277.
[65] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1251–1258.
[66] S. Han, J. Pool, J. Tran, and W. J. Dally, "Learning both weights and connections for efficient neural networks," 2015, *arXiv:1506.02626*.
[67] K. Han, Y. Wang, Q. Tian, J. Guo, C. Xu, and C. Xu, "GhostNet: More features from cheap operations," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 1580–1589.
[68] X. Ding, Y. Guo, G. Ding, and J. Han, "ACNet: Strengthening the kernel skeletons for powerful CNN via asymmetric convolution blocks," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 1911–1920.
[69] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2818–2826.
[70] M. Lebedev, Y. V. Vizilter, O. Vygolov, V. Knyaz, and A. Y. Rubis, "Change detection in remote sensing images using conditional adversarial networks," *ISPRS Int. Arch. Photogramm., Remote Sens. Spatial Inf. Sci.*, vol. 42, no. 2, pp. 565–571, May 2018.
[71] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.
[72] B. Hou, Q. Liu, H. Wang, and Y. Wang, "From W-Net to CDGAN: Bitemporal change detection via deep learning techniques," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 3, pp. 1790–1802, Mar. 2020.
[73] 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.



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 Post-Doctoral Research Fellow with the School of Electronics and Information, Northwestern Polytechnical University. From 2015 to 2016, he was a Visiting Scholar with the Quantum Computation and Intelligent Systems Group, University of Technology Sydney, Ultimo, NSW, Australia. He has authored or coauthored more than 80 research papers, including IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE TRANSACTIONS ON FUZZY SYSTEMS, IEEE TRANSACTIONS ON GEOSCIENCE AND REMOTE SENSING, IEEE GEOSCIENCE AND REMOTE SENSING LETTERS, the International Conference on Acoustics, Speech, and Signal Processing, the IEEE International Conference on Image Processing, and IEEE International Conference on Automatic Face and Gesture Recognition. He is currently a Professor with the School of Electronic Information and Artificial Intelligence, Shaanxi University of Science and Technology, Xi'an. His research interests include image processing, pattern recognition, and machine learning.

Jie Wang received the bachelor's degree from the Shaanxi University of Science and Technology, Xi'an, China, in 2019, where she is going to pursue the M.S. degree at the School of Electrical and Control Engineering.

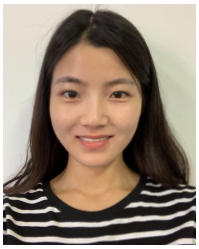
Her research interests include image processing and pattern recognition.

Hailong Ning received the Ph.D. degree in signal and information processing from the University of Chinese Academy of Sciences (UCAS), 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 main research interests include pattern recognition, machine learning, computer vision, and multimodal learning.

Xingwu Wang received the bachelor's degree in information security from Xidian University, Xi'an, China, in 2019. He is going to pursue the M.S. degree at the School of Electronic Information and Artificial Intelligence, Shaanxi University of Science and Technology, Xi'an.

His research interests include image processing and pattern recognition.



Dinghua Xue received the M.S. degree from the Shaanxi University of Science and Technology, Xi'an, China, in 2019, where she is currently pursuing the Ph.D. degree with the School of Electrical and Control Engineering.

Her research interests include image processing and pattern recognition.

Qi Wang (Senior Member, IEEE) received the B.E. degree in automation and the Ph.D. degree in pattern recognition and intelligent systems from the University of Science and Technology of China, Hefei, China, in 2005 and 2010, respectively.

He is currently a Professor with the School of Artificial Intelligence, Optics and Electronics (iOPEN), Northwestern Polytechnical University, Xi'an, China. His research interests include computer vision, pattern recognition, and remote sensing.

Asoke K. Nandi (Life Fellow, IEEE) received the Ph.D. degree in physics from the Trinity College, University of Cambridge, Cambridge, U.K., in 1979.

He held academic positions in several universities, including Oxford University, Oxford, U.K., Imperial College London, London, U.K., the University of Strathclyde, Strathclyde, U.K., and the University of Liverpool, Liverpool, U.K., as well as Finland Distinguished Professorship in Jyväskylä University, Jyväskylä, Finland. In 2013, he moved to Brunel University London, Uxbridge, U.K., to become the Chair and the Head of Electronic and Computer Engineering. He is currently a Distinguished Visiting Professor with Xi'an Jiaotong University, Xi'an, China, and an Adjunct Professor with the University of Calgary, Calgary, AB, Canada. In 1983, he codiscovered the three fundamental particles known as W^+, W, and Z^0 (by the UA1 Team, CERN, Meyrin, Switzerland), providing the evidence for the unification of the electromagnetic and weak forces, for which, in 1984, the Nobel Committee for Physics 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 250 journal articles and 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 lie in signal processing and machine learning, with applications to communications, image segmentations, biomedical data, and so on.

Prof. Nandi is also a fellow of the Royal Academy of Engineering, U.K., and seven other institutions, including the Institution of Engineering and Technology. Among the many awards, he received 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 was an IEEE EMBS Distinguished Lecturer from 2018 to 2019.