IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received February 16, 2022, accepted March 12, 2022, date of publication March 16, 2022, date of current version March 24, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3160163

MSF-Net: A Multiscale Supervised Fusion Network for Building Change Detection in High-Resolution Remote Sensing Images

JIAHAO CHEN[1], JUNFU FAN[1,2,3], MENGZHEN ZHANG[1], YUKE ZHOU[4], AND CHEN SHEN[2,3]

[1]School of Civil and Architectural Engineering, Shandong University of Technology, Zibo, Shandong 255000, China
[2]Shandong Tianyunhe Information Technology Company Ltd., Zibo, Shandong 255000, China
[3]High-Resolution Earth Observation System Data and Application Center of Zibo, Zibo, Shandong 255000, China
[4]Ecology Observing Network and Modeling Laboratory, Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing 100101, China

Corresponding authors: Junfu Fan (fanjf@sdut.edu.cn) and Yuke Zhou (zhouyk@igsnrr.ac.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 42171413 and Grant 42101306, in part by the Shandong Provincial Natural Science Foundation under Grant ZR2020MD015 and Grant ZR2020MD018, in part by the National Key Research and Development Program of China under Grant 2021XJKK0303, and in part by the Young Teacher Development Support Program of Shandong University of Technology under Grant 4072-115016.

ABSTRACT Building change detection is a primary task in the application of remote sensing images, especially in city land resource management and urbanization process assesment. Due to the rich textural features of remote sensing images and the multiscale characteristics of buildings, it is still a huge challenge to effectively filter out irrelevant change information (e.g., roads) and fuse multiscale building features. To date, deep learning-based methods have demonstrated powerful capabilities in this field. To fill these gaps, this study proposes a multiscale supervised fusion network (MSF-Net), which is an attention mechanism-based approach for building change detection using bi-temporal high-resolution satellite imagery. Especially, we built a dual-context fusion module to obtain abundant global context information of buildings and suppressing irrelevant features. We also used channel attention mechanism, selective kernel convolution and multiscale supervision module to fuse multiscale feature of buildings. The ablation experiments verified the availability of these modules. The MSF-Net model has been tested on the LEVIR-CD building change detection dataset. Compared with other state-of-the-art change detection methods, the study showed that our method obtained 0.8866 and 0.8130 in F1-score and Intersection over Union (IOU), respectively. The results indicate that the MSF-Net method has stronger multiscale building feature extraction capability and suppression ability of irrelevant features, which could produce clearer building boundaries and more accurate building change maps.

INDEX TERMS Remote sensing, building change detection, deep learning, attention mechanism, multiscale feature.

I. INTRODUCTION

Building change detection is a task that makes use of satellite-based remote sensing images of the same area in different periods, to identify the generation or disappearance of building objects. Since Weismiller proposed the image difference method for coastal zone environmental monitoring [1], remote sensing change detection has been developed for more than forty years and has been playing an important role in land surveys [2], [3], natural environment monitoring [4], [5], disaster assessment [6] and urban research [7], [8]. High-resolution remote sensing images have become an important data source for change detection, due to their rich ground object texture features and ground object multiscale features. Meanwhile, how to fuse multiscale features of ground objects and remove irrelevant ground objects change information are huge challenges.

There are two main branches of change detection: traditional methods and deep learning-based methods [9]. Traditional change detection methods can be divided into

The associate editor coordinating the review of this manuscript and approving it for publication was Yongjie Li.

three classes: methods based on image arithmetic, methods based on image transformations and methods based on post-classification. Image arithmetic methods include image difference [10], image ratio [11], Change Vector Analysis (CVA) [12], [13], among others. These methods obtain a feature map by subtraction or division, and then determine a segmentation threshold to generate the change map. They ignore the context information and cause a lot of salt-and-pepper noise [14]. Methods based on image transformation include Principal Component Analysis (PCA) [15], Multivariate Alteration Detection (MAD) [16], Iteratively Reweighted Multivariate Alteration Detection (IRMAD) [17], Kernel Slow Feature Analysis (KSFA) [18], etc. They transform the image into a specific feature space, highlight the changed area and suppress the unchanged area. However, it is a hard process to choose the most appropriate method for different area. The post-classification change detection method, classify the ground objects of the images firstly, then generates the change maps by comparative analysis of these images. The change detection accuracy depends on the classification accuracy, and the classification process requires manual labeling of samples. With the continuous launching of high resolution optical remote sensing satellites (WorldView-3, GF-2), the accuracy of remote sensing images is increasing, the huge amount of data and the fine and complex texture features in high-resolution images bring new challenges to traditional change detection [19]–[21].

Due to the excellent processing capability of big data, deep learning [22], [23] is becoming more widely applied in remote sensing [24]–[26], and achieved excellent results in change detection [27]–[32]. Daudt *et al.* proposed Fully Convolutional Early Fusion (FC-EF) [33] based Fully Convolutional Networks (FCN) [34], Peng *et al.* proposed Unet++ Multiple Side-Outputs Fusion network (Unet++_MSOF) [35] based on Unet++ [36], [37]. They combine two single-temporal images with channel number C into 2C data in channel-wise, then input the data into the network to identify the changed areas. They did not extract the depth feature of the single temporal image, limit the change detection accuracy. To solve this problem, Zhan *et al.* used a deep Siamese convolutional network to detect changes in optical aerial images [38], the network simultaneously extracts the feature information of two images, generates the distance map, and finally obtains the change map through threshold segmentation, which achieved good results. Daudt *et al.* first proposed end-to-end fully convolutional Siamese networks for change detection, named Fully Convolutional Siamese-Concatenation (FC-Siam-conc) and Fully Convolutional Siamese-Difference (FC-Siam-diff) [33]. Thereafter, dual-task constrained deep Siamese convolutional network (DTCDSCN) [39], pyramid feature-based attention-guided Siamese network (PGA-SiamNet) [40], Siamese NestedUNet Networks [41], NestNet [20] and others have been proposed, improve the accuracy of change detection.

Building change detection need filter irrelevant changes, the network should pay attention to building information and suppresses other information. The wide application of attention mechanism [42]–[45] in deep neural network (DNN) [46]–[50] brings further inspiration. Chen *et al.* built a dual attentive fully convolutional Siamese networks (DASNet) [14], DASNet uses the dual-attention mechanism to reconstruct the features of two single temporal images separately, it can obtain long-range dependencies and more differential feature representations. Jiang proposed an attention-guided Siamese network (PGA-SiamNet) based on a pyramidal structure [40], it captures possible variations using a convolutional neural network in the pyramid. The global co-attentive mechanism was introduced to emphasize the importance of correlations between input feature pairs. Zhang *et al.* proposed deeply supervised image fusion network (DSIFN) [19], introduced a spatial attention mechanism (SAM) [51] and a channel attention mechanism (CAM) [52] in the network. The attention module fuse the depth original image features with the image difference features, effectively improved the accuracy of the change map. The deeply supervised attention metric-based network (DSAM-Net) [53] and attention-based deeply supervised network (ADS-Net) [54], introduced the convolutional block attention module (CBAM) [55], [56] for feature reconstruction of scale information, also achieved good results.

Remote sensing images have multiscale features, to fuse them effectively, multiscale fusion module is carried out in the change detection network. PSPNet-CONC [57] introduces the Pyramid Scene Parsing (PSP) module [58] for multiscale feature extraction. Unet++_MSOF and NestNet perform multiscale feature fusion based on the dense skip module of the Unet++ network, they also introduce a multi-output fusion strategy at the output module to improve the detection accuracy. ADS-Net fuses the scale features corresponding to other branches in the decoding module, calculates the F1-score of each scale output, and performs the weighted fusion of change maps based on the F1-score, to further improve the accuracy of building change detection. Fang et al [59] proposed SNUNet-CD, which improved the UNet++ network structure by fusing its four outputs. It can generate a change map with multiscale information through the Ensemble Channel Attention Module (ECAM) attention module in the output stage of the network.

Taking the LEVIR-CD dataset [60] as an example, the state-of-the-art network achieves a maximum of 88% and 79% in F1-score and IOU, and there is still some space for improvement. In order to better fuse multiscale feature of buildings and remove irrelevant ground objects change information, we proposed MSF-Net: A building change detection network based on attention mechanism. We enhance the building feature extraction capability and filter out irrelevant information by introducing multiple attention mechanisms, it can make MSF-Net focus on the building information. To enhance the multiscale feature fusion capability, we introduced multiscale fusion module

and multiscale supervision module. These modules also make the detected building boundaries more complete. We used the LEVIR-CD building change detection dataset to verify the accuracy of our model.

II. MATERIALS AND METHODS

A. NETWORK STRUCTURE

The MSF-Net model consists of encoding module (Fig. 1(a)), decoding module and output module (Fig. 2). Specifically, in the encoding module, five different scales of building information are extracted from bi-temporal remote sensing images simultaneously. Then, the extracted feature information of five different scales are sent to five dual-context fusion module respectively, to obtain the image difference feature map. Finally, the image difference feature map is sent to the decoding module to extract the building change information. The decoding module reconstructs the feature maps by fusing multiscale features through the channel attention mechanism, then generates four building change feature output maps with the same size but different scale information. In the output module, the four building change feature output maps are merged using the channel attention mechanism, to obtain the final building change map of the network, while a multiscale supervision strategy is added in the output module to improve the accuracy of building change detection.

1) ENCODING MODULE

The encoding module is composed of Siamese network with shared weights, each branch of which has five layers of encoding blocks (Fig. 1(a)). Encode1 includes two 3×3 convolutional layers, while encode2 is composed of a 2×2 max pooling downsampling layer and two 3×3 convolutional layers. To enhance the extraction ability of deep building features, encode3-5 are all composed of a 2×2 max pooling downsampling layer and three 3×3 convolutional layers. In the five layers of encoding blocks, the BatchNorm regularization and ReLU activation functions are added after each convolution operation. It is worth noting that most state-of-the-art network direct uses a difference operation or channel series operation to fuse bi-temporal images. However, MSF-Net inputs the bi-temporal feature map into dual-context fusion module for bi-temporal images fusion, to focus on the characteristic information of buildings, as shown in Fig. 1(b). Specifically, in the dual-context fusion module, two context modules extract the global building context information and local building context information of the single-phase image respectively, as shown in Fig. 1(c). Then the global and local building context information are sent to the two channel attention mechanism respectively, to obtain the important features of buildings on the channel dimension. Finally, the difference operation is performed, to obtain the difference feature map of buildings in the bi-temporal images, and reduce the difficulty of feature extraction of changing buildings in the decoding module.

In Fig. 1, (a) represents the basic structure of encoding module, Dual-context Fusion1 - 5 represents dual-context fusion module as shown in (b). Each dual-context fusion module is composed of the context module (c) and feature fusion module, feature fusion module is composed of two CAM modules. The CAM module in (c) is shown in Fig. 3. Scale1 - 5 represents five different scales of output.

2) DECODING MODULE

The decoding module of MSF-Net consists of four branches, each of which has a different number of decoding layers according to the different input feature scales (Fig. 2). Except for the last layer of decoding, the feature map size is restored through the upsampling module after each decoding layer. The features corresponding to the same scale in the encoding module are fused through the channel attention mechanism, to recover the original edge feature information of the building. After decoding at the last layer, the selective kernel convolution mechanism [61] is used for the final feature extraction of changed buildings, and the network learns the appropriate convolutional kernel size by itself to enhance the feature extraction effect of different scales. In the decoding module, each decoding layer is composed of a 3×3 convolution and the BatchNorm regularization and ReLU activation functions. Each upsampling module consists of bilinear interpolation and 1×1 convolution. The channel attention mechanism module is shown in Fig. 3; the selective kernel convolution module is shown in Fig. 4.

In Fig. 2, Scale1-5 are the five different scales of encoding output, the blue arrow represents the decoding layer of the network. CAM represents the channel attention mechanism (Fig. 3), SK represents the selective kernel convolution (Fig. 4).

3) OUTPUT MODULE

A multi-output fusion module is designed in MSF-Net, as shown at the bottom of Fig. 2. After the SK decoding layer of each branch, the building change map of each branch is obtained through 1×1 convolution layer. To integrate the building change characteristics of each branch, MSF-Net performs a BatchNorm regularization operation on the outputs of the four branches. Then, the four outputs are combined into multiscale feature maps according to the channel. Finally, the building change results with different scales are given different weights through the channel attention module, and the final building change map of the network is generated through 3×3 convolution. The BatchNorm regularization operation is added after the branch output, to reduce the over-fitting phenomenon caused by continuous convolution in the output module. In the network output module, a multi-output supervision strategy is added to calculate the loss of four branch outputs and final outputs with different weights, to improve the multiscale building detection capability of the network.

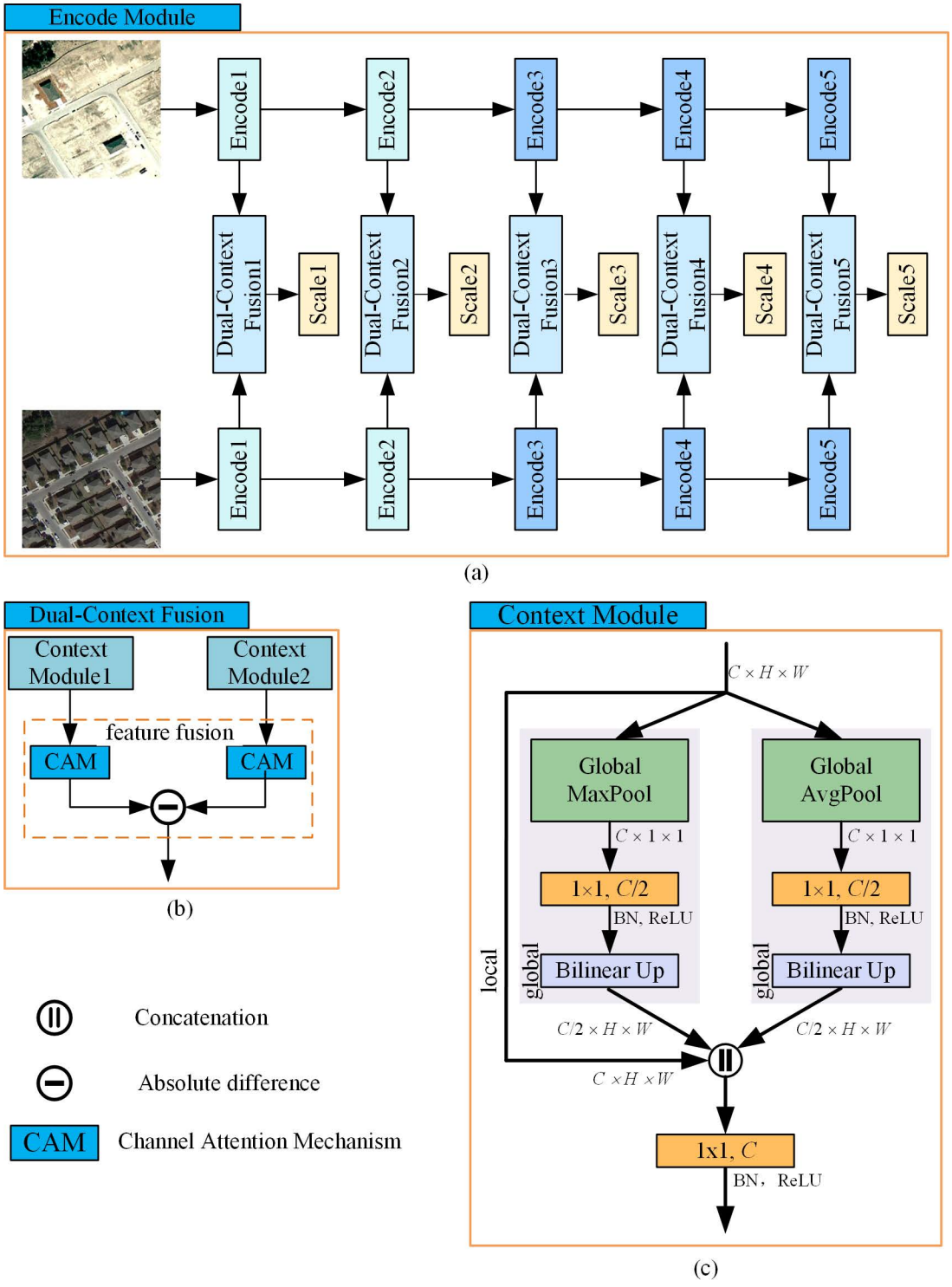


FIGURE 1. Diagram of Encode Module. (a) basic structure of encoding module, (b) dual-context fusion module, (c) context module.

B. DUAL-CONTEXT FUSION MODULE

The dual-context fusion module consists of two parts, context module and feature fusion module. The context module consists of a local context information branch and two global context information branches. The local context information branch retains original information. In the two global context information branches, firstly, the 1×1 feature map is obtained by global max pooling and global average

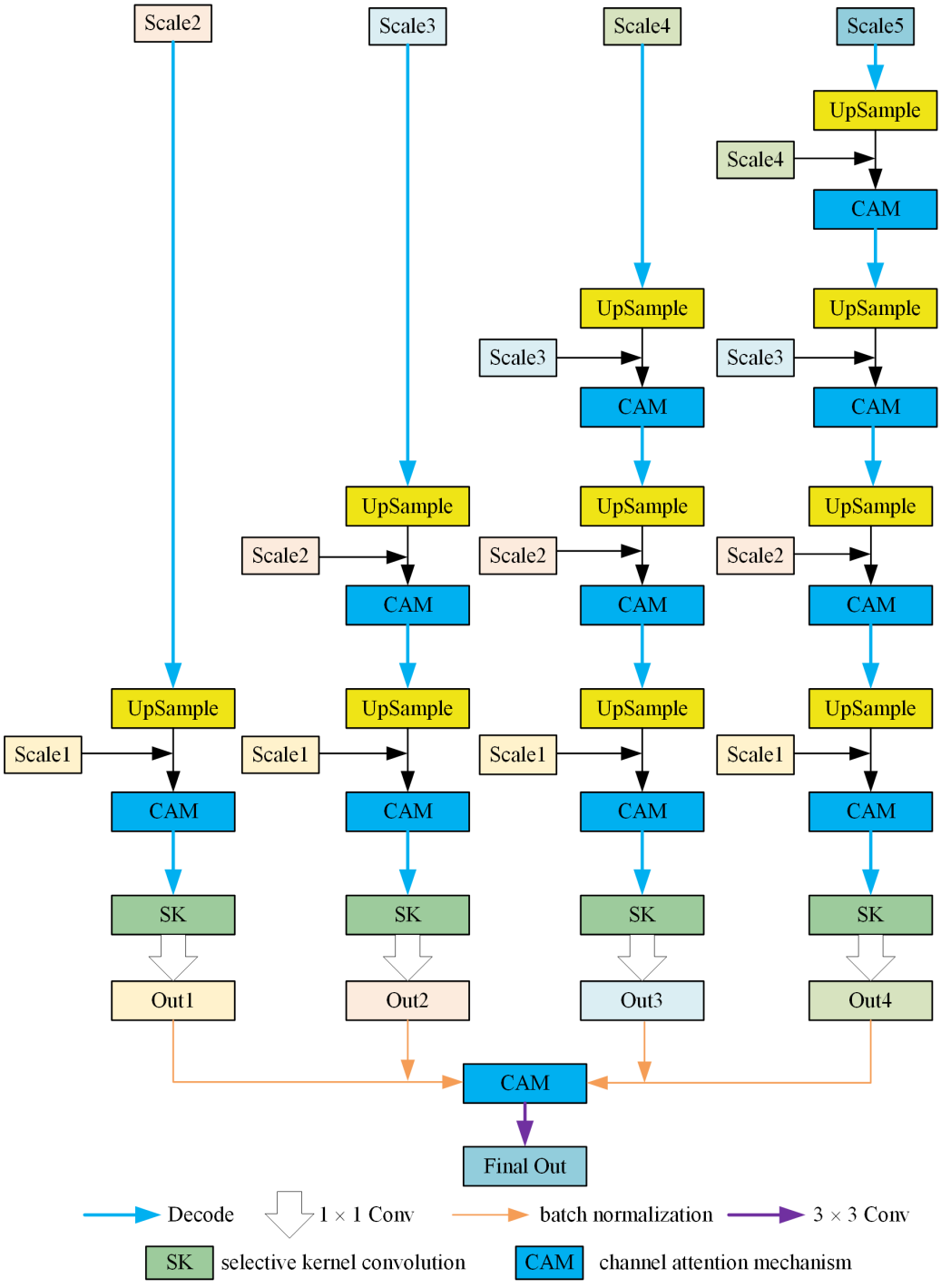


FIGURE 2. Diagram of decode module.

pooling respectively. After 1 × 1 convolution, BatchNorm regularization operation and ReLU activation function, it is passed into the bilinear interpolation upsampling module, to recover the feature map with the same size as the input feature size. Finally, the concatenation operation is used to merge the information of the three branches according to the channel, and then through 1 × 1 convolution, BatchNorm regularization operation and ReLU activation function, a single temporal image feature map with context information is generated. The feature fusion module is composed of two channel attention mechanism. The bi-temporal images feature maps processed by the context module are passed into the two channel attention mechanism branches, and after obtaining important information on each

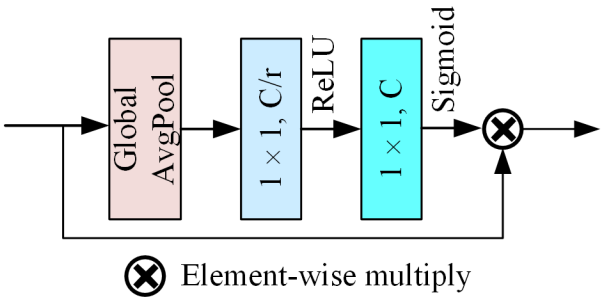


FIGURE 3. Flowchart of channel attention mechanism.

temporal image channel dimension, the difference operation is performed to obtain the bi-temporal images building difference feature maps.

C. CHANNEL ATTENTION MECHANISM

A multiscale feature fusion on the channel dimension introduces some unnecessary features, which adds difficulty to the extraction of changing buildings. To make the network focus on important channel information and suppress unimportant channel information, MSF-Net introduces a channel attention mechanism. The process of the whole channel attention mechanism is shown in Fig. 3. First, the feature map of size (H × W × C) is changed into a global feature vector of size (1 × 1 × C) using the global average pooling operation (H denotes the height of the feature map, W denotes the width of the feature map, and C denotes the number of feature map channels), and then two two-dimensional convolution operations are performed on the global feature vector to make it have nonlinear features. The size of the convolution kernel for both convolution operations is 1 × 1, and the number of convolution kernels is C/r and C, respectively (r denotes the scaling ratio, and r is 16 in the dual-context and decoding module, and r is 2 in the output fusion module). The results after the second convolution are passed through the Sigmoid activation function to obtain the weight coefficients of different channels, which are multiplied with the original feature map to obtain the feature map with channel attention.

D. SELECTIVE KERNEL CONVOLUTION

In a standard convolutional network, the receptive field size (convolutional kernel size) of each layer of the network neurons is the same. However, in MSF-Net, the four decoding branches have different scales of changing buildings. To make each branch find the perceptual field size suitable for the given scale, MSF-Net introduces a selective kernel convolution module after the last decoding layer, which can adjust the convolutional kernel size adaptively to improve the ability of building change detection. The flow of the convolutional kernel selection module is shown in Fig. 4. First, the input feature map is convolved through three convolutional branches to generate three scales of feature maps, each convolutional branch consists of two-dimensional convolution, BatchNorm regularization, and ReLU activation function. To reduce the number of parameters, the 5 × 5 convolution is replaced with the dilated convolution with a 3 × 3 kernel and dilation size 2, while 7 × 7 convolution is replaced with the dilated convolution with a 3 × 3 kernel and dilation size 3. After that, the three scales of the feature maps are summed up to obtain the multiscale feature map with three scales of information. Next, the global average pooling operation is performed on the multiscale feature map to obtain the channel-wise feature vectors, and then the feature vectors are squeezed using a two-dimensional convolution operation with a convolution kernel size of 1 × 1, the squeezed feature vectors are then subjected to three separate feature excitation operations using a 1 × 1 convolution operation, and finally the channel weight coefficients of each branch are obtained by the Softmax activation function. The weighting coefficients of the three branches are multiplied with the corresponding three scale feature maps to obtain the weighted three-branch feature maps. Then add the three-branch weighted feature maps to generate the output feature maps of each decoding branch of MSF-Net.

In Fig. 4, gp represents global average pooling operation, fc1 and fc2 represents fully connected operations with convolution kernel of 1 × 1, sm represents the softmax activation function.

E. LOSS FUNCTION

Loss function is an important component of DNN. In the remote sensing building change detection task, only a small portion of buildings usually change, which creates a problem of extreme imbalance between positive and negative samples. In order to overcome the negative impact of sample imbalance on detection accuracy, two loss functions (Binary Cross Entropy Loss and Dice Loss) were selected for loss calculation in this paper. The effectiveness of this combination has been demonstrated in the literature [19], [20]. For our multi-output network structure, we define a new loss function calculation formula. First, we calculate the loss of each output (Out1-4 and Final Out), and then sum them according to different weights to get the final loss. Finally the back propagation [62]–[64] is applied to optimize the model parameters. The total loss function is defined as follows:

$$L = \sum_{i=1}^{5} \omega_i L_{branch}^{i} + L_{final\_out} \tag{1}$$

where ω_i represents the loss weight of i^{th} branch output, after many experiments, we assigned five weights as (0.5, 0.5, 0.75, 0.75, 1). L_{branch}^{i} and L_{final_out} represents the output loss and final output loss of each branch respectively. Each branch is composed of Binary Cross Entropy Loss and Dice Loss, as shown in Equation (2):

$$L_{branch}^{i} = L_{bce}^{i} + L_{dice}^{i} \tag{2}$$

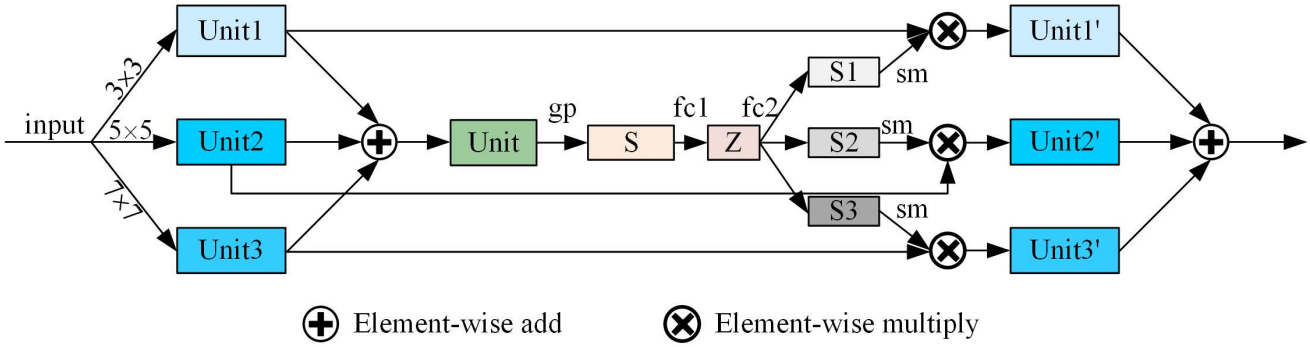


FIGURE 4. Flowchart of selective kernel convolution.

1) BINARY CROSS ENTROPY LOSS

Due to the superiority of Binary Cross Entropy Loss in handling unbalanced classification samples, we use it as part of the loss function, which is calculated as:

$$L_{bce} = -\frac{1}{N}\sum_{n=1}^{N}[y_n \log \sigma(x_n) + (1 - y_n)\log(1 - \sigma(x_n))] \tag{3}$$

where N represents the total number of pixels, y_n represents the true value of pixel point N(0 or 1, 0 represents unchanged pixel and 1 represents changed pixel), σ represents the sigmoid activation function, x_n represents the predicted value of pixel point N.

2) DICE LOSS

Dice Loss can improve segmentation performance and weaken the effect of class imbalance problems. It can measure the similarity between the forecast map and the real value on the ground, and when combined with the binary cross entropy Loss, can improve the stability of training losses. Dice Loss is calculated as follows:

$$L_{dice} = 1 - \frac{2 \cdot Y \cdot \text{softmax}(X)}{Y + \text{softmax}(X)} \tag{4}$$

where Y denotes the true value and X denotes the predicted value.

III. EXPERIMENTS

A. DATASETS AND PREPROCESSING

To verify the availability of the proposed network, we used the LEVIR-CD building change detection dataset to compare it with the other state-of-the-art networks. The dataset consists of 637 pairs of high-resolution remote sensing images with a size of 1024 × 1024 pixels and a resolution of 0.5m/pixel. It covers various types of architectural changes, such as villas and large garages, taken at different times from 2002 to 2018 in Texas, USA. To reduce the pressure on the GPU memory, we clip the dataset into 256 × 256 image pairs, and used the same method as the original paper to segment the data sets, generating a total of 7120 pairs of training data sets, 1024 pairs of validation data sets, and 2048 pairs of test data sets. To prevent over-fitting in the train process and enhance the robustness of the network, a random data enhancement was carried out on the training set during training.

B. PARAMETER SETTING

MSF-Net is built with the Pytorch library and the programming environment is PyCharm. The training period is set to 150 epochs in the training process, the batch size is 10, and the initial learning rate is 0.001. The learning rate is adjusted using the equal-interval adjustment strategy (StepLR), which is reduced to half of the original rate every 10 epochs of iteration. The experiments were run on a workstation with AMD Ryzen 9 5950X 16-Core (3.4GHz, 128RAM) CPU and Nvidia GeForce RTX 3090 (24GB) GPU.

C. EVALUATION METRIC

The accuracy of the model was evaluated using Precision (P), Recall (R), F1-score (F1), and IOU. Precision reflects the precision of the model to detect change pixels, the higher the Precision, the higher the correct rate of change pixels detected by the model. Recall reflects the check-all rate of the model, the higher Recall, the more change pixels the model detects. F1-score takes both indicators into account, so a higher F1-score indicates a better model. IOU indicates the overlap rate between the change map and the ground truth, a higher IOU means a better detection for the model. The metrics are calculated as shown below:

$$P = \frac{TP}{TP + FP} \tag{5}$$

$$R = \frac{TP}{TP + FN} \tag{6}$$

$$F_1 = \frac{2P \times R}{P + R} \tag{7}$$

$$IOU = \frac{TP}{TP + FP + FN} \tag{8}$$

where TP represents the number of pixels that have changed and are predicted to change, TN represents the number of pixels that have not changed and are predicted not changed,

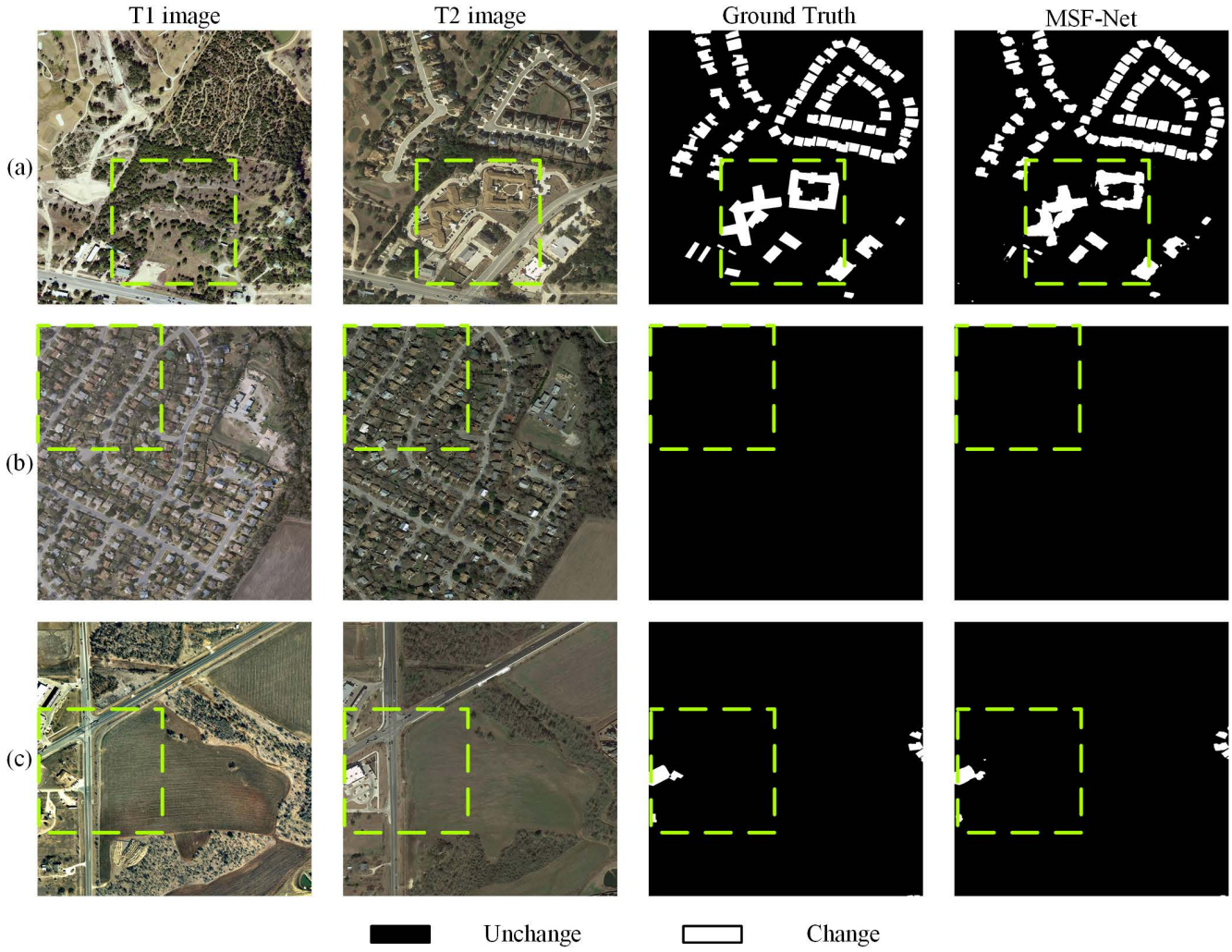


FIGURE 5. Test results of MSF-Net in three scenes.

FP represents the number of pixels that have not changed but are predicted to change, and FN represents the number of pixels that have changed but are predicted to remain unchanged.

D. COMPARISON METHODS

We selected the classical change detection networks and the state-of-the-art network to conduct a comparative test on the proposed MSF-Net.

(1) Unet++_MSOF: the Unet++ model is improved by adding a multiple side-outputs fusion module to improve the accuracy of change detection.
(2) Siam-conc: Combining the Unet++ network with the Siamese structure, the bi-temporal images are processed using concatenation operations in the encoding stage and then fed into the decoder for feature extraction.
(3) Siam-diff: Proposed at the same time as Siam-conc, it differs from it by using a difference operation to process the bi-temporal images in the encoding stage and then feeding it to the decoder for feature extraction.
(4) Siam-conc-diff: Proposed simultaneously with the first two networks, the difference with them is that the bi-temporal images and their differences are concatenated before being fed into the decoder for feature extraction.
(5) STA-Net: A Siamese network based on the spatial-temporal attention neural network, introducing a basic spatial–temporal attention module and a pyramid spatial–temporal attention module, it can captures long-range spatial–temporal dependencies and learns multi-scale feature information through the pyramid structure.
(6) SNUNet-CD: Based on Siamese UNet++, an Ensemble Channel Attention Module (ECAM) is added to combine multi-branch outputs into one output, and the ECAM module can refine the most representative features at different semantic levels to improve detection accuracy.
(7) NestNet: Using absolute different operation to process bi-temporal remote sensing images at each scale, the

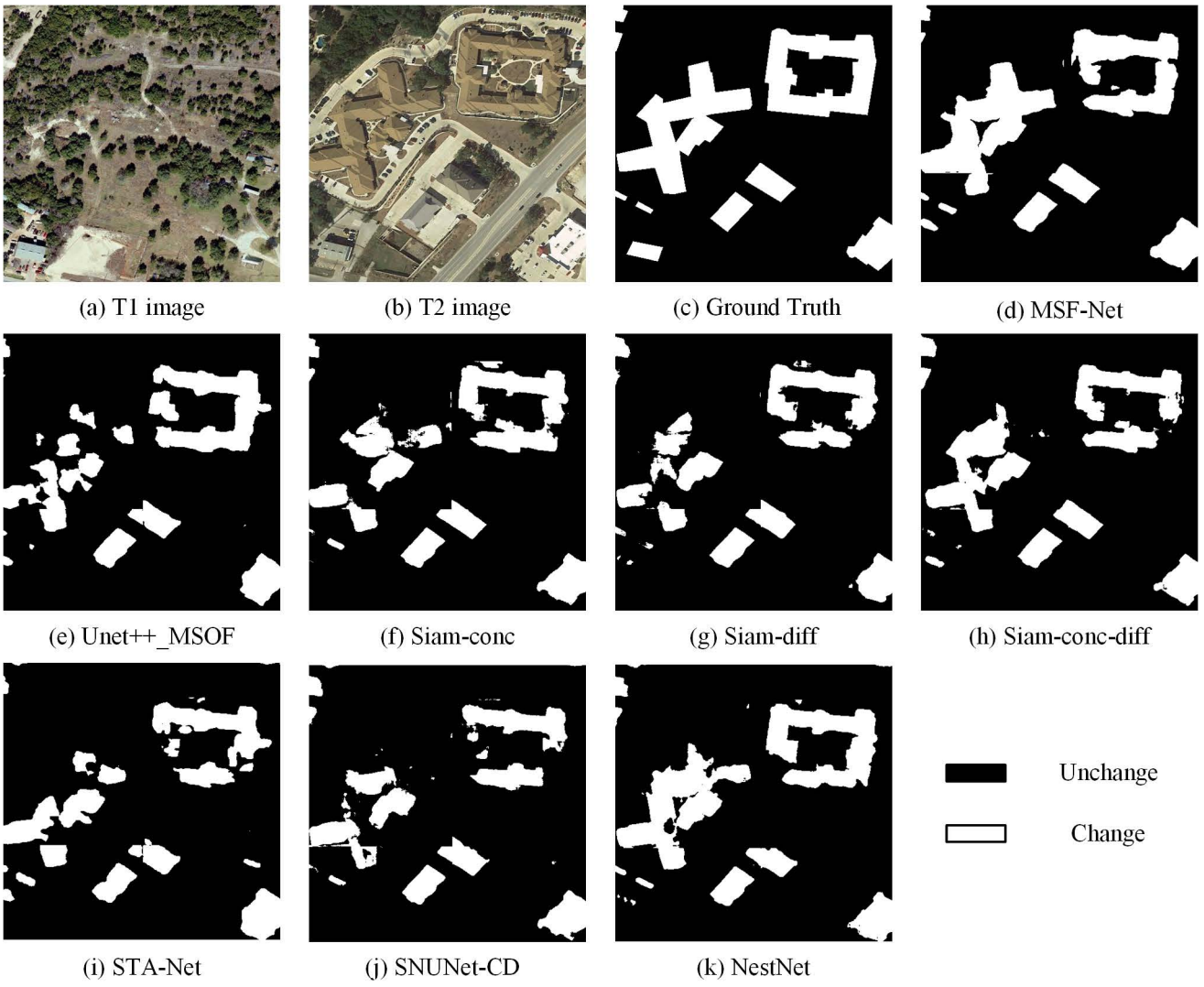


FIGURE 6. Comparison of building change detection with different networks for scene (a) in Fig. 5.

dense skip connections module is redesigned based on the Unet++ model to learn multiscale feature information.

E. COMPARISON OF EXPERIMENTAL RESULTS

We selected three scenes to test the accuracy of the MSF-Net, and the test results are shown in Fig. 5.

The experimental result shows that MSF-Net get better performance in detecting building boundary integrity, as shown in scene (a). Scene (b) shows MSF-Net can avoid false detection caused by light and other factors, filter out irrelevant information. Scene (c) shows MSF-Net can accurately distinguish buildings from roads and identify changes in small buildings. It also shows that MSF-Net can detect different scale building changes.

To further verify the effectiveness of the MSF-Net proposed in this paper, we conducted comparison experiments between the above seven methods and our proposed method on the LEVIR-CD building change detection dataset, and analyzed the effectiveness of the proposed network in the detection of building changes in high-resolution remote sensing images in both qualitative and quantitative aspects. The results of the qualitative analysis of the comparison experiments for the scenes in the three dashed boxes in Fig. 5 are shown in Fig. 6-8, and the results of the quantitative analysis are shown in Table 1.

Table 1 shows MSF-Net performs better than the other state-of-the-art networks. Although Recall is 2.82% less than for STA-Net, Precision, F1-score and IOU are 7.1%, 1.34% and 3.84% higher than it. Compared to other networks with the best results Siam-conc-diff, Precision, Recall, F1-score, and IOU are 0.39%, 2.84%, 2.34%, and 2.50% higher, respectively. F1-score and IOU of MSF-Net are 3.37% and 4.29% higher than those of Unet++_MSOF, indicating that early fusion affects the extraction of deep features of buildings and reduces the accuracy of building change detection.

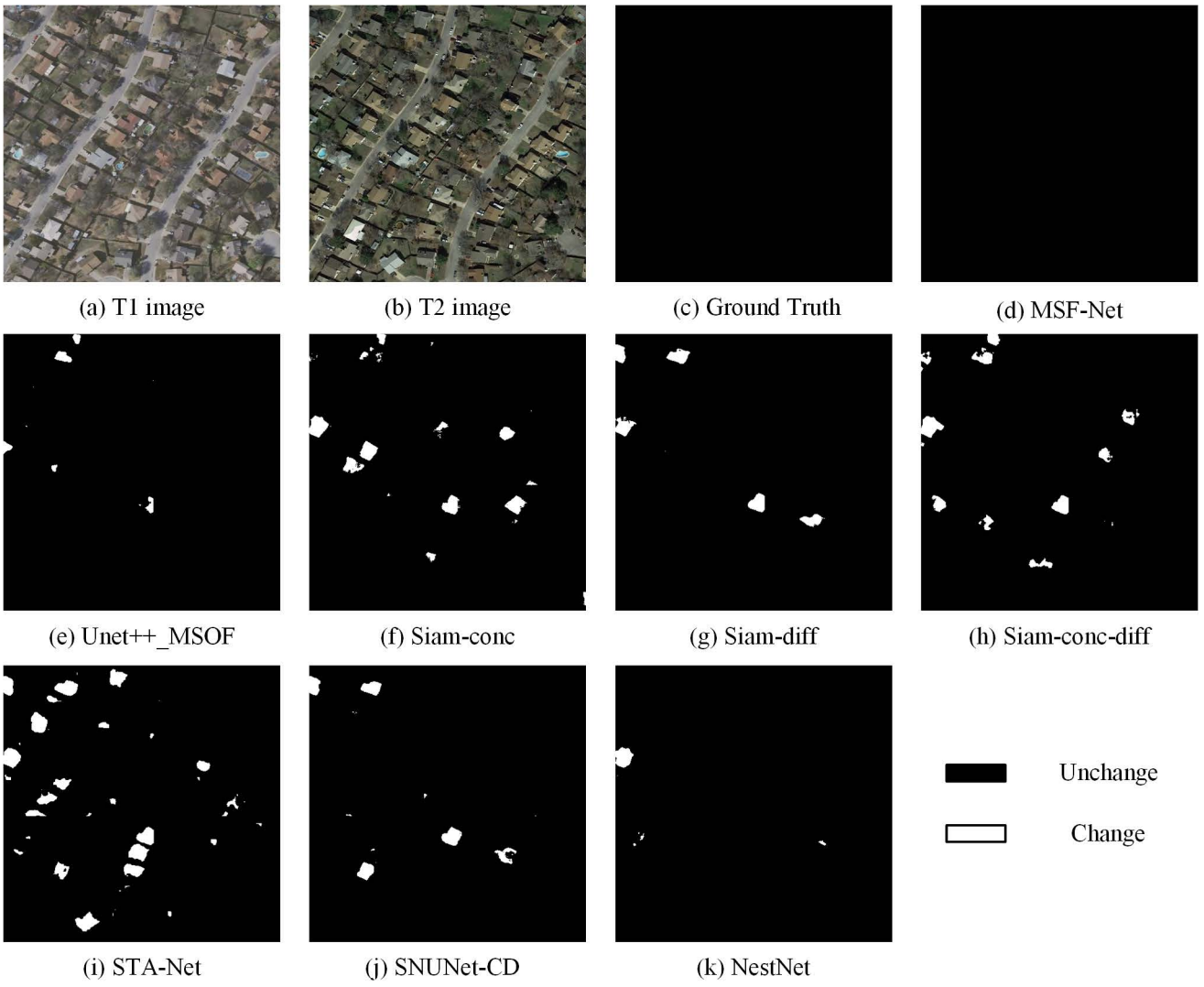


FIGURE 7. Comparison of building change detection with different networks for scene (b) in Fig. 5.

TABLE 1. Experimental results on the LEVIR-CD building change detection dataset.

Methods	Precision	Recall	F1-score	IOU
Unet++ _ MSOF	0.8950	0.8412	0.8529	0.7701
Siam-conc	0.8948	0.8596	0.8624	0.7849
Siam-diff	0.9019	0.8483	0.8579	0.7811
Siam-conc-diff	0.9053	0.8542	0.8632	0.7880
STA-Net	0.8382	0.9108	0.8732	0.7746
SNUNet-CD	0.8955	0.8565	0.8593	0.7803
NestNet	0.8930	0.8596	0.8608	0.7800
MSF-Net	0.9092	0.8826	0.8866	0.8130

The comparison between Siam-conc, Siam-diff, Siam-conc-diff and MSF-Net shows that the dual-context fusion module can accurately obtain the global building context information in the encoding stage, and avoid the influence of factors such as light on building feature extraction. For STA-Net, the spatial–temporal attention module was added to emphasize the characteristics of changed areas. Although the Recall of the model was greatly improved, the Precision of the change pixels of buildings was reduced. The comparison between SNUNet-CD and NestNet and MSF-Net shows that multi-output fusion helps the model detect multiscale changed architectural features and improve detection Precision.

MSF-Net has improved the boundary detection capability due to the fusion of multiscale raw features, and detected clear and more complete boundaries of changing buildings (Fig. 6). The Siam-conc-diff encoding module concatenates bi-temporal features and their differences feature maps, NestNet uses absolute different operation to process bi-temporal images, and both have achieved good results. Unet++_MSOF lost the original feature information due to the use of early fusion, the encoding modules of Siam-conc, Siam-diff, STA-Net and SNUNet-CD lost a lot of original information due to simple bi-temporal image concatenating and difference operations, resulting in poorer change building

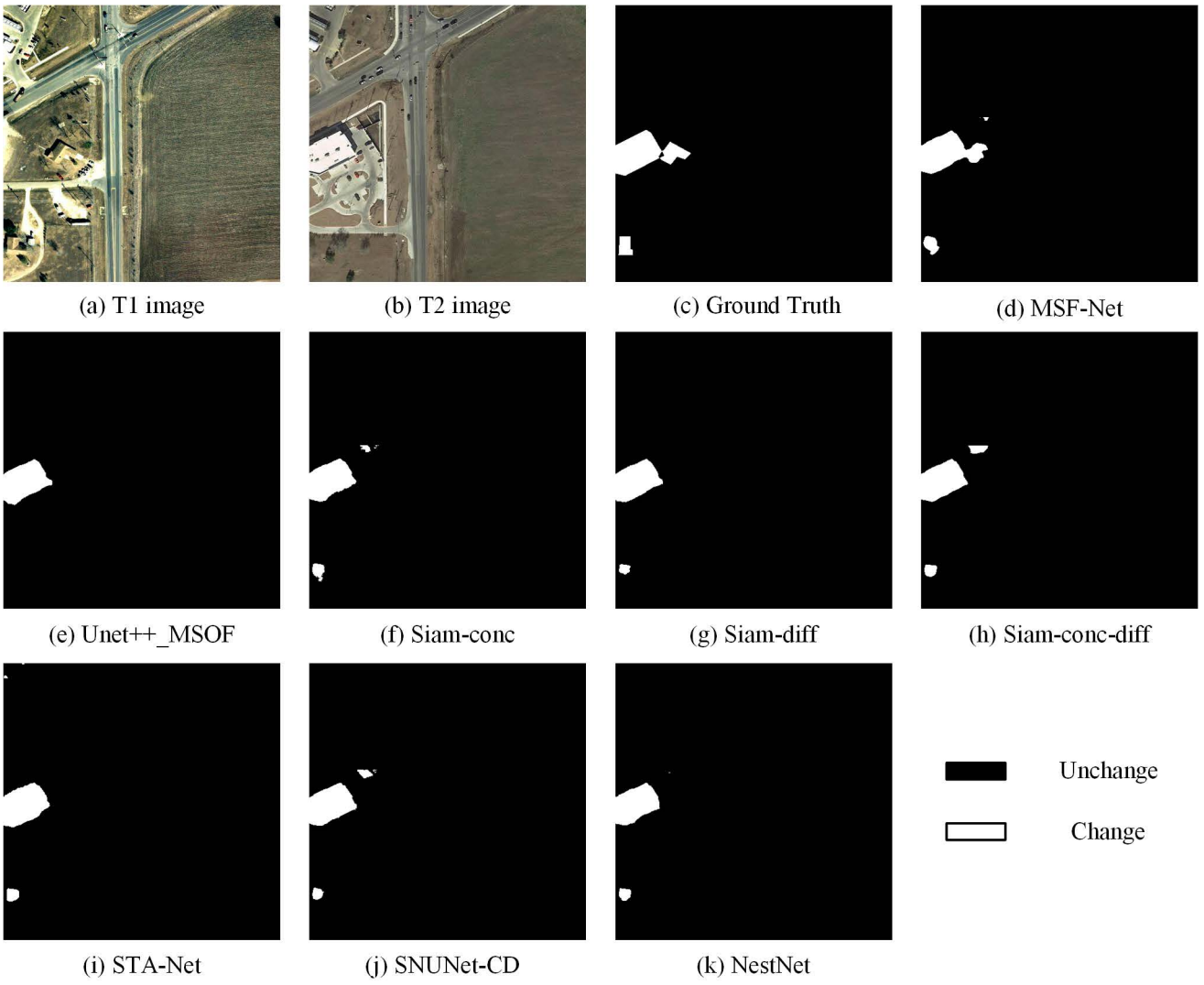


FIGURE 8. Comparison of building change detection with different networks for scene (c) in Fig. 5.

detection. It illustrates that MSF-Net improves the drawback of fuzzy building boundaries extracted by state-of-the-art networks.

MSF-Net has shown excellent ability in filtering irrelevant change information. In Fig. 7, the buildings in the two periods images did not change, although the colors were different due to the effects of light and other factors. MSF-Net did not mistakenly detect buildings change because the dual-context fusion module was used to remove the effects of these factors during the encoding module. In contrast, other models fail to remove the effects of light factors on building change detection and incorrectly identify unchanged buildings as changed. It indicates that MSF-Net has strong building feature extraction ability, and can effectively suppress the influence of other feature changes on building change detection.

MSF-Net detected relatively complete change information of buildings at different scales (Fig. 8), while other change detection methods did not detect the change of small-scale buildings due to the influence of light and roads. The results indicate that MSF-Net can suppress unnecessary change information, focusing on building information to improve the detection accuracy of buildings. It also shows that MSF-Net can simultaneously extract changing buildings at different scales in the same scene.

F. ABLATION EXPERIMENTS

To verify the effectiveness of the multiple attention mechanisms and multi-output fusion modules proposed in this paper, we designed three networks for the ablation experiments. A baseline network with all attention mechanisms and multi-output fusion modules removed, a baseline network with multi-output fusion, and a baseline network with both multi-output fusion and attention mechanisms. The comparison results of the three networks are shown in Table 2.

The ablation experiments verified the availability of these modules. In Table 2, after adding the multi-output fusion module, the network has a decrease in Recall. The main

TABLE 2. Verification result of the attention module and multi-output fusion effectiveness.

Methods	Precision	Recall	F1-score	IOU
Baseline	0.8845	0.8647	0.8608	0.7814
Baseline + multi-output fusion	0.9118	0.8483	0.8638	0.7845
Baseline + multi-output fusion + attention	0.9092	0.8826	0.8866	0.8130

reason is that too few samples of changing buildings will lead to low Recall of the network. The multi-output fusion module, on the other hand, performs multiple feature extractions of buildings at multiple scales, which aggravates the negative effect of too few samples of changing buildings to a certain extent, thus reducing the Recall. But Precision got a great improvement, and the two integrated indexes, F1-score and IOU have also been improved, indicating that the multi-output fusion module focuses on improving the detection Precision of changing buildings. After continuing to add the attention module, the model has a slight decrease in Precision, but Recall improves by 3.43%, and F1-score and IOU improve by 2.28% and 2.85%, respectively, indicating that the model has stronger learning ability after adding the dual-context fusion module, the channel attention module, and the selective kernel convolution module. The ablation experiment demonstrates that the proposed multi-output fusion module and attention module proposed in this paper are effective. These two modules enhance MSF-Net's ability to extract building features, filter irrelevant change information, and also improve multiscale fusion capabilities, making the detected building change boundaries more complete. It improves the shortcomings of the state-of-the-art network.

IV. CONCLUSION

In this paper we proposed a multiscale supervised fusion network based on the attention mechanism for building change detection in high-resolution remote sensing images. We introduced a new dual-temporal image fusion module to limit the effects brought by factors such as illumination on building extraction. We also introduced multiple attention mechanisms to exclude the irrelevant changes and focus on building changes, enhanced building feature extraction capabilities. In addition, we designed a multi-output fusion module to enhance multiscale information fusion capabilities, increased the precision of building change detection. We also designed an ablation experiment to verify the effectiveness of the proposed module. Combining the quantitative and qualitative analysis, we can see that our network focuses on extracting the change information of buildings, and can extract building features at different scales at the same time, which improves the extraction ability of change buildings. At the same time, the extracted boundaries are more complete and clear. The F1-score and IOU have improved significantly compared to state-of-the-art networks, demonstrating the contribution of MSF-Net to building change detection.

Although our network achieved good results, it still has some limitations. The network introduces attention mechanisms and multi-output fusion module to enhance the extraction of buildings, which suppresses the influence of other feature changes and improves the detection precision of the model, but also makes the network too strict in detecting buildings, and there is the phenomenon of missing detection of changing buildings. This makes the Recall of the network lower than that of STA-Net, which will be the direction of our future efforts.

Our future work will incorporate the use of multi-source multispectral image data and radar data to improve further the ability of building change detection. In addition, we intend to apply building change detection to specific types of changes and explore what kind of types of changes have occurred.

REFERENCES

[1] R. A. Weismiller, S. J. Kristof, D. K. Scholz, P. E. Anuta, and S. A. Momin, "Change detection in coastal zone environments," *Photogramm. Eng. Remote Sens.*, vol. 43, no. 12, pp. 1533–1539, 1978.

[2] M. C. Hansen and T. R. Loveland, "A review of large area monitoring of land cover change using Landsat data," *Remote Sens. Environ.*, vol. 122, pp. 66–74, Jul. 2012.

[3] J. R. Hu and Y. Z. Zhang, "Seasonal change of land-use/land-cover (LULC) detection using MODIS data in rapid urbanization regions: A case study of the pearl river delta region (China)," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 6, no. 4, pp. 1913–1920, Aug. 2013.

[4] E. M. Nielsen, S. D. Prince, and G. T. Koeln, "Wetland change mapping for the US mid-Atlantic region using an outlier detection technique," *Remote Sens. Environ.*, vol. 112, no. 11, pp. 4061–4074, Nov. 2008.

[5] C. Song, B. Huang, L. Ke, and K. S. Richards, "Remote sensing of Alpine lake water environment changes on the Tibetan plateau and surroundings: A review," *ISPRS J. Photogramm. Remote Sens.*, vol. 92, pp. 26–37, Jun. 2014.

[6] F. Bovolo and L. Bruzzone, "A split-based approach to unsupervised change detection in large-size SAR images," *Proc. SPIE*, vol. 6365, Sep. 2006, Art. no. 63650I.

[7] N. Sofina and M. Ehlers, "Building change detection using high resolution remotely sensed data and GIS," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 9, no. 8, pp. 3430–3438, Aug. 2016.

[8] Z. Liangpei and W. Chen, "Advance and future development of change detection for multi-temporal remote sensing imagery," *Acta Geodaetica Cartographica Sinica*, vol. 46, no. 10, pp. 1447–1459, 2017.

[9] S. Haigang, F. Wenqing, L. Wenzhuo, S. Kaimin, and X. Chuan, "Review of change detection methods for multi-temporal remote sensing imagery," *Geomatics Inf. Sci. Wuhan Univ.*, vol. 43, no. 12, pp. 1885–1898, 2018.

[10] A. Singh, "Change detection in the tropical forest environment of northeastern India using Landsat," *Remote Sens. Tropical Land Manage.*, vol. 44, pp. 254–273, Feb. 1986.

[11] P. J. Howarth and G. M. Wickware, "Procedures for change detection using Landsat digital data," *Int. J. Remote Sens.*, vol. 2, no. 3, pp. 277–291, Jul. 1981.

[12] J. Chen, P. Gong, C. He, R. Pu, and P. Shi, "Land-use/land-cover change detection using improved change-vector analysis," *Photogramm. Eng. Remote Sens.*, vol. 69, no. 4, pp. 369–379, 2003.

[13] J. Chen, X. Chen, X. Cui, and J. Chen, "Change vector analysis in posterior probability space: A new method for land cover change detection," *IEEE Geosci. Remote Sens. Lett.*, vol. 8, no. 2, pp. 317–321, Mar. 2010.

[14] J. Chen, Z. Yuan, J. Peng, L. Chen, H. Huang, J. Zhu, Y. Liu, and H. Li, "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, 2021.

[15] H. Wei, H. Jinliang, W. Lihui, H. Yanxia, and H. Pengpeng, "Remote sensing image change detection based on change vector analysis of PCA component," *Remote Sens. Land Resour.*, vol. 28, no. 1, pp. 22–27, 2016.

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

[17] G. Xu, H. Li, Y. Zang, L. Xie, and C. Bai, "Change detection based on IR-MAD model for GF-5 remote sensing imagery," in *Proc. RSWC*, Shanghai, China, 2019, p. 8.

[18] W. Chen, L. Zhang, and D. Bo, "Kernel slow feature analysis for scene change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 4, pp. 2367–2384, Apr. 2017.

[19] C. Zhang, P. Yue, D. Tapete, L. Jiang, B. Shangguan, L. Huang, and G. Liu, "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.

[20] X. Yu, J. Fan, J. Chen, P. Zhang, and L. Han, "NestNet: A multiscale convolutional neural network for remote sensing image change detection," *Int. J. Remote Sens.*, vol. 42, no. 13, pp. 4902–4925, 2021.

[21] H. Zhang, M. Wang, F. Wang, G. Yang, Y. Zhang, J. Jia, and S. Wang, "A novel squeeze-and-excitation W-net for 2D and 3D building change detection with multi-source and multi-feature remote sensing data," *Remote Sens.*, vol. 13, no. 3, p. 440, Jan. 2021.

[22] G. Hinton, S. Osindero, and Y. Teh, "A fast learning algorithm for deep belief nets," *Neural Comput.*, vol. 18, no. 7, pp. 1527–1554, Jul. 2006.

[23] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst. (NIPS)*, vol. 25, Dec. 2012, pp. 1097–1105.

[24] Y. Ren, C. Zhu, and S. Xiao, "Small object detection in optical remote sensing images via modified faster R-CNN," *Appl. Sci.*, vol. 8, no. 5, p. 813, 2018.

[25] H. S. Alhichri, A. S. Alswayed, Y. Bazi, N. Ammour, and N. A. Ajlan, "Classification of remote sensing images using EfficientNet-B3 CNN model with attention," *IEEE Access*, vol. 9, pp. 14078–14094, 2021.

[26] Y. Wei, K. Zhang, and S. Ji, "Simultaneous road surface and centerline extraction from large-scale remote sensing images using CNN-based segmentation and tracing," *IEEE Trans. Geosci. Remote Sens.*, vol. 58, no. 12, pp. 8919–8931, Dec. 2020.

[27] S. R. Maiya and S. C. Babu, "Slum segmentation and change detection : A deep learning approach," 2018, *arXiv:1811.07896*.

[28] Q. Wang, X. Zhang, G. Chen, F. Dai, Y. Gong, and K. Zhu, "Change detection based on faster R-CNN for high-resolution remote sensing images," *Remote Sens. Lett.*, vol. 9, no. 10, pp. 923–932, 2018.

[29] 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.

[30] S. Saha, F. Bovolo, and L. Bruzzone, "Change detection in image time-series using unsupervised LSTM," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, pp. 1–5, 2022.

[31] W. Huang, S. Zhang, and H. H. Wang, "Efficient GAN-based remote sensing image change detection under noise conditions," in *Proc. Int. Conf. Image Process. Capsule Netw.* Cham, Switzerland: Springer, 2020, pp. 1–8.

[32] 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.*, Jul. 2019, pp. 214–217.

[33] 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.

[34] 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.

[35] 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, p. 1382, 2019.

[36] 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.* Cham, Switzerland: Springer, 2015, pp. 234–241.

[37] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "U-Net++: A nested U-Net architecture for medical image segmentation," in *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*. Cham, Switzerland: Springer, 2018, pp. 3–11.

[38] 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, Oct. 2017.

[39] Y. Liu, C. Pang, Z. Zhan, X. Zhang, and X. Yang, "Building change detection for remote sensing images using a dual-task constrained deep Siamese convolutional network model," *IEEE Geosci. Remote Sens. Lett.*, vol. 18, no. 5, pp. 811–815, May 2021.

[40] H. Jiang, X. Hu, K. Li, J. Zhang, J. Gong, and M. Zhang, "PGA-SiamNet: Pyramid feature-based attention-guided Siamese network for remote sensing orthoimagery building change detection," *Remote Sens.*, vol. 12, no. 3, p. 484, Feb. 2020.

[41] K. Li, Z. Li, and S. Fang, "Siamese NestedUNet networks for change detection of high resolution satellite image," in *Proc. Int. Conf. Control, Robot. Intell. Syst.*, Oct. 2020, pp. 42–48.

[42] H. Xia, J. Ma, J. Ou, X. Lv, and C. Bai, "Pedestrian detection algorithm based on multi-scale feature extraction and attention feature fusion," *Digit. Signal Process.*, vol. 121, Mar. 2022, Art. no. 103311.

[43] Y. Chi, J. Li, and H. Fan, "Pyramid-attention based multi-scale feature fusion network for multispectral pan-sharpening," *Appl. Intell.*, vol. 52, pp. 5353–5365, Aug. 2021.

[44] Q.-L. Zhang and Y.-B. Yang, "SA-net: Shuffle attention for deep convolutional neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2021, pp. 2235–2239.

[45] A. Vaswani *et al.*, "Attention is all you need," in *Proc. 31st Conf. Neural Inf. Process. Syst. (NIPS)*, Long Beach, CA, USA, 2017, pp. 5998–6008.

[46] Q. Dou, Y. Lu, P. Manakul, X. Wu, and M. J. F. Gales, "Attention forcing for machine translation," 2021, *arXiv:2104.01264*.

[47] H. I. Liu and W. L. Chen, "Re-transformer: A self-attention based model for machine translation," *Procedia Comput. Sci.*, vol. 189, no. 8, pp. 3–10, 2021.

[48] L. Zhou, C. Xu, P. Koch, and J. J. Corso, "Watch what you just said: Image captioning with text-conditional attention," 2016, *arXiv:1606.04621*.

[49] S. Qu, Y. Xi, and S. Ding, "Visual attention based on long-short term memory model for image caption generation," in *Proc. 29th Chin. Control Decis. Conf. (CCDC)*, May 2017, pp. 4789–4794.

[50] J. Yang, B. Price, S. Cohen, and M.-H. Yang, "Context driven scene parsing with attention to rare classes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 3294–3301.

[51] H. Wang, Y. Fan, Z. Wang, L. Jiao, and B. Schiele, "Parameter-free spatial attention network for person re-identification," 2018, *arXiv:1811.12150*.

[52] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-excitation networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 8, pp. 2011–2023, Apr. 2019.

[53] Q. Shi, M. Liu, S. Li, X. Liu, F. Wang, and L. Zhang, "A deeply supervised attention metric-based network and an open aerial image dataset for remote sensing change detection," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, pp. 1–16, 2022.

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

[55] S. Woo, J. Park, J. Y. Lee, and I. S. Kweon, *CBAM: Convolutional Block Attention Module*. Cham, Switzerland: Springer, 2018.

[56] J. Park, S. Woo, J.-Y. Lee, and I. So Kweon, "BAM: Bottleneck attention module," 2018, *arXiv:1807.06514*.

[57] X. Yu, J. Fan, P. Zhang, L. Han, D. Zhang, and G. Sun, "Multi-scale convolutional neural network for remote sensing image change detection," in *Geoinformatics in Sustainable Ecosystem and Society*. Singapore: Springer, 2019, pp. 234–242.

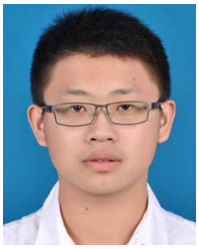
[58] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2881–2890.

[59] 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, pp. 1–5, 2022.

[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, p. 1662, May 2020.

[61] X. Li, W. Wang, X. Hu, and J. Yang, "Selective kernel networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 510–519.

[62] C. Chen and G. X. Gu, "Generative deep neural networks for inverse materials design using backpropagation and active learning," *Adv. Sci.*, vol. 7, no. 5, Mar. 2020, Art. no. 1902607.
[63] T. P. Lillicrap, A. Santoro, L. Marris, C. J. Akerman, and G. Hinton, "Backpropagation and the brain," *Nature Rev. Neurosci.*, vol. 21, pp. 335–346, 2020.
[64] A. Mukherjee, D. K. Jain, P. Goswami, Q. Xin, L. Yang, and J. J. Rodrigues, "Back propagation neural network based cluster head identification in MIMO sensor networks for intelligent transportation systems," *IEEE Access*, vol. 8, pp. 28524–28532, 2020.



JIAHAO CHEN was born in Qingdao, Shandong, China, in 1997. He received the B.S. degree in geographic information science from the Shandong University of Technology, Zibo, China, in 2019, where he is currently pursuing the M.S. degree with the Department of Surveying and Mapping Engineering.

His research interest includes spatiotemporal intelligent computing.

JUNFU FAN was born in Liaocheng, Shandong, China, in 1985. He received the B.S. and M.S. degrees in cartography and geographic information system from the Shandong University of Science and Technology, Qingdao, China, in 2008 and 2011, respectively, and the Ph.D. degree in cartography and geographic information system from the Institute of Geographical Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing, China, in 2014.

He is currently an Associate Professor with the Department of Surveying and Mapping Engineering, Shandong University of Technology. His research interests include high performance geo-computing, spatial analysis algorithm design and development, and remote sensing of urban environment.

MENGZHEN ZHANG was born in LinYi, Shandong, China, in 1997. She received the B.S. degree in geographic information science from the Shandong University of Technology, Zibo, China, in 2020, where she is currently pursuing the M.S. degree in science and technology of surveying and mapping.

Her research interests include intelligent geographic computing and change detection algorithm based on high resolution remote sensing images.

YUKE ZHOU was born in Jining, Shandong, China, in 1984. He received the Ph.D. degree in cartography and geographic information system from the Institute of Geographical Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing, China, in 2013.

He is currently an Associate Researcher with the Institute of Geographical Sciences and Natural Resources Research, CAS. His research interests include high performance geo-computing and development, and remote sensing of ecosystem environment.

CHEN SHEN was born in Zibo, Shandong, China, in 1989. He received the B.S. degree in surveying and mapping engineering from the Kunming University of Science and Technology, Kunming, China, in 2012.

He is currently working with Shandong Tianyunhe Information Technology Company, Ltd., which is also known as the Organizer of High-Resolution Earth Observation System Data and Application Center of Zibo. His research interest includes spatiotemporal intelligent computing.