SSCFNet: A Spatial-Spectral Cross Fusion Network for Remote Sensing Change Detection

Jiahao Wang, Fang Liu, *Senior Member, IEEE*, Licheng Jiao, *Fellow, IEEE*, Hao Wang, Hua Yang, Xu Liu, *Member, IEEE*, Lingling Li, *Senior Member, IEEE*, and Puhua Chen, *Senior Member, IEEE*

***Abstract*—Convolutional neural networks (CNNs) are data-driven methods that automatically extract the rich information embedded in remote sensing images. However, most current deep learning-based remote sensing image change detection methods prioritize high-level semantic features, while not enough attention is given to low-level semantic features, resulting in the loss of edges and details of the change region. To address this problem, this article constructs a spatial-spectral cross fusion network (SSCFNet), divided into the following three modules: 1) a feature extractor network module; 2) a combined enhancement module; 3) a semantic cross-fusion module. A new combined enhancement strategy is proposed to construct several semantic feature blocks in the combined enhancement module. Different convolution operations are applied to the newly constructed semantic feature blocks in the semantic cross fusion module, and the obtained semantic features at various levels are cross-fused. Experiments show that the proposed SSCFNet outperforms the other six state-of-the-art methods on four publicly available remote sensing image change detection datasets.**

***Index Terms*—Change detection, combined enhancement, convolutional neural network (CNN), cross-fusion, remote sensing image.**

I. Introduction

Convolutional neural networks (CNNs) excel in various fields of deep learning, mainly due to their highly effective feature extraction capabilities, as supported by several studies [1], [2], [3], [4], [5], [6], [7], [8], [9]. For complex tasks such as target detection, semantic segmentation, and change detection [10], [11], [12], the need for high-quality features is even more significant due to the diverse and intricate nature of the scenes. Consequently, numerous scholars have dedicated their research to explore how to achieve optimal feature extraction.

It is essential to utilize remote-sensing images to assist in decision-making. On the one hand, it decreases the human and material resources invested in the process of investigation [13]; on the other hand, dynamic monitoring of the land surface change using remote sensing technology is an important technique that is very commonly used in the agricultural investigation, land use [14], urban expansion [15], disaster monitoring [16], and other applications. Especially in agriculture, change detection is often used for arable land area control, plantation monitoring, disaster assessment [17], deforestation monitoring, forest resource control, etc. For urban areas, building change monitoring [18] is also a helpful task. It is of great interest in applications, such as urban environment, town expansion monitoring, urban development planning, and assessment of natural disasters like earthquakes [19].

Remote sensing change detection refers to the process of extracting change information in image pairs acquired from the same geographical location at different time phases [20]. In a multitemporal image, pixels are classified as either changed or unchanged, and a binary label is assigned to each pixel to indicate whether it has changed or not. Finally, a change map is obtained. The general process of remote sensing change detection is as follows.

1) Firstly, multitemporal remote sensing images are preprocessed to provide high-quality data input (correction, enhancement, and registration [21] of original data, etc.).
2) Secondly, feature extraction and selection (spectral, spatial, object, and scene) can be carried out [22], [23]. In this process, feature fusion [24] at different levels can be carried out, which is related to the change detection algorithm model. To a certain extent, deep networks are also automatic feature extraction processes.
3) Then, construct change indices, mainly interpolation calculations, ratios, similarities, etc., to help us integrate various change information from multitemporal data into discriminative feature maps.
4) Finally, the change detection algorithm models are performed on extracted feature maps. The final change map was obtained.

Traditional change detection methods rely greatly on manually designed feature descriptors [25], and the construction of descriptors highly depends on the experience of experts and domain knowledge. With the wide application of deep

Manuscript received 21 February 2023; revised 22 March 2023; accepted 11 April 2023. Date of publication 14 April 2023; date of current version 2 May 2023. This work was supported in part by the National Natural Science Foundation of China under Grant 62076192, in part by the Key Research and Development Program in Shaanxi Province of China under Grant 2019ZDLGY03-06, in part by the State Key Program of National Natural Science of China under Grant 61836009, in part by the Program for Cheung Kong Scholars and Innovative Research Team in University under Grant IRT_15R53, in part by The Fund for Foreign Scholars in University Research and Teaching Programs (the 111 Project) under Grant B07048, and in part by the Key Scientific Technological Innovation Research Project by Ministry of Education, the National Key Research and Development Program of China, and the CAAI Huawei MindSpore Open Fund. *(Corresponding author: Fang Liu.)*

The authors are with the Key Laboratory of Intelligent Perception and Image Understanding of Ministry of Education, International Research Center for Intelligent Perception and Computation, Joint International Research Laboratory of Intelligent Perception and Computation, School of Artificial Intelligence, Xidian University, Xi'an 710071, China (e-mail: jh_wang1024@163.com; f63liu@163.com; lchjiao@mail.xidian.edu.cn; stuhaowang@163.com; yanghuayts@163.com; xuliu361@163.com; llli@xidian.edu.cn; huaer1986710@126.com).

Digital Object Identifier 10.1109/JSTARS.2023.3267137

neural network models in the field of change detection, both the powerful feature extraction capability and pattern modeling competence of the CNNs are shown.

In the continuous development of change detection methods based on CNNs, the following two main feature extraction frameworks have been derived: 1) one is a single stream network, which takes image pairs as input to generate the change map directly; 2) the other is the siamese network [26]. Each branch of the siamese network shares the weights and the images of two periods go through two network branches for feature extraction. Subsequently, the obtained features are processed by the following processing.

In general, CNN extracts discriminative features from images layer by layer, from shallow to deep. During this process, a valid receptive field is crucial for the quality of extracted features. The perceptual field size of the network affects the spatial information and semantic representation of the output features. Some researchers have made some attempts in this field, such as the DeepLab family of networks for semantic segmentation [27], [28], [29], [30], pyramid scene parsing network (PSPNet) [31], and criss-cross attention network (CCNet) [32]. It is well known that in deep neural networks, deeper layers have larger receptive fields and shallow layers have smaller ones. Deeper features with larger receptive fields have a strong ability to represent semantic information but lose a part of spatial information and geometric details after many convolution operations; shallow features with smaller receptive fields have rich spatial information and high resolution but a weak ability to represent high-level semantic information [33].

Because of the powerful data pattern modeling and feature representation learning ability, CNNs were introduced and now are widely used in remote sensing change detection. Although CNNs have been applied to various change detection methods, most existing methods utilize a single layer of features integrated at the end of the backbone network without effectively utilizing the features' substantial and essential semantic representation information at different levels of the intermediate hidden layers.

Features at different levels and scales are vital for various downstream tasks. For example, in the object detection method single-shot multiBox detector [34], a multiscale feature map is used to predict targets, using shallow low-level features to predict small targets and deep high-level features to predict large targets. However, in predicting small targets, using only shallow low-level features will make the prediction of small targets unsatisfactory. Low-level features contain rich spatial features of small targets, while high-level features focus more on large targets. In the meantime, low-level features concentrate on spatial information, while high-level features focus on abstract semantic features. For the detection of small targets, both spatial information in low-level features and semantic information in high-level features are indispensable. Therefore, to reduce the false alarm rate of target detection, high-level and low-level features are necessary for detecting both small and large targets. As with target detection, change detection requires multiscale, multilevel information. Hence how to effectively exploit the feature at different semantic levels is the focus of this work.

A commonly used strategy is to perform multiscale feature fusion to account for low-level and high-level features. Low-level and high-level feature maps complement each other and guide each other to improve the final discrimination performance. Feature fusion strategies roughly include simple and direct concatenation or using high-level semantic information to assist the training of low-level feature maps and selectively fusing low-level features. In the change detection experiments in this work, we demonstrate that the shallow low-level features are more biased toward detecting small discrete targets, while the deep high-level features are more friendly to large individual targets, as shown in Fig. 1.

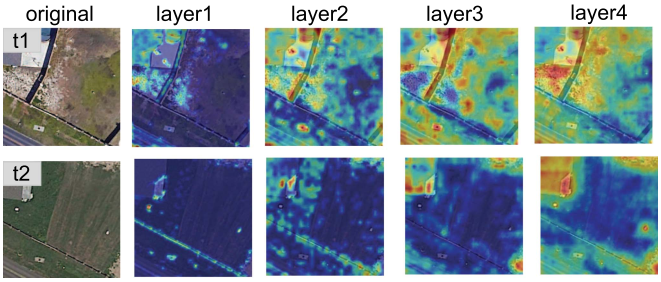


Fig. 1. Visualization graph of four levels of features extracted from remote sensing images in two time phases t1 and t2 are illustrated [35]. The feature extracting backbone network is based on ResNet101. As the network layers change from shallow to deep, the extracted features have different biases.

To address the limitations of existing change detection methods, which do not effectively utilize multiscale features and focus solely on increasing the diversity of discriminative features, we implement a novel spatial-spectral cross fusion module (SSCF) in this work, inspired by the pyramid squeeze attention module in efficient pyramid squeeze attention block on convolutional neural network (EPSANet) [36] and multiscale design strategies for networks [37]. The SSCF combines features from each hidden layer in the feature extractor module, and applies different attention mechanisms to the resulting feature maps. This enhances the feature maps with multiscale information and stronger semantic representation capability. By incorporating the proposed SSCF into a different backbone as the feature extraction module, we achieve state-of-the-art change detection performance on several public datasets, including LEVIR building change detection dataset (LEVIR-CD), WHU building dataset (WHU), season-varying, and Sun Yat-sen university change detection dataset (SYSU-CD).

The main contributions in this work are summarized below.

1) A novel remote sensing image change detection model spatial-spectral cross fusion network (SSCFNet) is implemented, which is superior to other advanced methods on multiple remote sensing image change detection datasets.
2) An effectively combined enhancement module is proposed. The representative ability of the combined features is enhanced by the misplaced combination of semantic features at different levels.
3) A novel semantic cross-fusion module is proposed. This module provides cross-fusion enhancement of semantic features by performing different convolution operations on each feature after the misplaced combination, and the semantic information in each feature sufficiently and differently interacts.

This article is organized as follows. Section II introduces several related works, while Section III provides a detailed description of the proposed SSCFNet. Section IV presents and analyzes the experimental setup and results. In Section V, the limitations of the proposed method are discussed. Finally, Section VI concludes this article.

II. Related Work

In this section, we will present several related works in remote sensing change detection. The general change detection pipeline includes feature extraction, change index construction, and a change detection algorithm model. Most traditional change detection methods employ hand-crafted feature descriptors that perform well under certain conditions. However, given the complexity of land cover in real-world situations, these manually designed feature descriptors may not always meet the necessary assumptions, leading to unsatisfactory performance.

By the time of the popularity of deep learning, deep neural network-based methods had primarily improved the detection accuracy and robustness. A more intuitive idea is to consider this task as a semantic segmentation task with dual-temporal inputs, so that we can obtain prior knowledge from the semantic segmentation literature. Network structures like ResNet [8], DenseNet [9], EfficientNet [38], and InceptionNet [7] that perform well in other fields are used to carry out remote sensing image change detection. Unlike traditional methods, CNN is a data-driven method that automatically extracts rich information from remote sensing images.

Semantic segmentation has successfully applied fully convolutional networks [11]. To some extent, the change detection task can be viewed as a semantic segmentation task. Therefore, many fully supervised change detection approaches are based on a two-branch siamese fully convolutional network (FCN) architecture. The FCN network represented by UNet [39] has been applied by many scholars in the task of remote sensing image change detection.

As far as we know, many CNN-based remote sensing image change detection methods use Siamese networks as their feature extracting network structure. However, some other methods leverage a single-stream architecture to generate the change map. In siamese networks [40], two weights-shared networks are used to extract features from each stream of input, which are then fed to the following operation modules to obtain the final change prediction map. Feature fusion operation and attention mechanisms are often combined with siamese networks. After feature fusion or attention mechanisms, information from different temporal stages can be meaningfully fused and aggregated. Considering the stage of fusion, feature fusion can be divided into early fusion and late fusion. From the perspective of fusion methods, it can be divided into single-scale fusion and multiscale fusion. Single-scale fusion only fuses the highest-level features, while multiscale fusion can map low-level spatial information to high-level semantic features.

In [41], the dual-temporal images are early-fused (concatenated) and fed into an improved UNet++. Daudt et al. [42] used different feature fusion strategies on the UNet for change detection. Based on the siamese structure, the authors fused the multilayer features extracted from UNet using feature concatenation and feature difference strategies, and constructed FC-Siam-Conc and FC-Siam-Diff, which are two methods to realize change detection of remote sensing images. In [43], Zhang and Shi performed multiscale feature extraction from pairs of input images using a pair of siamese very deep convolutionalnetworks (VGG) networks and fused the extracted features by performing differences and concatenations, generating the final change results. In [44], the authors directly concatenated two-branch features for triplet-loss-based training. In [45], the deep feature extracted from a pretrained multilayer CNN was utilized to construct change vector of multitemporal images and those features were processed through a layerwise feature selection mechanism to retain only the change-relevant features. In [46], the dual-temporal features were constrained both in bitemporal feature extraction and feature fusion. In addition, a nonlocal feature pyramid network, and a dense connection-based feature fusion module were used to fuse the bitemporal information. In [47], an unsupervised multimodal change detection framework based on structural relationships was proposed. The authors represented images with graphs and performed graph convolutions on constructed graphs to reconstruct vertex and edge information. Then, an adaptive variance-based mechanism was leveraged to fuse the local edge and nonlocal vertex information. In [48], based on their previous work, the authors extended the structural relationship analysis to the graph Fourier domain. The proposed framework exploits both local and nonlocal structural relationships and is implemented in the graph Fourier domain. A frequency-decoupling adaptive fusion method was leveraged to fuse local and nonlocal structural difference maps of high and low frequencies separately.

Attention mechanism are also often used in feature fusion. In [49], building extraction is set as an auxiliary task for change detection. The two-stream feature maps are fused by skip connections and a dual attention mechanism. In [49], the dual-attention mechanism captures long-range dependencies and allows for more representative features. Fang et al. [50] proposed siamese network for change detection (SNUNet), a densely connected siamese network for remote sensing image change detection. It reduces the loss of deep features through dense information transmission, and incorporates an improved attention module with channel attention, residual connection weights, and refinement of the most important features for change map generation. In [50], a pyramid-style feature fusion is leveraged, where feature maps at different scales are fused by concatenation, channel attention, and spatial attention. In [52], a pyramid pooling module is used to capture features with different receptive fields. In [53], both channel attention and spatial attention are integrated into each fusion node of the decoder to adaptively emphasize areas that may be relevant to change. In [54], attention-based feature pyramids are used to retain small-scale features and filter spatially and channel-wise useful features, leading to the fusion of bitemporal features.

There are also a lot of related works based on other feature fusion approaches. In [55], the authors use an long short term memory (LSTM) and skip connections to fuse the temporal and spatial information. The gating function of the LSTM is to retain or omit information from the deep features. Xu et al. [56]

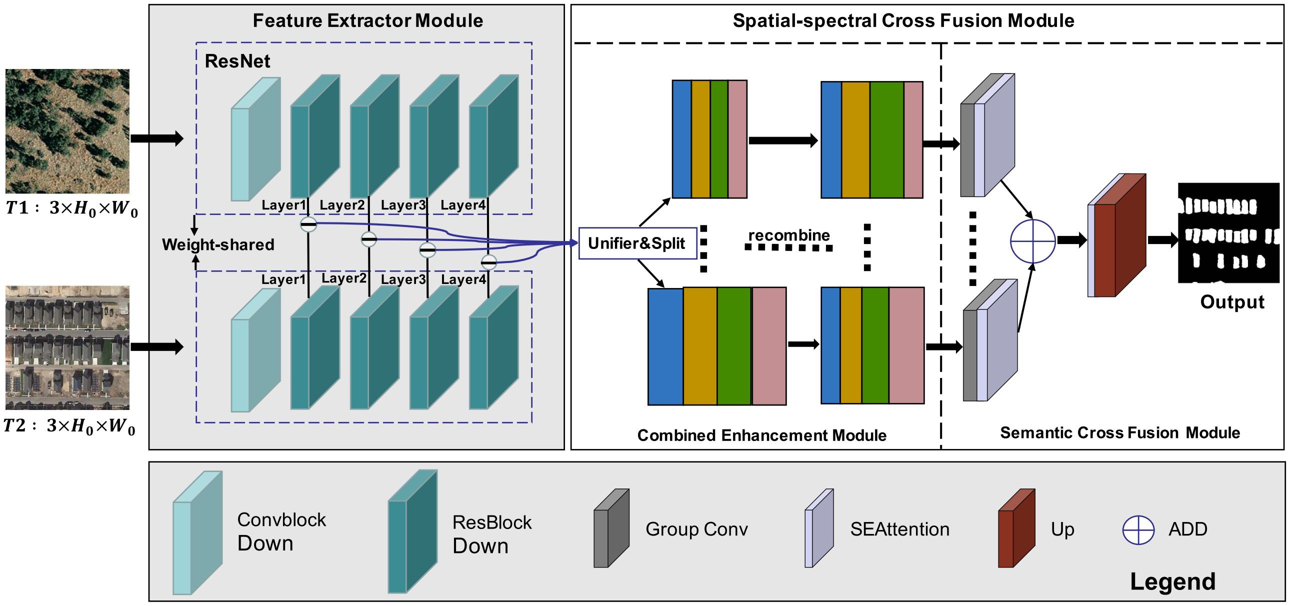


Fig. 2. Proposed network structure for remote sensing image change detection, SSCFNet. The network uses a weight-sharing siamese network structure to obtain semantic features of different layers from two remote sensing images at different time series. The proposed SSCF module is divided into a combined enhancement module for recombining multilayer semantic features and a semantic cross-fusion module for semantic feature fusion. Down indicates downsampling; Up refers to upsampling operation by transpose convolution; SEAttention indicates Squeeze-and-Excitation Block.

proposed multidirectional fusion pathway (MFP-Net), a network for remote sensing image change detection that incorporates a multidirectional fusion path and an adaptive weighted fusion strategy to enhance the flexibility and diversity of information pathways. The adaptive weighted fusion (AWF) policy weights at each fusion node are calibrated to emphasize representative feature maps and resolve semantic differences. Moreover, a new perceptual similarity module and perceptual losses are utilized to generate a high-quality change map. In [57], the authors use point-wise addition on the superpixel-based features to fuse the information in the dual-time phase.

These methods are effective in acquiring pixel-level change maps, but they also have some drawbacks. Although they consider the semantic feature information on different levels, the intermediate scale features are not directly expressed in the final discriminative features. Still, they are concatenated or fused with other features in the feature extraction process and enter the information transfer pathway again. Thus, the information extracted from each hidden layer in the information transfer of feature extraction is not utilized more effectively.

To address the existing methods' inefficient feature utilization problem, we propose the SSCFby using all features of different scales for change map generation. At the same time, the features of different scales are crossed and misplaced for fusion, which increases the diversity of features and achieves feature enhancement, and improves the quality of change maps to some extent.

III. Method

In this section, we first briefly introduce the overall structure of the proposed network in Section III-A. The SSCF is described in Section III-B, which can combine features from various levels in the backbone network to avoid missing small targets in change detection. Finally, the multiloss strategy is illustrated in Section III-C.

A. Overview of SSCFNet

Fig. 2 shows the overall structure of the proposed network in this article. The input to this network is two remote sensing images at different time series, and the output is a binary map of change predictions.

The proposed network SSCFNet consists of the feature extractor module and the SSCF module for spatial spectrum cross-fusion. We use ResNet and ResNext [58] as the backbone network module to extract generic features. Each of these backbone networks has five stages, among which Stage 0 has a simpler structure and can be regarded as a preprocessing of the input, and the last four stages are composed of bottlenecks. To use the semantic features of each layer for change detection, we feed the output features of the last four stages into the SSCF module for fusion. The SSCF is a key structure in our proposed SSCFNet, consisting mainly of a combined enhancement module and a semantic cross-fusion module. The details of these two modules will be provided in Section III-B.

B. Spatial-Spectral Cross Fusion Module

As shown in Fig. 3, the SSCF consists of the following three modules: 1) a feature size unifier; 2) a combined enhancement module; 3) a semantic cross-fusion module.

1) Feature Size Unifier: As shown in Fig. 3(a), the output features from the backbone network's last four stages represent information at different scales. The feature size unifier upsamples these features at different scales to the same scale by nearest

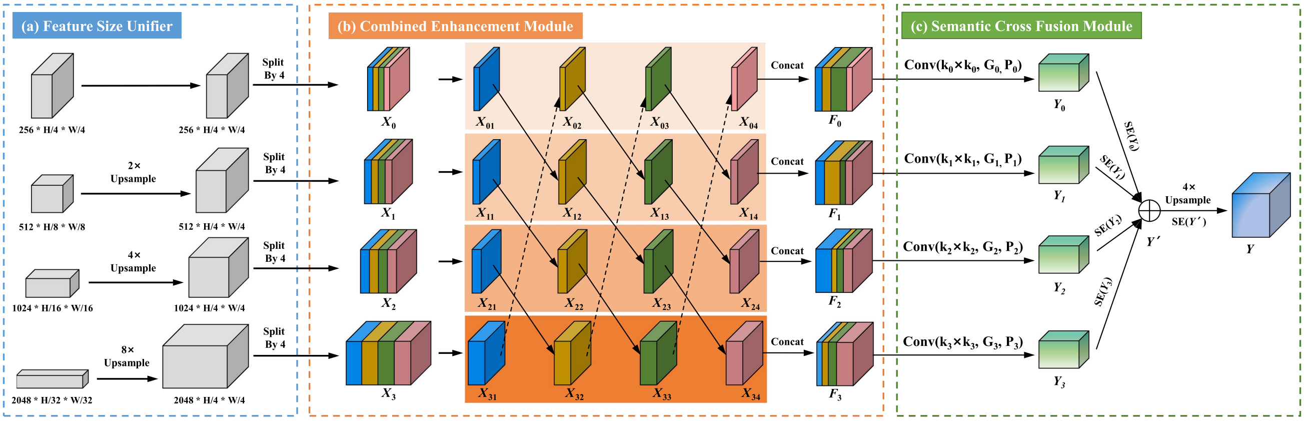


Fig. 3. Illustration of the proposed SSCF.

neighbor interpolation. Here, the size of the first stage features is used as the unifying scale to facilitate subsequent feature cross-fusion

$$S^i_{\text{out}} = Upsample(S^i_{in}, (H_1, W_1)) \quad i \in 1, 2, 3, 4 \tag{1}$$

where H_1, W_1 represent the size of the output features of the first stage of ResNet, and $S_{\text{in}}, S_{\text{out}}$ represent the input and output feature size of the module for each stage, respectively.

2) Combined Enhancement Module: In the combined enhancement module, a combined enhancement strategy is proposed for recombining semantic features at different levels of the backbone network.

The combined enhancement strategy separates the four-level feature maps X_0, X_1, X_2, X_3 equally into four semantic feature blocks in the channel dimension. Then, it recombines them into new semantic feature blocks $F_0 \sim F_3$ according to (2). The new combined semantic feature blocks do not intersect with each other and are different from each other. Such splitting and recombination can obtain richer location information of semantic features and process them on multiple scales in a parallel manner, which makes the low-level semantic features interact more closely with the high-level semantic features. As each feature block in $F_0 \sim F_3$ contains a part of each of X_0, X_1, X_2, X_3, and each input feature map contains semantic and spatial information of different layers, the semantic features are essentially combined and enhanced to obtain more discriminative (more representative of the change region) features at different levels. The combined enhancement strategy is expressed as follows:

$$\begin{cases} F_0 = Concat\left[X_{11}\ X_{22}\ X_{33}\ X_{04}\right] \\ F_1 = Concat\left[X_{21}\ X_{32}\ X_{03}\ X_{14}\right] \\ F_2 = Concat\left[X_{31}\ X_{02}\ X_{13}\ X_{24}\right] \\ F_3 = Concat\left[X_{01}\ X_{12}\ X_{23}\ X_{34}\right] \end{cases} \tag{2}$$

$$F_i = \underset{(S=4)}{Concat}\left(\left[X_{(i+1)\%S,1}\ X_{(i+2)\%S,2}\ \dots\ X_{i,S}\right]\right). \tag{3}$$

In the equation, F_i is the multiscale feature map obtained from the ith layer.

3) Semantic Cross-Fusion Module: In the semantic cross-fusion module, different convolution operations are performed on the newly constructed semantic feature blocks $F_0 \sim F_3$, resulting in feature maps $Y_0 \sim Y_3$, which are then cross-fused to obtain Y'. This approach allows the features in the same channel to consider both the small targets, which are focused on by lower-level features, and the larger targets, which are the focus of higher-level features, thus capturing both rich spatial information and more robust semantic information. Next, we use the channel attention mechanism to selectively strengthen more beneficial features and weaken less significant features in the fusion result and finally up-sample the feature map to produce the final result Y. After our experimental tests, the cross-fusion of features from different channels are more efficient than the sequential fusion of features.

As shown in Fig. 3(b), each branch of the fusion path contains deep and shallow features, so it learns features at all four levels. Then, as shown in Fig. 3(c), convolution kernels with different sizes are used to extract features at different scales for different branches. We use group convolution to address the problem of increasing the number of parameters due to the large size of convolution kernels. To maintain consistency of features across different branches after the convolution operation, padding is used to keep the input and output sizes consistent during the convolution process. The generation function of the feature map after convolution can be written as follows:

$$Y_i = Conv(k_i \times k_i, P_i, G_i)(X) \quad i = 0, 1, 2, 3. \tag{4}$$

In the equation, define $k_i = 2 \times (i + 1) + 1$ as the calculation method of the size of the ith convolution kernel, define $P_i = \lfloor \frac{2\times(i+1)+1}{2} \rfloor$ as the calculation method of the size of the ith padding size, and define $G_i = 2^{i+1}$ as the calculation method of the size of the ith group. Next, we use the channel attention mechanism to select the focus location and produce a more discriminative feature representation. Then, to cross-fuse the convolved results, the spatial and semantic information of each semantic feature block is fused without increasing the computational effort, thus obtaining the entire multiscale channel attention vector Y' in a direct summation as

$$Y' = \sum_{i=0}^{3} SEAttention(Y_i). \tag{5}$$

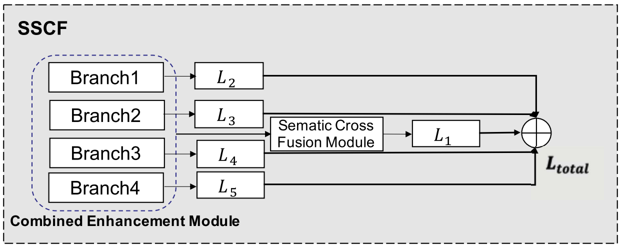


Fig. 4. Illustration of a multiloss module design, where “SSCF” represents spatial-spectral cross fusion.

The cross-fused result Y' is further applied to the channel attention mechanism (SEAttention) [59] then upsampled to obtain the final result Y, Y can be expressed as

$$Y = Upsample(SEAttention(Y')). \tag{6}$$

For the details of SEAttention, take $\mathbf{Y}'$ as an example, the SEAttention of $\mathbf{Y}'$ can be formulated like

$$SEAttention(\mathbf{Y}') = \sum_{i=1}^{C} s_c \cdot Y'_c \tag{7}$$

where the Y'_c is the cth channel of the feature map $\mathbf{Y}'$, the C is the number of channels of $\mathbf{Y}'$, and the s_c can be formulated like

$$s_c = fc_2\left(fc_1\left(GAP(Y'_c)\right)\right). \tag{8}$$

In the equation, the fc_1 is a fully connected layer with rectified linear unit (ReLU) activation function, and the fc_2 is a fully connected layer with Sigmoid activation function. The global average pooling (GAP) can be formulated as

$$GAP(Y'_c) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} Y'_c(i,j). \tag{9}$$

C. Design of Multiloss

An improved version of the proposed SSCF is obtained by adding auxiliary losses after each branch in the combined enhancement module. In addition to using the final prediction graph of the semantic cross-fusion module to calculate the losses, the auxiliary losses are also back-propagated to optimize the parameters jointly. Experimental results show that this leads to more effective cross-fusion of the newly constructed semantic feature blocks. All losses are calculated using the binary cross-entropy (BCE) loss function. For samples in dataset k, the loss function $\mathcal{L}$ can be defined as

$$\mathcal{L} = -\frac{1}{N^k} \sum_{m=1}^{N^k} \sum_{i,j} \left(y_{i,j}^m \log\left(\hat{y}_{i,j}^m\right) + \left(1 - y_{i,j}^m\right) \log\left(1 - \hat{y}_{i,j}^m\right)\right) \tag{10}$$

where $\hat{y}_{i,j}^m$ represents the confidence map predicted by the SSCFNet at the location (i, j) for sample m of dataset k, and $y_{i,j}^m$ represents the ground-truth label of the corresponding pixel in the input image.

As shown in Fig. 4, L_2, L_3, L_4, and L_5 are all auxiliary losses, and loss1 is the main loss. The auxiliary losses help to optimize the learning process, and the main loss is still the main optimization direction. During training, different weights w_i are given to the auxiliary and the main losses to balance the auxiliary losses. In the testing phase, the auxiliary branches are dropped, and only the optimized main branch is used for the final prediction. For the final optimization of the network, the weighted sum of all these losses is denoted as $\mathcal{L}_{\text{total}}$

$$\mathcal{L}_{\text{total}} = \sum_{i=1}^{5} w_i \mathcal{L}_i \tag{11}$$

where w_1 is set to 1 and (w_2, w_3, w_4, w_5) is set to 0.5.

IV. Experiments

This section describes the experimental setup used to evaluate the proposed algorithm in a change detection task. We begin by introducing the four datasets used in the evaluation: LEVIR-CD, WHU, season-varying, and SYSU-CD. Next, we provide an overview of six state-of-the-art comparison methods. We then describe the implementation details of the training process, followed by a quantitative and qualitative comparison of the results obtained by our algorithm on the four datasets with those of the six comparison methods. Finally, we discuss the effectiveness of our proposed method. The source code will be released at https://github.com/Wprofessor/SSCFNet.

A. Dataset Description

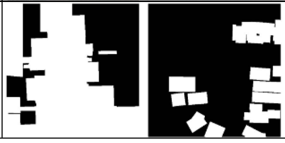
1) LEVIR-CD [60]: LEVIR-CD is a new massive building open change detection dataset [60], as shown in some samples in Table I(a). It contains 637 high-resolution (50 cm/pixel) image pairs of 1024 × 1024 pixels. These dual-temporal images cover changes in various buildings, such as garages, warehouses, and villas spanning from 5 to 14 years. In addition, the remote sensing images originated from 20 different areas in respective cities in TX, USA, containing Austin, Buda, Bee Cave, Dripping Springs, Manor, Pflugervilletx, Kyle, Lakeway, and others. Its authors divided the dataset into a training set, a validation set, and a test set. We cropped each sample into 16 small-size blocks of 256 × 256 pixels using a nonrepeating sliding window, generating 7120 pairs of image blocks for training, 1024 for validation, and 2048 for testing.

2) WHU [61]: This dataset contains aerial images acquired in April 2012 and includes 12 796 buildings over 20.5 km^2 (a new dataset released in 2016 includes 16 077 buildings in the same area). Some examples are shown in Table I(b). The subdataset was geo-corrected to an aerial dataset with a 1.6-pixels accuracy by manually selecting 30 ground control points (GCPs) on the ground surface. This subdataset and the corresponding images from the original dataset are now publicly available, along with the constructed vector and raster maps.

It contains a high-resolution aerial image of size 32 507 × 15 354. No data decomposition scheme is given in [61]. We crop the image into small 512 × 512 pixel blocks without overlapping by sliding window and divide it into three parts: Training set, validation set, and test set, containing 1189, 319, and 319 pairs of image blocks, respectively.

3) Season-Varying [62]: The season-varying dataset [62] consists of seven pairs of 4725 × 2200 pixel high-resolution

TABLE I
SAMPLE DIACHRONIC IMAGES AND GROUND TRUTH FROM LEVIR-CD, WHU DATASET, SEASON-VARYING, AND SYSU-CD CHANGE DETECTION DATASETS

	(a) LEVID-CD	(b) WHU dataset	(c) Season-varying	(d) SYSU-CD
T1				
T2				
GT				

The first row indicates the image of the T1 time phase, the second row shows the image of the T2 time phase, and the third row means the ground truth.

seasonal change images for manual creation of ground truth and four pairs of 1900 × 1000 pixel images for manual addition of other objects. The spatial pixel density of the obtained images is between 3 and 100 cm/px. Some examples are shown in Table I(c). The dataset considers objects of various sizes (e.g., from cars to large building constructions) and seasonal variations of natural things (e.g., from individual trees to vast forest areas). The dataset originated by clipping 256 × 256 randomly rotated segments (0–2) with at least a portion of the target objects. Thus, the target center coordinates are unique, and the distance between target centers is 32 pixels for each axis. Finally, the dataset contains 16 000 pairs of 256 × 256 pixel images: 10 000 training sets, 3000 test, and validation sets.

4) SYSU-CD [62]: The dataset [62] was made up of 20 000 pairs of 0.5 resolution high aerial images recorded in Hong Kong between 2007 and 2014. Some samples are shown in Table I(d). In constructing the dataset, the authors first divided the 800 acquired original image pairs into a training set, a validation set, and a test set. The ratio of training, validation, and test sets was 6:2:2. Then, 25 sample pairs were randomly selected from each image pair, each of which was 256 × 256 in size, to generate the final dataset for use. Random flips and rotations were used to extend the data. After the preprocessing, 20 000 pairs of aerial image patches of size 256 × 256 were obtained. The main types of changes in the dataset included new urban construction, suburban sprawl, preconstruction foundations, vegetation changes, road expansion, and marine construction.

B. Comparison Method

1) FC-Siam-Conc [42]: A feature fusion method that fuses the multilayer features extracted from a siamese full convolutional network by concatenating features from a different branch of the siamese network and skip connection.

2) FC-Siam-Diff [42]: A feature fusion method that uses a siamese full convolutional neural network to extract multilevel features and use feature differences to fuse bitemporal information.

3) Dual Task Constrained Deep Siamese Convolutional Network (DTCDSCN) [49]: A multiscale feature fusion method combines the channel and spatial attention mechanisms in FCN to obtain more discriminative features.

4) SNUNet [50]: A densely connected remote sensing change detection network based on siamese nested UNet. The SNUNet uses many skip connections to effectively transfer information between encoders and decoders in the backbone network. After feature extracting, the extracted multilevel semantic feature maps are fused by an improved attention module based on the channel attention mechanism and residual connection.

5) Deeply Supervised Image Fusion Network (DSIFN) [51]: A deeply supervised network that first uses a two-branch fully convolutional network for feature extraction and then uses a deeply supervised difference discriminative network CNN to detect changes in the input image pairs from extracted features. To enhance the integrity of change map boundaries and internal densities, the extracted multilevel feature maps are fused with the different maps of bitemporal images through an attention mechanism.

6) MFPNet [56]: A multidirectional fusion and perception network for change detection of dual-time sequence high-resolution remote sensing images. An MFPNet consisting of MFP and AWF strategies is proposed. MFP increases the versatility of information ways and simplifies information dissemination. The AWF module stresses the significant feature maps along with irrelevant feature maps that inhibit dependable information transfer. As a result, significant and comprehensive features can be aggregated at each fusion node.

C. Evaluation Metrics

The F1 is the weighted reconciled average of precision and recall, which considers both precision and recall to balance the conflict and can more comprehensively reflect the performance of the change detection model. Therefore, we use the average F1 score of the change category and background as the primary evaluation metric, which is calculated as follows:

$$F1 = \frac{2 \times precision \times recall}{precision + recall}. \tag{12}$$

In addition, we also used Precision, Recall, and intersection over union (IoU) as auxiliary evaluation metrics, which are calculated as follows:

$$Precision = \frac{TP}{TP+FP} \tag{13}$$

$$Recall = \frac{TP}{TP+FN} \tag{14}$$

$$IoU = \frac{TP}{TP+FN+FP}. \tag{15}$$

Among them, TP, FP, TN, and FN represent the number of true positives, false positives, true negatives, and false negatives, respectively.

D. Experiment setup

To train the SSCFNet proposed in this article, ResNet50, ResNeXt50, and ResNet101 are chosen as backbones. We use BCE loss as the loss function. Stochastic gradient descent (SGD) [63] is used as an optimizer and sets the momentum to 0.9. The initialization learning rate is set to 0.01. Warm-up is used to linearly increase the preset learning rate in the first five epochs and then decay according to the cosine function value, and the weight decay is set to 0.0005. We used a minibatch size of eight and trained the models for 100 epochs. For all tasks, we utilize the PyTorch [64] deep learning framework and rely on four Nvidia 2080Ti graphics processing units (GPUs) for efficient processing. We also intend to utilize SSCFNet on MindSpore [65], a new deep learning computing framework.

E. Ablation Experimental Study

To compare the effect of selecting a backbone with different feature layers on the change detection results, we set ResNet50 as the feature extraction network for the ablation experiments. In addition, we compare the proposed SSCF with other methods and visualize the results to better assess the effect of the fusion module on the change detection results. We conduct relevant ablation experiments to verify the effectiveness of group convolution and multiloss, all of which are performed on the LEVIR-CD dataset. Furthermore, we use two composite metrics (mean intersection over union (MIoU) [66] and F1 score) to evaluate the results of the ablation experiments quantitatively in this section.

1) Ablation Experiment for Selection of Feature Layers: To explore whether all feature layers in the feature extraction module affect the results of change detection, we divide the five feature layers into the following five combinations:

1) (0, 1, 2, 3, 4);
2) (1, 2, 3, 4);
3) (2, 3, 4);
4) (3, 4);
5) (4).

Input each of these combinations into the SSCF to explore whether high-level features combined and low-level features positively impact the change detection task. The results are used to select the optimal combination of feature layers.

TABLE II
COMPARATIVE STUDY OF ABLATION EXPERIMENTS

Layers	T_{train} (s)	F1 score	MIoU
(a) 0,1,2,3,4	265	94.94	90.36
(b) 1,2,3,4	160	94.89	90.30
(c) 2,3,4	103	93.91	88.94
(d) 3,4	95	93.87	88.96
(e) 4	110	93.77	88.81

We report the effect of SSCFNet using features from different layers of ResNet50 on change detection results in dataset LEVIR-CD, where T_{train} refers to the time consumed to train an epoch.

TABLE III
RESULTS OF THE ABLATION EXPERIMENTS OF THE PROPOSED METHOD WITH OTHER FUSION MODULES ON THE LEVIR-CD DATASET, WITH THE HIGHEST SCORES MARKED IN BOLD AND ALL SCORES DESCRIBED IN PERCENTAGES

Method	F1 score	MIoU
PSPNet	92.04	86.25
CCNet	92.73	87.15
DeeplabV3+	92.80	87.84
Ours (ResNet50)	**94.89**	**90.30**

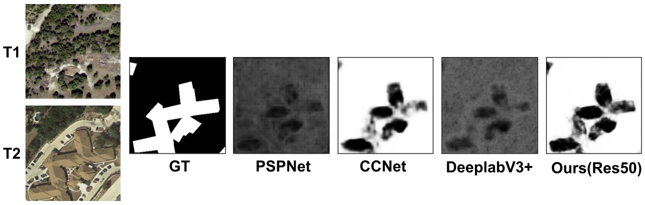


Fig. 5. Comparison of feature visualization graphs for different network fusion modules in the LEVID-CD dataset. “T1” and “T2” denote T1 time phase and T2 time phase, respectively. “GT” is meant as ground truth.

Table II shows the results of the ablation experiments. It can be found that more the number of feature layers, the more the semantic features are enhanced and the higher the main evaluation metric F1 score. However, since the bottom layer features are more inclined to focus on small targets and detailed information, which account for a relatively small proportion in the ratio change detection, especially the layer 0 features contain very little effective information compared with other layers, the F1 score of combination in Table II is not significantly improved compared with that of combination b. On the contrary, the computational resources consumed are substantially higher than that of combination b in the same experimental environment. The F1 score and MIoU are improved by (1.04%, 1.53%), (1.09%, 1.51%), and (1.19%, 1.68%) for combination b compared to combination c, combination d, and combination e, respectively. There is no significant improvement in the computational resources consumed, so in the proposed network SSCFNet, we select combination b as the input to the SSCF.

2) Ablation Experiment for SSCF Module: To demonstrate the effectiveness of the SSCF module, we implemented PSPNet, CCNet, and DeeplabV3+ in the change detection (CD) task to facilitate the comparison of the fusion modules.

As shown in Table III, the results of the models containing different fusion modules were evaluated separately on the LEVIR-CD dataset. It can be observed that the proposed method in this article achieves the best results on both MIoU and F1 score evaluation metrics compared to other models.

The visual results of some representative examples in this ablation study reinforce the findings presented in Table III.

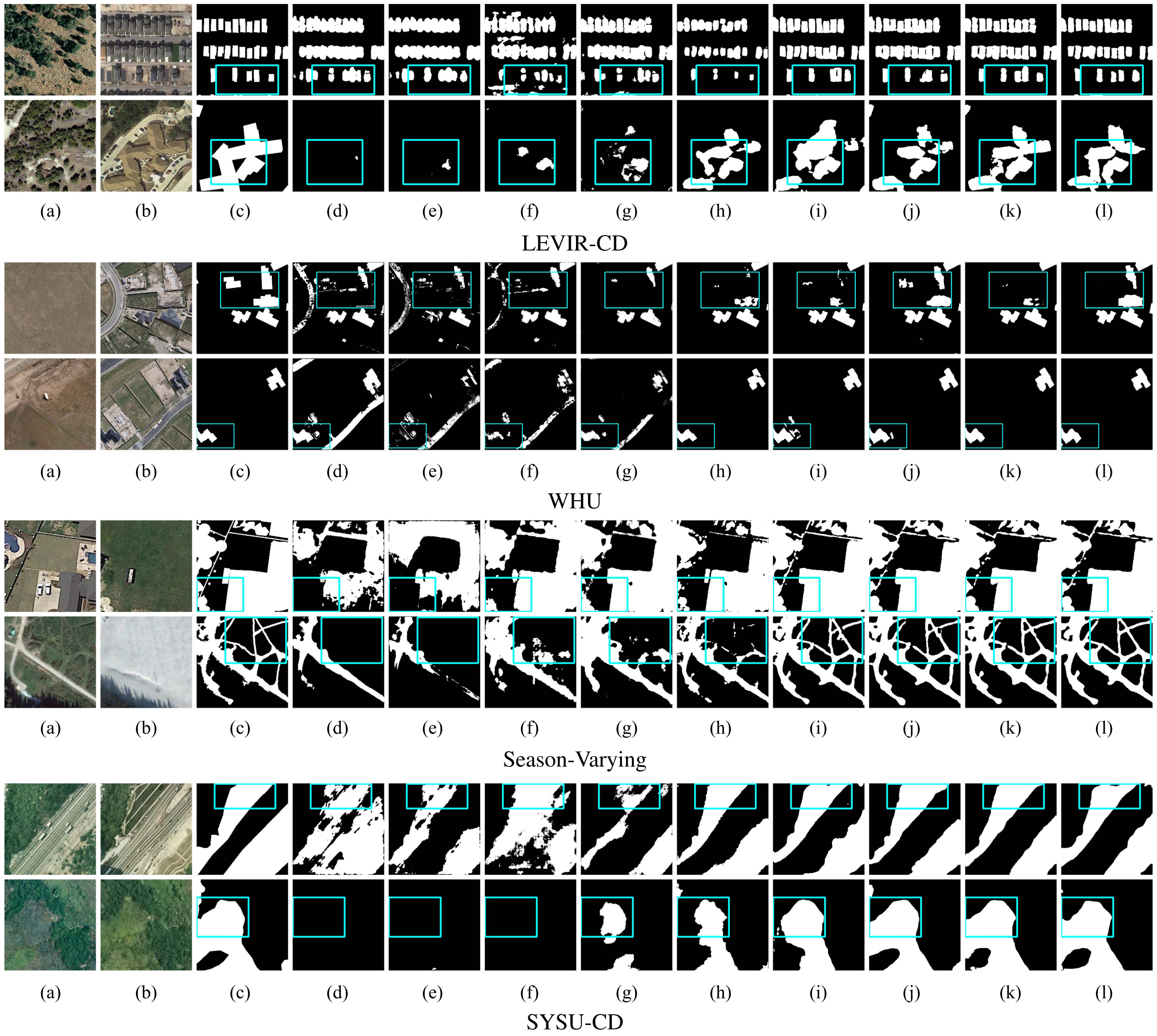


Fig. 6. Qualitative comparison results of change detection in the LEVIR-CD, WHU, size-varying, and SYSU-CD datasets. (a) t1 time phase images. (b) t2 time phase images. (c) Ground truth. (d) FC-Siam-Conc. (e) FC-Siam-Diff. (f) DTCDSCN. (g) SNUNet. (h) DSIFN. (i) MFP-Net. (j) Ours (ResNet50). (k) Ours (ResNeXt50). (l) Ours (ResNet101). The changed area is white and the unchanged area is black.

Fig. 5 provides a visual representation of the feature visualization diagrams of each model's fusion module. Our proposed method outperforms the other models in processing details and detecting large-area targets more completely, demonstrating the effectiveness of the SSCF.

3) Ablation Experiment for Group Convolution and Multiloss: To explore the effect of group convolution and multiloss, a set of ablation experiments was set up to gradually combine the group convolution and multiloss strategies and compared them with the CD task.

TABLE IV
RESULTS OF ABLATION EXPERIMENTS FOR EACH COMPOSITION ON THE LEVIR-CD DATASET

Method	G_conv	Multi loss	F1 score	MIoU
Ours (ResNet50)			94.10	89.45
Ours (ResNet50)	✓		94.51	90.20
Ours (ResNet50)		✓	94.38	89.73
Ours (ResNet50)	✓	✓	94.89	90.30

The "G conv" in the table header means group convolution.

As shown in Table IV, we evaluated the results of different combinations on the F1 score and MIoU metrics for the CD task on the LEVIR-CD dataset. When only group convolution is combined with the proposed SSCFNet for the CD task, performance gains in F1 score and MIoU can be observed (0.44% and 0.84%), due to the sparse relationship between the filters and the fact that dividing the groups can have some regularization effect on the model afterward. When only the multiloss is combined with the proposed SSCFNet for the CD task, performance gains in F1 score and MIoU can be observed

TABLE V
QUANTITATIVE COMPARISON OF CHANGE DETECTION METHODS ON THE LEVIR-CD, WHU, SEASON-VARYING, AND SYSU-CD DATASETS, WITH THE HIGHEST SCORE FOR EACH EVALUATION METRIC IS MARKED IN BOLD BLACK, AND ALL SCORES ARE EXPRESSED AS PERCENTAGES

Dataset	Method	Precision	Recall	F1 score	MIoU
LEVIR-CD dataset	FC-Siam-Conc [42]	85.01	91.64	88	80.27
	FC-Siam-Diff [42]	85.62	91.1	88.14	80.46
	DTCDSCN [49]	83.49	93.43	87.74	79.94
	SNUNet [51]	91.96	93.55	92.74	87.15
	DSIFN [52]	82.74	**96.91**	88.43	80.88
	MFP-Net [57]	93.3	95.18	94.48	89.96
	Ours (ResNet50)	**93.75**	95.16	94.89	90.30
	Ours (ResNeXt50)	93.66	95.88	95.11	90.69
	Ours (ResNet101)	93.71	96.5	**95.31**	**90.87**
WHU dataset	FC-Siam-Conc [42]	89.9	68.27	74.34	64.49
	FC-Siam-Diff [42]	81.91	75.77	78.47	69.06
	DTCDSCN [49]	87.32	71.75	77.18	67.54
	SNUNet [51]	89.12	87.80	88.44	80.95
	DSIFN [52]	90.84	**98.75**	94.42	89.89
	MFP-Net [57]	95.28	95.30	95.29	91.34
	Ours (ResNet50)	95.79	95.72	95.80	91.87
	Ours (ResNeXt50)	96.02	95.65	95.87	91.92
	Ours (ResNet101)	**96.54**	95.92	**96.58**	**92.89**
Season-varying dataset	FC-Siam-Conc [42]	70.8	92.12	76.73	66.14
	FC-Siam-Diff [42]	68.82	92.54	74.74	64.14
	DTCDSCN [49]	85.53	93.09	88.8	80.94
	SNUNet [51]	91.47	93.46	92.43	86.44
	DSIFN [52]	91.92	**98.38**	94.99	90.84
	MFP-Net [57]	97.23	98.1	97.16	95.26
	Ours (ResNet50)	97.54	97.83	97.58	95.57
	Ours (ResNeXt50)	97.56	98.00	97.88	95.81
	Ours (ResNet101)	**97.98**	98.22	**98.20**	**96.52**
SYSU-CD dataset	FC-Siam-Conc [42]	79.94	82.37	81.04	69.44
	FC-Siam-Diff [42]	78.28	83.38	80.34	68.65
	DTCDSCN [49]	77.98	83.34	80.13	68.39
	SNUNet [51]	83.92	85.15	84.51	74.08
	DSIFN [52]	87.22	88.78	87.96	79.12
	MFP-Net [57]	87.68	88.36	88.01	79.18
	Ours (ResNet50)	87.91	88.61	88.59	79.50
	Ours (ResNeXt50)	87.95	89.28	88.61	79.55
	Ours (ResNet101)	**87.77**	**90.04**	**89.12**	**80.58**

The bold values indicate the highest score for each evaluation metric.

(0.30% and 0.31%). Finally, when both group convolution and multiloss are combined into SSCFNet, he achieves the best accuracy for the LEVIR-CD dataset (94.89% and 90.30%). These results show that combining group convolution and multiloss can improve the model's performance.

F. Experimental Comparison and Analysis

To evaluate the superiority of the proposed SSCFNet network, we conducted a quantitative and qualitative comparison with six existing methods, including FC-Siam-Diff, FC-Siam-Conc, DTCDSCN, SNUNet, DSIFN, and MFPNet, on four datasets: LEVIR-CD, WHU, season-varying, and SYSU-CD. To more fully demonstrate the performance of our proposed method, we compared it with other methods using different backbones, such as ResNet50, ResNeXt50, and ResNet101, in the same experimental environment.

TABLE VI
COMPARISON STUDY OF MODEL EFFICIENCY

Method	Parameters (M)	FLOPs (G)
FC-Siam-Conc [42]	1.55	10.62
FC-Siam-Diff [42]	1.35	9.42
DTCDSCN [49]	41.07	40.76
SNUNet [51]	12.03	109.63
DSIFN [52]	50.44	164.53
MFP-Net [57]	85.97	257.52
Ours (ResNet50)	29.21	73.06
Ours (ResNeXt50)	28.68	72.12
Ours (ResNet101)	48.20	111.95

We report six different models and the proposed SSCFNet network for comparison of the number of parameters (parameters) and floating point operations per second (FLOPs). The input of the model is adjusted to 256 × 256 × 3 to compute floating point operations.

1) Quantitative Comparison: We first performed a quantitative comparison with the other six state-of-the-art methods in terms of Precision, Recall, F1 score, and MIoU, with higher F1 scores indicating better detection. Table V shows the quantitative comparison of the four datasets LEVIR-CD, WHU, season-varying, and SYSU-CD. It can be seen that the proposed network SSCFNet is better than the current state-of-the-art models in terms of F1 score and MIoU. Such results illustrate the effectiveness of fusing shallow features with deep features. To further verify the efficiency of our proposed method, a comparison was conducted with six other models, and the results are presented in Table VI. The scores of each model were evaluated using various metrics and compared against DSIFN and MFP-Net, which demonstrated comparable performance in Table V. Our method outperformed these two models while requiring a smaller number of parameters and less computational resources. Thus, our approach not only achieves better experimental results, but also offers efficiency advantages in terms of model size and computational complexity. In backbone ResNet101, which has the largest number of parameters, the number of parameters and computation of our method are reduced by 4.4% and 31.96%, respectively, compared to DSIFN, and by 43.93% and 56.53%, respectively, compared to MFP-Net. Therefore, our proposed network SSCFNet is the best.

2) Qualitative Comparison: Qualitative comparison results of change detection between the proposed method and other state-of-the-art methods are shown in Fig. 6. It is clear from the figure that our method has significant advantages over other methods in detail processing for small target detection. For large area change detection, the proposed method outperforms other methods in terms of detection continuity, accuracy, and is closer to the ground truth. Therefore, in terms of visual effects, our proposed method, SSCFNet, is superior to the best methods currently available.

V. Discussion

Our proposed remote sensing image change detection model, SSCFNet, effectively utilizes semantic feature information extracted from different levels of the backbone network. The model performed best in both quantitative and qualitative evaluations across four public remote sensing image change detection datasets. However, our model still has some areas for improvement. First, the loss function we currently use is relatively simple, and we plan to design a more unique loss function in future work to further improve detection accuracy for remote sensing image change detection problems. Second, our method currently only uses visual feature information, which may be relatively limited, so we will consider adding prior knowledge for multimodal related work in future work.

VI. Conclusion

This article proposes a novel remote sensing image change detection model, SSCFNet. To effectively utilize the different semantic feature information extracted from the lower and deeper layers of the backbone network, the SSCF is proposed. The SSCF can recombine the output features of each layer of the backbone network to achieve the effect of semantic feature enhancement, then add different attention mechanisms into each combination to obtain features with richer representational power. Finally, the obtained semantic features are cross-fused to enhance the global context information ability. To make the cross-fusion of the newly constructed semantic feature blocks more adequate, we introduce a multiloss strategy to assist the optimization. Extensive experiments on four public remote sensing image change detection datasets verify that our method achieves the best results in both quantitative and qualitative evaluations.

In future work, we intend to make improvements in the following two aspects: 1) we will consider adding a priori information to the proposed method to improve the accuracy of the model even further; 2) we will prepare to extend the method to semisupervised, weakly supervised, or unsupervised, considering that the labels of remote sensing change detection datasets are difficult to obtain.

References

[1] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” *Nature*, vol. 521, no. 75538, pp. 436–444, 2015.

[2] J. Schmidhuber, “Deep learning in neural networks: An overview,” *Neural Netw.*, vol. 61, pp. 85–117, 2015.

[3] X. Zhang, “Deep learning-based multi-focus image fusion: A. survey and a comparative study,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 9, pp. 4819–4838, Sep. 2022.

[4] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based learning applied to document recognition,” *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.

[5] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” *Adv. Neural Inf. Process. Syst.*, vol. 25, pp. 1097–1105, 2012.

[6] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” 2014, *arXiv:1409.1556*.

[7] 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.*, 2016, pp. 2818–2826.

[8] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[9] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, “Densely connected convolutional networks,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 4700–4708.

[10] R. Girshick, J. Donahue, T. Darrell, and J. Malik, “Rich feature hierarchies for accurate object detection and semantic segmentation,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2014, pp. 580–587.

[11] J. Long, E. Shelhamer, and T. Darrell, “Fully convolutional networks for semantic segmentation,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 3431–3440.

[12] F. Liu, L. Jiao, X. Tang, S. Yang, W. Ma, and B. Hou, “Local restricted convolutional neural network for change detection in polarimetric SAR images,” *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 3, pp. 818–833, Mar. 2019.

[13] H. Zhang, X. Tang, X. Han, J. Ma, X. Zhang, and L. Jiao, “High-resolution remote sensing images change detection with siamese holistically-guided FCN,” in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2021, pp. 4340–4343.

[14] H. A. T. Nguyen, T. Sophea, S. H. Gheewala, R. Rattanakom, T. Areerob, and K. Prueksakorn, “Integrating remote sensing and machine learning into environmental monitoring and assessment of land use change,” *Sustain. Prod. Consumption*, vol. 27, pp. 1239–1254, 2021.

[15] S. Hafner, A. Nascetti, H. Azizpour, and Y. Ban, “Sentinel-1 and sentinel-2 data fusion for urban change detection using a dual stream U-net,” *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2021, Art. no. 4019805.

[16] 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.

[17] L. Khelifi and M. Mignotte, “Deep learning for change detection in remote sensing images: Comprehensive review and meta-analysis,” *IEEE Access*, vol. 8, pp. 126385–126400, 2020.

[18] T. Liu et al., “Building change detection for VHR remote sensing images via local-global pyramid network and cross-task transfer learning strategy,” *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2021, Art. no. 4704817.

[19] C. Wu et al., “Building damage detection using U-net with attention mechanism from pre- and post-disaster remote sensing datasets,” *Remote. Sens.*, vol. 13, no. 5, p. 905, 2021.

[20] M. Yang, L. Jiao, B. Hou, F. Liu, and S. Yang, “Selective adversarial adaptation-based cross-scene change detection framework in remote sensing images,” *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 3, pp. 2188–2203, Mar. 2021.

[21] J. Ma, J. Jiang, H. Zhou, J. Zhao, and X. Guo, “Guided locality preserving feature matching for remote sensing image registration,” *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 8, pp. 4435–4447, Aug. 2018.

[22] D. Amitrano, R. Guida, and P. Iervolino, “Semantic unsupervised change detection of natural land cover with multitemporal object-based analysis on sar images,” *IEEE Trans. Geosci. Remote Sens.*, vol. 59, no. 7, pp. 5494–5514, Jul. 2021.

[23] X. Hai Wang, C. Xing, Y. Feng, R. Song, and Z. Mu, “A novel hyperspectral image change detection framework based on 3D-wavelet domain active convolutional neural network,” in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2021, pp. 4332–4335.

[24] R. Cao, L. Fang, T. Lu, and N. He, “Self-attention-based deep feature fusion for remote sensing scene classification,” *IEEE Geosci. Remote Sens. Lett.*, vol. 18, no. 1, pp. 43–47, Jan. 2021.

[25] G. Liu, Y. Gousseau, and F. Tupin, “A contrario comparison of local descriptors for change detection in very high spatial resolution satellite images of urban areas,” *IEEE Trans. Geosci. Remote Sens.*, vol. 57, no. 6, pp. 3904–3918, Jun. 2019.

[26] 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.

[27] L.-C. Chen, G. Papandreou, I. Kokkinos, K. P. Murphy, and A. L. Yuille, “Semantic image segmentation with deep convolutional nets and fully connected CRFs,”, 2015, *arXiv:1412.7062*.

[28] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, “Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2017.

[29] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, “Rethinking atrous convolution for semantic image segmentation,” 2017, *arXiv1706.05587*.

[30] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, “Encoder-decoder with atrous separable convolution for semantic image segmentation,” in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 801–818.

[31] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, “Pyramid scene parsing network,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2881–2890.

[32] Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei, and W. Liu, “Ccnet: Criss-cross attention for semantic segmentation,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2019, pp. 603–612.

[33] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, “Feature pyramid networks for object detection,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2117–2125.

[34] W. Liu et al., “SSD: Single shot multibox detector,” in *Proc. Eur. Conf. Comput. Vision.* Springer, 2016, pp. 21–37.

[35] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, “Learning deep features for discriminative localization,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 2921–2929.

[36] H. Zhang, K. Zu, J. Lu, Y. Zou, and D. Meng, “EpsaNet: An efficient pyramid squeeze attention block on convolutional neural network,” in *Proc. Asian Conf. Comput. Vis.*, 2022, pp. 1161–1177.

[37] L. Jiao, J. Gao, X. Liu, F. Liu, S. Yang, and B. Hou, “Multi-scale representation learning for image classification: A survey,” *IEEE Trans. Artif. Intell.*, vol. 4, no. 1, pp. 23–43, Feb. 2023.

[38] M. Tan and Q. V. Le, “Efficientnet: Rethinking model scaling for convolutional neural networks,” in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 6105–6114.

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

[40] 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.*, 2005, vol. 1 pp. 539–546.

[41] 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.

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

[43] 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.

[44] M. Zhang, G. Xu, K. Chen, M. Yan, and X. Sun, “Triplet-based semantic relation learning for aerial remote sensing image change detection,” *IEEE Geosci. Remote Sens. Lett.*, vol. 16, no. 2, pp. 266–270, Feb. 2019.

[45] S. Saha, F. Bovolo, and L. Bruzzone, “Unsupervised deep change vector analysis for multiple-change detection in VHR images,” *IEEE Trans. Geosci. Remote. Sens.*, vol. 57, no. 6, pp. 3677–3693, Jun. 2019.

[46] P. Chen, B. Zhang, D. Hong, Z. Chen, X. Yang, and B. Li, “Fccdn: Feature constraint network for VHR image change detection,” *ISPRS J. Photogrammetry Remote Sens.*, vol. 187, pp. 101–119, 2022.

[47] H. Chen, N. Yokoya, C. Wu, and B. Du, “Unsupervised multimodal change detection based on structural relationship graph representation learning,” *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art. no. 5635318.

[48] H. Chen, N. Yokoya, and M. Chini, “Fourier domain structural relationship analysis for unsupervised multimodal change detection,” *ISPRS J. Photogrammetry Remote Sens.*, vol. 198, pp. 99–114, 2023.

[49] Y. Liu, C. Pang, Z. Zhan, X. Zhang, and X. Yang, “Building change detection for remote sensing images using a dual-task constrained deep siamese convolutional network model,” *IEEE Geosci. Remote Sens. Lett.*, vol. 18, no. 5, pp. 811–815, May 2020.

[50] S. Fang, K. Li, J. Shao, and Z. Li, “SNUNet-CD: A. densely connected siamese network for change detection of VHR images,” *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2021, Art. no. 8007805.

[51] C. Zhang et al., “A deeply supervised image fusion network for change detection in high resolution bi-temporal remote sensing images,” *ISPRS J. Photogrammetry Remote Sens.*, vol. 166, pp. 183–200, 2020.

[52] S. Xiang, M. Wang, X. Jiang, G. Xie, Z. Zhang, and P. Tang, “Dual-task semantic change detection for remote sensing images using the generative change field module,” *Remote. Sens.*, vol. 13, no. 16, 2021, Art. no. 3336.

[53] J. Wu, C. Xie, Z. Zhang, and Y. Zhu, “A deeply supervised attentive high-resolution network for change detection in remote sensing images,” *Remote. Sens.*, vol. 15, no. 1, p. 45, 2022.

[54] J. Dong, W. Zhao, and S. Wang, “Multiscale context aggregation network for building change detection using high resolution remote sensing images,” *IEEE Geosci. Remote. Sens. Lett*., vol. 19, 2021, Art. no. 8022605.

[55] M. Papadomanolaki, S. Verma, M. Vakalopoulou, S. Gupta, and K. Karantzalos, “Detecting urban changes with recurrent neural networks from multitemporal sentinel-2 data,” in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2019, pp. 214–217.

[56] J. Xu, C. Luo, X. Chen, S. Wei, and Y. Luo, “Remote sensing change detection based on multidirectional adaptive feature fusion and perceptual similarity,” *Remote Sens.*, vol. 13, no. 15, 2021, Art. no. 3053.

[57] H. Zhang, M. Lin, G. Yang, and L. Zhang, “Escnet: An end-to-end superpixel-enhanced change detection network for very-high-resolution remote sensing images,” *IEEE Trans. Neural Netw. Learn. Syst*., vol. 34, no. 1, pp. 28–42, Jan. 2023.

[58] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, “Aggregated residual transformations for deep neural networks,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 1492–1500.

[59] J. Hu, L. Shen, and G. Sun, “Squeeze-and-excitation networks,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7132–7141.

[60] 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.

[61] 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.

[62] M. Lebedev, Y. V. Vizilter, O. Vygolov, V. Knyaz, and A. Y. Rubis, “Change detection in remote sensing images using conditional adversarial networks,” *Int. Arch. Photogrammetry, Remote Sens. Spatial Inf. Sci.*, vol. 42, no. 2, 2018.

[63] S. Ruder, “An overview of gradient descent optimization algorithms,” 2016, *arXiv:1609.04747*.

[64] A. Paszke et al., “Pytorch: An imperative style, high-performance deep learning library,” in *Proc. Adv. Neural Inf. Process. Syst.*, 2019.

[65] C. Lei, “Deep learning and practice with mindspore,” in *Cogn. Intell. Robot.*, 2021.

[66] A. Garcia-Garcia, S. Orts, S. Oprea, V. Villena-Martinez, and J. G. Rodríguez, “A review on deep learning techniques applied to semantic segmentation,” 2017, *arXiv:1704.06857*.

Jiahao Wang received the B.S. degree in computer science and technology from the North University of China, Taiyuan, China, in 2021. He is currently working toward the Ph.D. degree in computer science and technology with Xidian University, Xi’an, China.

His current research interests include remote sensing image processing and computer vision.

Fang Liu (Senior Member, IEEE) received the B.S. degree from Xi'an Jiaotong University, Xi'an, China, and the M.S. degree from Xidian University, Xi'an, China, in 1984 and 1995, respectively, both in computer science and technology.

She is currently a Professor with the School of Computer Science, Xidian University. Her research interests include signal and image processing, synthetic aperture radar image processing, multiscale geometry analysis, learning theory and algorithms, optimization problems, and data mining.

Prof. Liu was the recipient of the second prize of the National Natural Science Award in 2013.

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

Since 1992, he has been a Professor with the School of Electronic Engineering, Xidian University, Xi'an, China, where he is currently the Director of the Key Laboratory of Intelligent Perception and Image Understanding of the Ministry of Education of China. His research interests include image processing, natural computation, machine learning, and intelligent information processing.

Dr. Jiao is a Foreign Member of the Academia Europaea and the Russian Academy of Natural Sciences. He is a Fellow of IET, CAAI, CIE, CCF, and CAA. He is also a Councilor of the Chinese Institute of Electronics, a Committee Member of the Chinese Committee of Neural Networks, an Expert of the Academic Degrees Committee of the State Council, the Chairman of the Awards and Recognition Committee, and the Vice Board Chairperson of the Chinese Association of Artificial Intelligence.

Hao Wang received the B.S. degree in computer science and technology from the North University of China, Taiyuan, China, in 2021. He is currently working toward the Ph.D. degree in computer science and technology with Xidian University, Xi'an, China.

His current research interests include remote sensing image processing and computer vision.

Hua Yang received the B.S. degree in electronic information science and technology, in 2020, from Xidian University, Xi'an, China, where he is currently working toward the master's degree in electronic information.

His current research interests include intelligent remote sensing and computer vision.

Xu Liu (Member, IEEE) received the B.S. degrees in mathematics and applied mathematics from the North University of China, Taiyuan, China, and the Ph.D. degree in electronic science and technology from Xidian University, Xi'an, China, in 2013 and 2019, respectively.

He is currently an Associate Professor of Huashan elite and a Postdoctoral Researcher with the Key Laboratory of Intelligent Perception and Image Understanding of Ministry of Education, School of Artificial Intelligence, Xidian University. He is also the Chair of IEEE Xidian University Student Branch (2015–2019). His current research interests include machine learning and image processing.

Lingling Li (Senior Member, IEEE) received the B.S. degrees in electronic information engineering and Ph.D. degree in intelligent information processing from Xidian University, Xi'an, China, in 2011 and 2017, respectively.

From 2013 to 2014, she was an Exchange Ph.D. Student with the Intelligent Systems Group, Department of Computer Science and Artificial Intelligence, University of the Basque Country (UPV/EHU), Leioa, Spain. She is currently a Postdoctoral Researcher with the School of Artificial Intelligence, Xidian University. Her current research interests include quantum evolutionary optimization, machine learning, and deep learning.

Puhua Chen (Senior Member, IEEE) received the B.S. degree in environmental engineering from the University of Electronic Science and Technology of China, Chengdu, China, and the Ph.D. degree in circuit and system from Xidian University, Xi'an, China, in 2009 and 2016, respectively.

She is currently an Associate Professor with the School of Artificial Intelligence, Xidian University. Her current research interests include machine learning, pattern recognition, and remote sensing image interpretation.