Deep Depthwise Separable Convolutional Network for Change Detection in Optical Aerial Images

Ruochen Liu, Dawei Jiang, Langlang Zhang, and Zetong Zhang

***Abstract*—In this article, a remote sensing image change detection method based on depthwise separable convolution with U-Net is proposed, which omits the tedious steps of generating and analyzing the difference map in the traditional remote sensing image change detection method. First, two images having *c*-channel each can be specifically stacked into a 2*c*-channel image, and the change detection can be converted to an image segmentation problem, an improved full convolution network (FCN) called U-Net is exploited to directly separate the changing regions. Because the capability of the deep convolution network is proportional to the depth of the network and a deeper convolution network means the increase of the training parameters, we then replace the original convolution in FCN by the depthwise separable convolution, making the entire network lighter, while the model performs slightly better than the traditional convolution operation. Besides that, another innovation in our proposed method is to use a preference control loss function to meet the different needs of precision and recall rate. Experimental results validate the effectiveness and robustness of the proposed method.**

***Index Terms*—Change detection, depthwise separable convolution, image segmentation, optical aerial images.**

I. Introduction

IMAGE change detection is to detect the change of the two images taken at different times in the same place. With the continuous development of aerial photography and satellite radar technology, the acquisition of remote sensing images becomes easier, which facilitates the wide application of change detection technologies in land cover detection [1], building change detection [2], disaster assessment, and other fields. Compared with the time-consuming and tedious manual processing, it is very important to detect the changed area automatically in real time.

According to the basic unit of data processed, the change detection methods are divided into four main categories: pixel based, parcel based, window based, and patch based. The pixel-based method takes the pixel as the basic unit of image analysis, making full use of spectral characteristics without considering the spatial context. This method is simple and easy to understand, but it has poor robustness to noise. The parcel-based change detection method uses the object as the analysis unit [3], which requires relatively low registration accuracy. It can directly obtain the change target and facilitate subsequent processing. However, the challenge for this method lies in it is difficult to extract objects. The window-based approach first obtains the difference map, then slides the window to get the block on the difference map and classifies the pixels in the center of the window or the pixels of the whole window. In recent years, this approach becomes increasingly since it can be combined with deep learning. Gong *et al.* [4] proposed an unsupervised method based on deep neural networks (DNN), which can obtain final change detection map directly from the two original images. According to the selection criteria, a training set is constructed by taking blocks with sliding windows on two original images, then a stacked restricted Boltzmann machines network is learned for binary classification. Similar window-based methods can be found in [5] and [6]. This method has significant advantages over pixel-based method for considering the spatial neighborhood and the contextual information, and has better robustness to noise. Unfortunately, with the expansion of the data scale in high-resolution remote sensing images, window-based method becomes poor in performance. In addition, the application of the sliding window mechanism to large-size images takes a lot of time, and when the image resolution is increased and the terrain environment is complex, the window-based method is sensitive to noise. Reasonably, the patch-based method appears for this case. In [7], Gong *et al.* innovatively use generative adversarial networks (GAN) for image change detection. GAN acts as a generative model to learn the distribution between the training data and their corresponding image patches, and then capture the whole difference map by the generator of GAN, later traditional fuzzy local information c-means algorithm is used for post analysis. In [8], a supervised Siamese convolutional neural network (SCNN) is proposed to extract features from the input image pairs, and in order to identify the changed and unchanged pixels more effectively and reduce the impact of data imbalance, the authors employ a weighted contrastive loss.

With the development of deep learning, various neural network structures have been applied to change detection. In [9], a feature learning method using stacked contractive autoencoder is presented to extract temporal change feature from super pixel with noise suppression. In [6], an unsupervised deep learning detection method, symmetric convolution coupled network (SCCN), is proposed to detect varying and invariant regions in

Manuscript received May 13, 2019; revised September 26, 2019, December 9, 2019, and February 10, 2020; accepted February 11, 2020. Date of publication March 16, 2020; date of current version April 8, 2020. This work was supported in part by the National Natural Science Foundation of China under Grants 61876141 and 61373111 and in part by the Provincial Natural Science Foundation of Shaanxi of China under Grant 2019JZ-26. *(Corresponding author: Ruochen Liu.)*

The authors are with the Laboratory of Intelligent Perception and Image Understanding of Ministry of Education, Xidian University, Xi'an 710071, China (e-mail: ruochenliu@xidian.edu.cn; 1656556479@qq.com; 984195898@qq.com; 503202358@qq.com).

Digital Object Identifier 10.1109/JSTARS.2020.2974276

nonuniform optical and SAR images. Unlike existing methods based on labeled pixels to learn the potential relationship between two heterogeneous images, SCCN is completely unsupervised and does not use any labeled pixels. The unsupervised deep learning method also appears in [10]. The authors proposed an unsupervised context sensitive framework named depth change vector analysis. DCVA uses multilayer CNN segmented semantically to train the aerial optical images and obtain multilayer depth features at the pixel level. Lyu *et al.* [11] made use of an RNN-based network to solve the multispectral change detection task, in which, the joint spectral–temporal feature representation is learned from a bitemporal image sequence using long short-term memory network. Mou *et al.* [12] proposed a novel network architecture, which is trained to learn a joint spectral–spatial–temporal feature representation in a unified framework for change detection of multispectral images. For this purpose, they combined CNN and RNN into an end-to-end network framework. The former is responsible for extracting the rich spectral–spatial features of bitemporal images, whereas the latter is effective in analyzing the temporal dependence of bitemporal images. In order to solve the difficulty of marking a large amount of labeled data, Gong *et al.* [13] proposed a generative discriminatory classified network for multispectral image change detection, in which labeled data, unlabeled data, and new fake data generated by GANs are used. Feng *et al.* [14] presented a novel change detection method for multitemporal synthetic aperture radar images based on PCANet, which exploits representative neighborhood features from each pixel using PCA filters as convolution filters.

Deep learning is widely used in computer vision because of its powerful representation ability. But for change detection, the cost of manually annotating data is very expensive, and the training of CNN or DNN requires a large amount of data to avoid overfitting. Although we use rotation, distortion, adding noise, and even more efficient data enhancement methods to extend the training set, these may not solve the question. In order to address the problem, we exploit a depthwise separable convolution [15], which requires fewer parameters and involves less computation to learn better representations with less data.

The traditional change detection method first preprocesses the image and then generates and analyzes the difference map. The quality of these methods is closely related to the quality of the difference map. If the generated difference map carries a lot of noise, the results are generally poor. Considering the joint distribution of the images, we stack the two-phase images along the channel. After such processing, an image change detection problem can be transformed into an image segmentation problem, and then the U-Net structure commonly used in image semantic segmentation is exploited as the framework to segment stacked images. Image semantic segmentation sets a category label for each pixel in the image, and then uses image segmentation algorithm to segment the original image into different regions with the same semantic. Different from the semantic segmentation method, the proposed method directly classifies the pixels of the stacked image. Our main contributions can be summarized as follows.

1) We implemented a deep full convolution network (FCN) based on depthwise separable convolution to get the change detection map from the two original images directly. The application of depthwise separable convolution improves the convolution efficiency, greatly reduces the model training parameters. The stacked image is directly fed into the first convolutional layer of our proposed model, and then we train the whole model in an end-to-end manner, where the training labels are the corresponding change of the input stacked image.
2) We define a joint loss function that combines image segmentation problems with image binary classification problems. For the sake of meeting different needs to precision and recall rate flexibly, we adopt a preference control loss function.

The rest of this article is organized as follows. The proposed framework is described in Section II. Section III shows the experimental results on optical aerial datasets and a comparison with other existing methods. Finally, the conclusion of this article is drawn in Section IV.

II. Proposed Framework

A. U-Net Architecture With Separable Convolution

Image semantic segmentation can be defined to classify images into predefined categories at the pixel level. Some traditional CNN-based segmentation methods typically use one image block around the center pixel as input to the CNN for training and prediction in order to classify a pixel. However, these methods are not only expensive in terms of time and storage space, but also limit the size of the perception area. To solve this problem, FCN [16] was proposed. It recovers the class of each pixel from the extracted high-dimensional features, and replaces the fully connected layer with the convolutional layer to obtain the spatial feature map instead of the classification scores. This method allows training a CNN model in the end-to-end manner for image segmentation with free-size input image. Our proposed network architecture is based on the improved version of FCN called U-Net [17], which was proposed for medical image segmentation. As shown in Fig. 1, the U-Net architecture consists of contracting path and expansive path, in which combines low-level feature maps with high-level feature maps using skip connection to bring precise pixel-level positioning. Admirably, such connection can help the decoder better repair target details. In the expansive path, numerous of feature channels propagate contextual information to higher resolution layers. In the binary image segmentation tasks, such as satellite image analysis and medical image analysis, the effectiveness of this structure has been well demonstrated.

In [15], the Xception model of separable convolution is designed and its performance is slightly better than the state-of-the-art methods. In Fig. 2, we show internal operations of depthwise separable convolution. It consists of a depthwise convolution and a pointwise convolution, the former performs the spatial convolution independently on each channel of the input, and the latter is a regular convolution with 1×1 windows. In traditional convolutional layers, the weights in the network

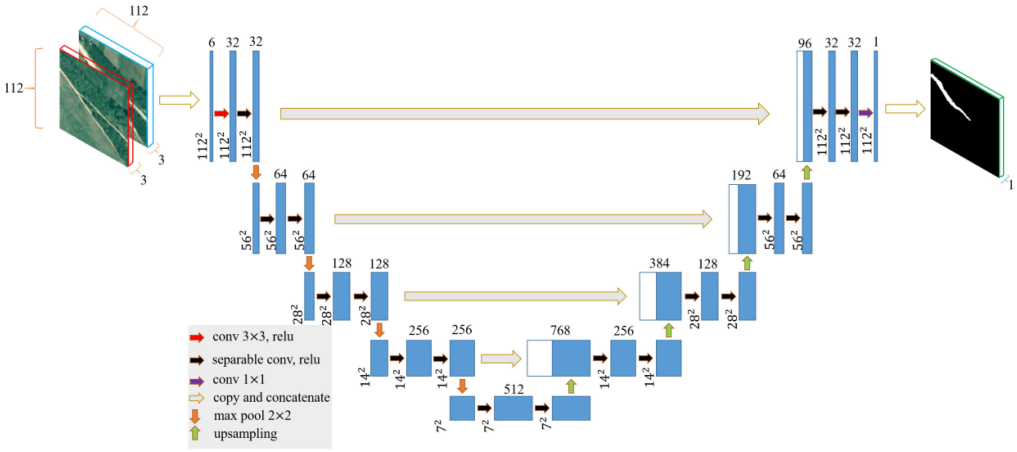


Fig. 1. Network structure of the proposed model based on U-Net and separable convolution.

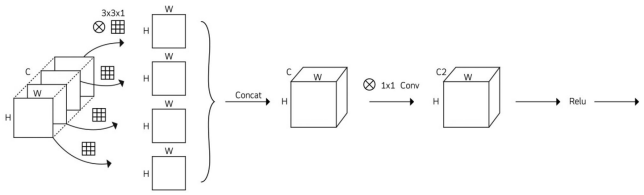


Fig. 2. Depthwise separable convolution: A depthwise convolution followed by a pointwise convolution.

are shared, feature extraction and feature fusion are performed simultaneously, the invariant function is used to sample the pool layer spatially, and a large number of parameters are generated. In contrast, the depthwise separable convolution separates the two steps by splitting the different channels from each other in the depth direction, it performs feature extraction first and then performs feature fusion. In this way, we can make full use of model parameters to carry out representation learning, so as to get a better model. It is noteworthy that depthwise separable convolution differs from spatially separable convolution in image processing field.

The traditional convolution, depthwise convolution, pointwise convolution, and depthwise separable convolution are shown in mathematical expression

$$\text{Conv}(W, y)_{(i,j)} = \sum_{k,l,m}^{K,L,M} W_{(k,l,m)} \cdot y_{(i+k,j+l,m)} \tag{1}$$

$$\text{PonitwiseConv}_{(i,j)} = \sum_{m}^{M} W_m \cdot y_{(i,j,m)} \tag{2}$$

$$\text{DepthwiseConv}(W, y)_{(i,j)} = \sum_{k,l}^{K,L} W_{(k,l)} \odot y_{(i+k,j+l)} \tag{3}$$

$$\text{SepConv}(W_p, W_d, y)_{(i,j)} = \text{PonitwiseConv}_{(i,j)} \\ ((W_p, \text{DepthwiseConv}(W, y)_{(i,j)}(W_d, y)) \tag{4}$$

where W is the input image. y represents a convolution kernel of size $K \times L$, whereas for pointwise convolution, its size is 1×1. M denotes the number of channels of the input picture.(i, j) is pixels of each image. From these definitions, we can conclude that pointwise convolution collects the characteristics of each point, whereas depthwise convolution collects the spatial characteristics of each channel.

B. Deep Depthwise Separable Convolutional Network for Change Detection

The network structure of the proposed model based on U-Net and separable convolution is shown in Fig. 1. The first layer of the network adopts traditional convolution and deep separable convolution is used in both shrink path and extension path of U-Net structure. The number of feature channels for each down-sampling step is doubled. The expansive path includes up-sampling operation of the feature maps, then performs the separable deep convolution with half of the feature channels and connects with the corresponding feature map from the shrinking path, we set $relu$ as activation function. In order to integrate the information of each pixel on different channels, we choose 1×1 convolution as the output layer of the whole model. In addition, because what we expect is a pixel-level binary map, here nonlinear activation function $sigmoid$ is used to obtain the probability map. Through the binary probability map, we can directly get the change detection map, with size of $w \times h \times 1$.

Considering the joint distribution of the two optical aerial images of $w \times h \times c$, we can stack them together along the third channel to form a $2c$-channel input image with new shape $w \times h \times 2c$. By using the proposed model to segment the new image, the detection results of the changed region can be obtained directly.

C. Loss Function

Since the binary image segmentation can be regarded as a pixel-level classification task, the common loss function of the binary image classification problem can be applied during network training, which is called binary cross entropy (BCE). BCE is defined as

$$\text{BCE} = -\frac{1}{n}\sum_{i=1}^{n}(y_i \log \hat{y}_i + (1 - y_i)\log(1 - \hat{y}_i)) \tag{5}$$

where n is the number of samples, and y_i and $\hat{y}_i$ represent the true mask and predicted probability value, respectively. Having set up our notation, $y_i \in \{0, 1\}$ and $\hat{y}_i \in [0, 1]$, we can use BCE to measure the dissimilarity between y_i and $\hat{y}_i$.

The performance of the model is assessed by using dice (DICE), which is also known as F_1 score. DICE can be defined as

$$\begin{aligned} \text{DICE} &= 2 \cdot \frac{|LT \cap PT| + \text{smooth}}{|LT| + |PT| + \text{smooth}} \\ &= 2 \cdot \frac{\sum_i y_i \hat{y}_i + \text{smooth}}{\sum_i y_i + \sum_i \hat{y}_i + \text{smooth}} \end{aligned} \tag{6}$$

where LT and PT represent the authentic mask and predicted image, respectively, and in order to prevent no change in some areas where the denominator is zero. We introduce the parameter smooth in the numerator and denominator, which plays a smoothing role, and we set it as 1 in the experiment. In general, for the purpose of achieving better results, evaluation metric, and training objective should be as close as possible, so we adopt the dice coefficient loss function, which is defined as

$$\text{DICE\_loss} = 1 - \text{DICE}. \tag{7}$$

For different applications, there is a different level of attention from precision and recall, some are more concerned with how many positive examples of predictions are true, whereas others are concerned with how many positive examples are detected. With this in mind, we fine-tune the original dice coefficient loss function by introducing a parameter ω to adjust the contribution of precision and recall to the metric. The adjusted DICE is denoted as DICE_loss_new and is defined as follows:

$$\text{DICE\_loss\_new} = 1 - \frac{(1+w)|LT \cap PT| + 1}{w|PT| + |LT| + 1} = 1 - F_\alpha. \tag{8}$$

The derivation process based on F_α is as follows:

$$\begin{aligned} F_\alpha &= \frac{1}{\alpha \cdot \frac{1}{\text{precision}} + (1-\alpha) \cdot \frac{1}{\text{recall}}} \\ &= \frac{\text{precision} \cdot \text{recall}}{\alpha \cdot \text{recall} + (1-\alpha) \cdot \text{precision}} \\ &= \frac{|LT \cap PT|}{\alpha|PT| + (1-\alpha)|LT|} \\ &= \frac{\frac{1}{1-\alpha} \cdot |LT \cap PT|}{\frac{\alpha}{1-\alpha}|PT| + |LT|} \\ &= \frac{(1 + \frac{\alpha}{1-\alpha}) \cdot |LT \cap PT|}{\frac{\alpha}{1-\alpha}|PT| + |LT|} \\ &= \frac{(1+\omega) \cdot |LT \cap PT|}{\omega|PT| + |LT|} \begin{cases} \alpha \in (0, 1) \\ \omega \in (0, \infty) \end{cases} \end{aligned}$$

where α is a controllable parameter, in extreme cases, we set α equal to 0, then $F_\alpha = \text{recall}$, and if we set α equal to 1, consequently $F_\alpha = \text{precision}$. We can easily draw a conclusion that the greater the value of α, the greater the contribution of precision to evaluation metric, the smaller the value of α, the greater the contribution of recall to evaluation metric. When $\alpha = 0.5$ and $\omega = 1$, both the contribution of the consistent, that is, we commonly used 1, and the corresponding loss function is dice coefficient loss. The effect of parameter ω on the result will be discussed later in the experiment.

Since the task we face is not only a binary classification problem, but also a binary image segmentation, naturally, the joint loss function can be defined as combination of (5) and (8)

$$\text{Loss} = \text{BCE} + \text{DICE\_loss\_new}. \tag{9}$$

For change detection, the number of changed and unchanged pixels in the image is quite different. In our constructed training data, the ratio of changed pixels to unchanged pixels is approximately equal to 1:24, which is extremely unbalanced. Thus, the weighted loss function seems to be a good idea, for example, Yang *et al.* [8] used a contrastive loss function to handle class imbalance problem. However, based on proposed model and the loss function in (9), class imbalance is no longer included in our consideration.

III. Experimental Study

A. Datasets

In order to assess the performance of our method, we choose the SZTAKI AirChange Benchmark set [18], which is a ground truth collection for change detection in optical aerial images taken with several years of time differences and in different seasonal conditions. To verify the robustness of our algorithm, we extend our experiments to ONERA Satellite Change Detection (OSCD) dataset [19]. The SZTAKI AirChange Benchmark set contains three groups of optical aerial image pairs, named TISZADOB, SZADA, and ARCHIEVE, in which, respectively, contains 5, 7, and 1 image pairs of size 952×640 and resolution 1.5 m/pixel and binary change masks (drawn by an expert). We used the first two group for analysis. Given the apparent differences in radiation condition, color histogram matching is applied to the two coregistered images. The matching results are shown in the first row of Figs. 3–5. The change detection data of ONERA satellite provides a comparative standard for the proposed single band, color, or multispectral change detection algorithm. It contains 24 regions around the world with approximately 600×600 pixels at 10-m resolution. The training set contains 14 pairs of labeled pictures of different channels, and the remaining 10 unlabeled data can be used for testing, the detection results can be uploaded to the website[1] to get result.

The mechanism for constructing training and testing data in SZTAKI AirChange Benchmark set is the same as [8], which aims to make a fair compressor. For each image pair, we crop the top-left corner to 784×448 for constructing testing data. The remaining area is used to construct training data, where the area can be overlapped sampling with a size of 112×112. Specifically, we randomly select a patch from the remaining area of each image pair and repeat it 300 times. After random cropping and data augmentation, we can get a total of 3600 training image pairs with a size of 112×112 and 12 testing

[1] http://dase.grss-ieee.org/

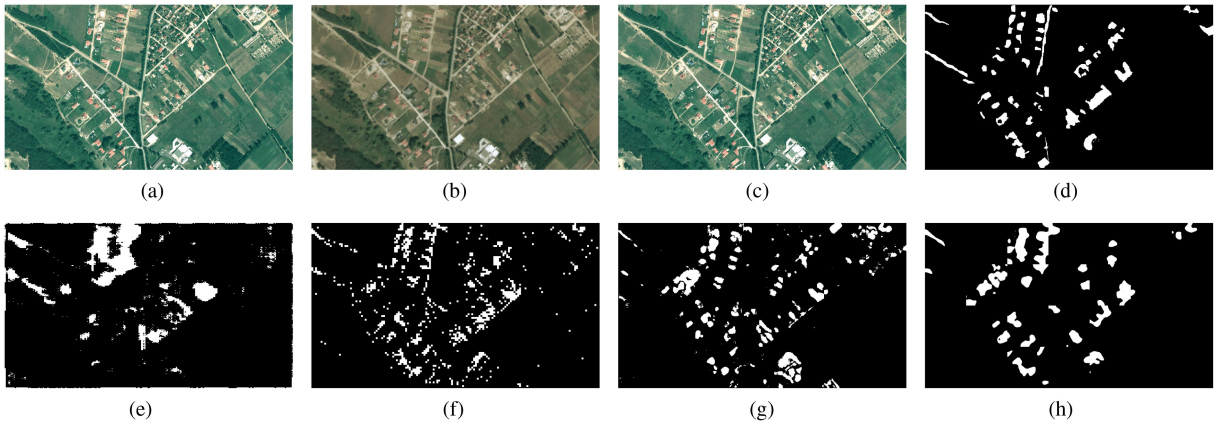


Fig. 3. (a)–(d) Optical aerial image pairs of SZADA/1 dataset and radiation correction result. (e)–(h) Experimental results by the proposed method and other methods on SZADA/1 dataset. (a) Image acquired at t_1. (b) Image acquired at t_2. (c) Radiation correction result for image at t_2. (d) GT. (e) GAN. (f) CNN. (g) SCNN. (h) Ours.

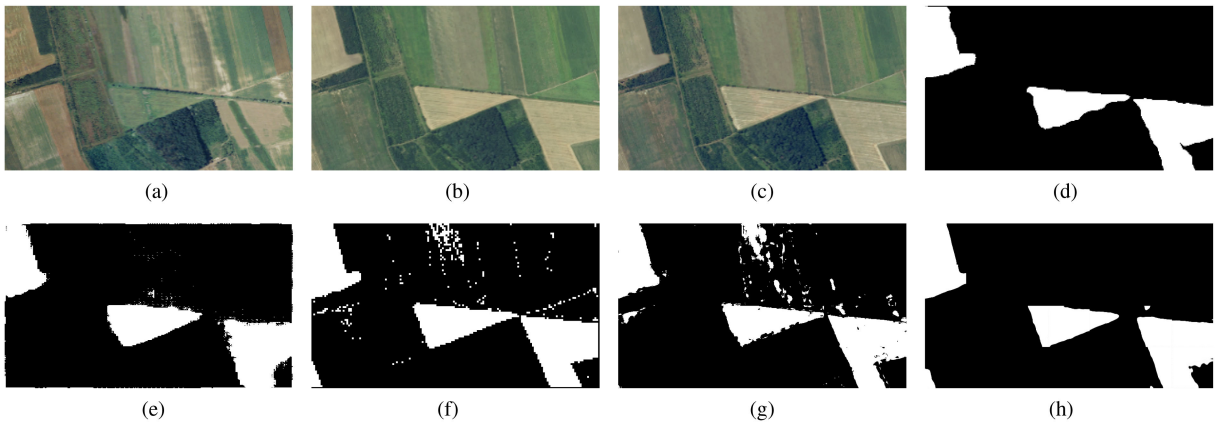


Fig. 4. (a)–(d) Optical aerial image pairs of TISZADOB/3 dataset and radiation correction result. (e)–(h) Experimental results by the proposed method and other methods on TISZADOB/3 dataset. (a) Image acquired at t_1. (b) Image acquired at t_2. (c) Radiation correction result for image at t_2. (d) GT. (e) GAN. (f) CNN. (g) SCNN. (h) Ours.

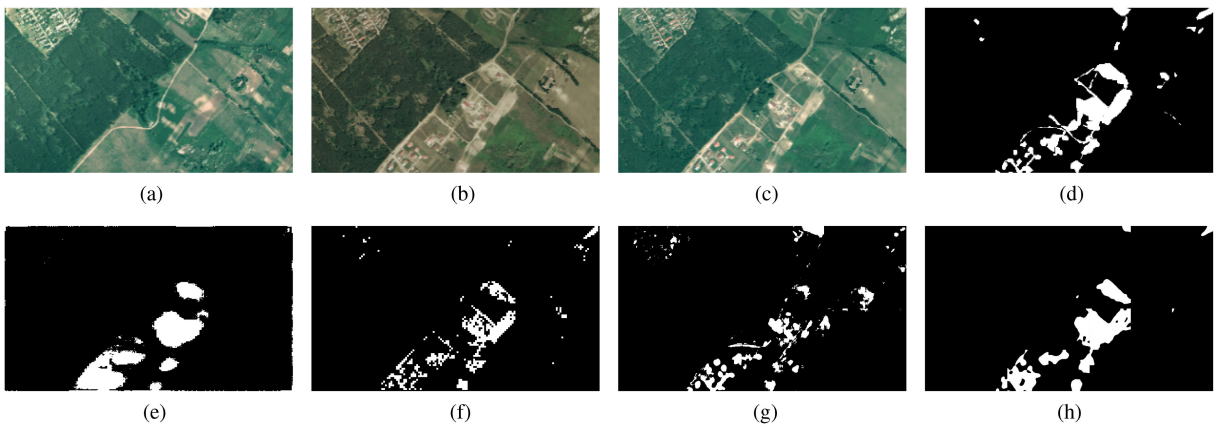


Fig. 5. (a)–(d) Optical aerial image pairs of SZADA/2 dataset and radiation correction result. (e)–(h) Experimental results by the proposed method and other methods on SZADA/2 dataset. (a) Image acquired at t_1. (b) Image acquired at t_2. (c) Radiation correction result for image at t_2. (d) GT. (e) GAN. (f) CNN. (g) SCNN. (h) Ours.

image pairs with a size of 784×448. For OSCD, we overlap the samples on the whole training set and test on unlabeled data. To augment training data, we use horizontal and vertical flipping and random rotations for each cropped training pair. In order to speed up the convergence and improve the detection accuracy, the data are fed into the model during the training and prediction stages are normalized to -1 to 1.

B. Optimization and Management of Training Details

We choose Adam [20] with learning rate 0.0001 as an optimization algorithm. Xavier initialization [21] is applied to initialize the weight of each layer of the network. The batch size of training data is set to 32, we train the network for 100 epochs. In order to prevent over fitting and reduce the training time of the model, we exploit the early-stopping mechanism, which plays an important role during the training.

C. Computational Time

The depthwise separable convolution has lower trainable parameters than the traditional convolution, the training time of our model is greatly reduced. We implement the proposed network based on the Keras framework, and a single NVIDIA GTX 1070ti GPU with 8 G memory is used for training and testing. In our implementation, the training takes about 50 min, and the inference on all test image pairs needs around 1.67 s.

D. Results and Evaluation

For SZTAKI AirChange Benchmark set, the precision(P), recall(R), and $F_{1-\text{score}}$ [22] are employed and the ROC curve [23] is analyzed graphically to evaluate the performance of the proposed method. For the OSCD dataset, we report the Overall acc, Average acc, and Change acc that obtained from the website after submitting our detection results. The Overall acc is the sum of all correctly classified pixels divided by the total number of pixels of the test dataset, the Average acc represents the average accuracy rate of the change and unchanged class, and the Change acc denotes the accuracy rate of changed class, which is considered as an important evaluation indicator.

1) Comparison Between Proposed Method and Three Other Methods: Since our method is patch-based, we select SCNN [8] and GAN-based [7] method as the comparison algorithms. In addition, a supervised CNN classification model is used to validate the superiority of the patch-based approach. When simulating the GAN-based approach, we allow the generator to generate a change map directly, rather than generating a difference map. This can be understood as a style conversion model transformed from an optical image to a binary image. The second row in Figs. 3–5 show the visual experimental results on our test data. From the change detection result graph, it can be seen that the CNN and SCNN algorithms have more noise points, the detected change area is not smooth. The details of the images generated by GAN are not well preserved, which is more obvious on the SZADA/1 and SZADA/2 datasets. The change region of the detection of the proposed algorithm is relatively smooth. This is mainly because U-Net with good segmentation performance and more efficient convolution are used in the design of the network model, which significantly improves the detection performance. There are some misdetections in some small areas, this may be due to the fact that we convert the change detection into a pixelwise classification problem. The quantitative values are exhibited in Table I. Meanwhile, the ROC curves of Fig. 6 visually show the performance for these algorithms. We can see that our proposed method has a smoother detection result and outperforms the other three methods in most evaluation criteria, especially in terms of $F_{1-\text{score}}$, which is a comprehensive evaluation criteria. Except that CNN-based classification method has competitive AUC value on TISZADOB/3 dataset, our proposed method achieves the best AUC value over all three datasets. Furthermore, the results of the proposed method are with very fewer burrs compared to other methods. This is mainly because U-Net with good segmentation performance and more efficient convolution are used when designing the network model, which significantly improves the detection performance. On the OSCD dataset, our approach still achieves better performance. From the results, SCNN and CNN have lower accuracy in Change acc, which show that these two algorithms have serious miss detection. GAN is relatively smooth to some extent, but the overall performance is lower than our proposed algorithm. In addition, even with very few changed pixels in the image pairs, our method can also get robust results.

TABLE I
QUANTITATIVE COMPARISON AMONG DIFFERENT METHODS ON THREE IMAGE PAIRS

Datasets	Metrics	GAN	CNN	SCNN	ours
SZADA/1	Precision(%)	28.3	47.1	38.2	48.2
	Recall(%)	40.9	53.4	46.6	56.1
	$F_{1-score}(\%)$	33.5	50.1	42.0	51.8
TISZADOB/3	Precision(%)	87.9	83.4	83.6	93.4
	Recall(%)	92.5	97.4	92.9	91.5
	$F_{1-score}(\%)$	90.1	89.8	88.0	92.4
SZADA/2	Precision(%)	56.6	78.1	60.8	71.1
	Recall(%)	62.9	53.5	41.8	70.3
	$F_{1-score}(\%)$	59.6	63.5	49.5	70.9
OSCD	Average acc(%)	74.7	71.3	71.5	83.3
	Overall acc(%)	85.0	89.5	92.5	91.7
	Changed acc%)	64.4	50.9	48.2	73.8

TABLE II
COMPARISON OF PARAMETERS AND TIME COST BETWEEN CONV AND DEPTHWISE SEPARABLECONV

Convolution Type	Trainable Parameters	Inference Time
Conv	7760961	1.81s
Depthwise SeparableConv	1512097	1.67s

2) Depthwise Separable Conv Versus Conv: As described in Section II, with the same network architecture, compared to the traditional convolution, the depthwise separable convolution has lower trainable parameters and can reduce the theoretical calculation of the model. In this section, we first compare the number of the parameters and the inference time (in second) of using two convolutions. The results are shown inTable II. It is easy to see that the parameters of depthwise separable convolution are only about one-fifth of the traditional convolution parameters and the inference time of the former on the whole test datasets is less than that of the latter. This shows that the space

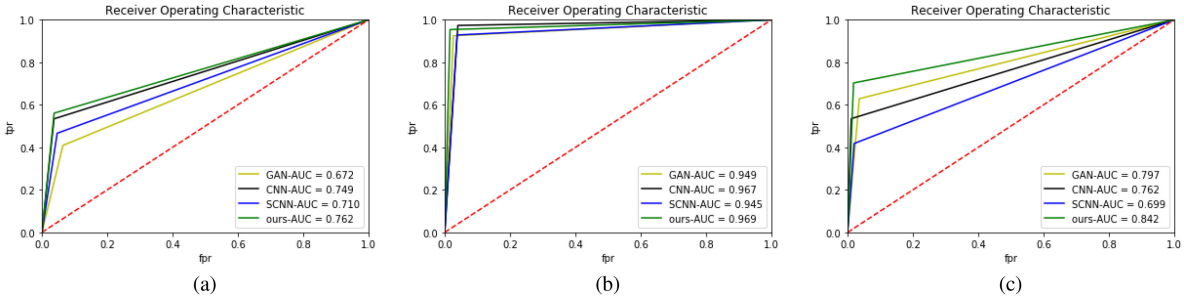


Fig. 6. ROC curves of the four different methods for the three datasets. (a) ROC curves for SZADA/1 dataset. (b) ROC curves for TISZADOB/3 dataset. (c) ROC curves for SZADA/2 dataset.

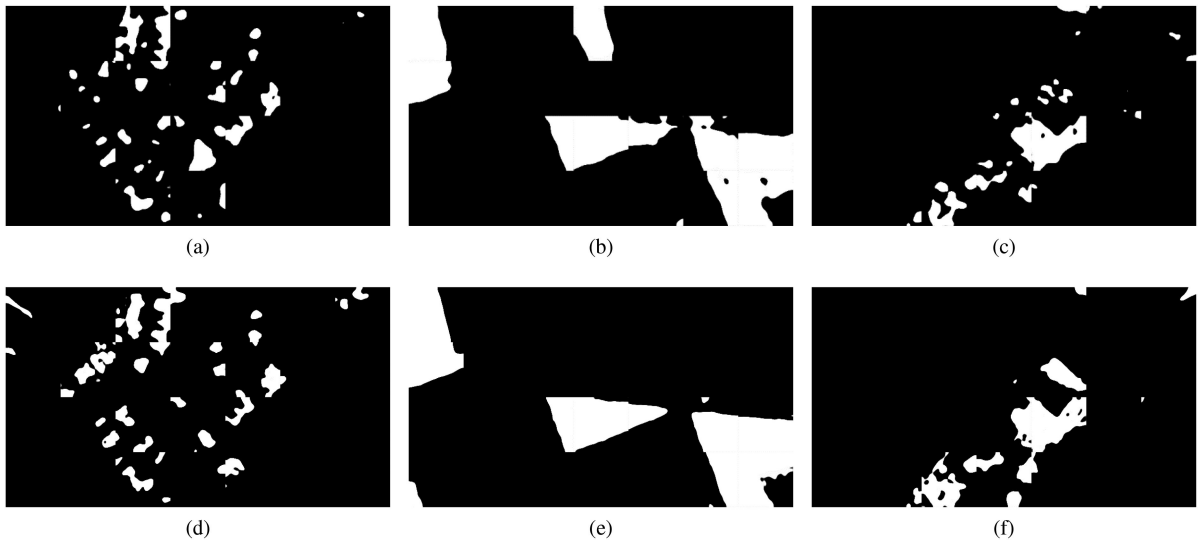


Fig. 7. Results by different convolutions. (a)–(c) Results by the traditional convolution. (d)–(f) Results by depthwise separable convolution.

TABLE III
QUANTITATIVE COMPARISON CONV AND DEPTHWISE SEPARABLECONV

Datasets	Metrics	Conv	Depthwise SeparableConv
SZADA/1	Precision(%)	52.9	48.2
	Recall(%)	55.5	56.1
	$F_{1-score}(\%)$	54.2	51.8
TISZADOB/3	Precision(%)	81.2	93.4
	Recall(%)	92.9	91.5
	$F_{1-score}(\%)$	86.6	92.4
SZADA/2	Precision(%)	73.3	71.1
	Recall(%)	54.6	70.3
	$F_{1-score}(\%)$	62.6	70.9
OSCD	Average acc(%)	80.0	83.3
	Overall acc(%)	93.7	91.7
	Changed acc%)	66.4	73.8

complexity and time complexity of the deep separable convolution are much lower. Fewer parameters can fit the training data better, we then verify its effectiveness on test dataset through experiments. For fair comparison, we only change the type of convolution and get the results of Table III and Fig. 7. For the TISZADOB/3 and SZADA/2 datasets, the depthwise separable convolution performs better than the traditional convolution on most evaluation criteria. Due to overly complicated lighting conditions and statistical properties, both convolutions perform poorly on the SZADA/1 datasets, the traditional convolution slightly better than the depthwise separable convolution. For the OSCD datasets, the performance of our algorithm is also better than the traditional convolution. Although the overall accuracy is higher than ours, this is because the area of change is relatively small compared to the area without change. In terms of indicators for detecting the changed area, our algorithm is higher than the traditional convolution.

3) Performance of Different Depth of the Network: Under the same U-Net structure, the depthwise separable convolution has lower trainable parameters and better robustness than the traditional convolution. The proposed network (in terms of layers) may be redundant compared to the given images. In this case, simply reducing the number of layers can reduce the number of parameters as well as improve performance. To further demonstrate the usefulness of our model, we explore the effects of different network depths on performance. For fair comparison, we only change the depths of convolution layers and the results are given in Table IV and Fig. 8. For all four datasets, the performance of deeper networks is better than that of shallower networks on most evaluation criteria. Especially in

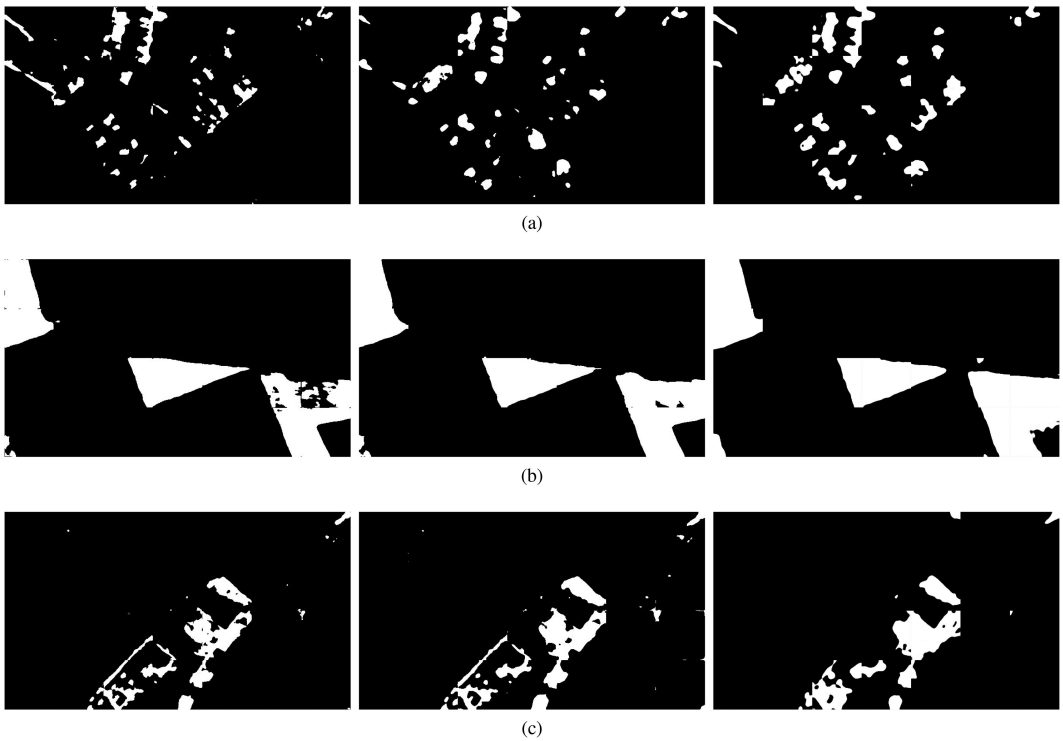


Fig. 8. Results by different depths that increase from left to right in depth. (a) For SZADA/1 dataset. (b) For TISZADOB/3 dataset. (c) For SZADA/2 dataset.

TABLE IV
QUANTITATIVE COMPARISON BETWEEN DIFFERENT DEPTHS

Datasets	Metrics	13layers	18layers	23layers
SZADA/1	Precision(%)	48.0	49.9	48.2
	Recall(%)	37.6	42.3	56.1
	$F_{1-score}(\%)$	42.2	45.8	51.8
TISZADOB/3	Precision(%)	97.1	96.8	93.4
	Recall(%)	83.6	88.2	91.5
	$F_{1-score}(\%)$	89.9	92.3	92.4
SZADA/2	Precision(%)	77.7	73.1	71.1
	Recall(%)	59.8	65.4	70.3
	$F_{1-score}(\%)$	67.6	69.1	70.9
OSCD	Average acc(%)	69.7	81.1	83.3
	Overall acc(%)	89.1	92.0	91.7
	Changed acc(%)	48.1	68.9	73.8

TABLE V
QUANTITATIVE COMPARISON AMONG DIFFERENT LOSS

Datasets	Metrics	BCE	DICE	ours
SZADA/1	Precision(%)	61.4	41.4	48.2
	Recall(%)	28.5	32.9	56.1
	$F_{1-score}(\%)$	38.9	36.7	51.8
TISZADOB/3	Precision(%)	90.2	97.2	93.4
	Recall(%)	97.7	82.6	91.5
	$F_{1-score}(\%)$	93.8	89.3	92.4
SZADA/2	Precision(%)	88.3	76.3	71.1
	Recall(%)	47.9	56.7	70.3
	$F_{1-score}(\%)$	62.1	65.1	70.9

the OSCD dataset, The structure with only 13 layers of network depth is quite weak, because the network is too shallow to fit the data. The experimental results show that reducing the network depth will inevitably reduce the network performance. This shows that it is necessary to introduce depthwise separable convolution, and the proposed method can effectively increase the depth and obtain a better performance model under the condition of limited data volume.

E. Perforcement of Different Loss

As described in Section II, we define a joint loss function that combines BCE and DICE. BCE solves the binary image classification task, whereas DICE is widely used for image segmentation. The image change detection task is not only a binary classification problem, but also a binary image segmentation problem. Using these loss functions alone is not enough for our prediction task. In this section, we prove the superiority of our joint loss function through experiments. For fair comparison, we only change the loss function and the results are shown in Table V and Fig. 9. It can be seen from Table V that our joint loss function perform better than BCE and DICE. According to Fig. 9, it is easy to see that the joint loss function can achieve more robust effects in segmentation, and DICE will miss some details of the picture.

F. Influence of Parameter ω

Except for the network structure, ω is an important parameter affecting the experimental result, it controls the preference of change detection result to precision or to recall rate. We set ω to 0.5, 1.0, and 2.0, respectively, to explore its impact on the evaluation metrics and the results are shown in Fig. 10 and Table VI. Because all algorithms behave poorly on SZADA/2 dataset, it is not meaningful to show the effect of parameter ω on the result, so set it aside. When ω is equal to 0.5, the

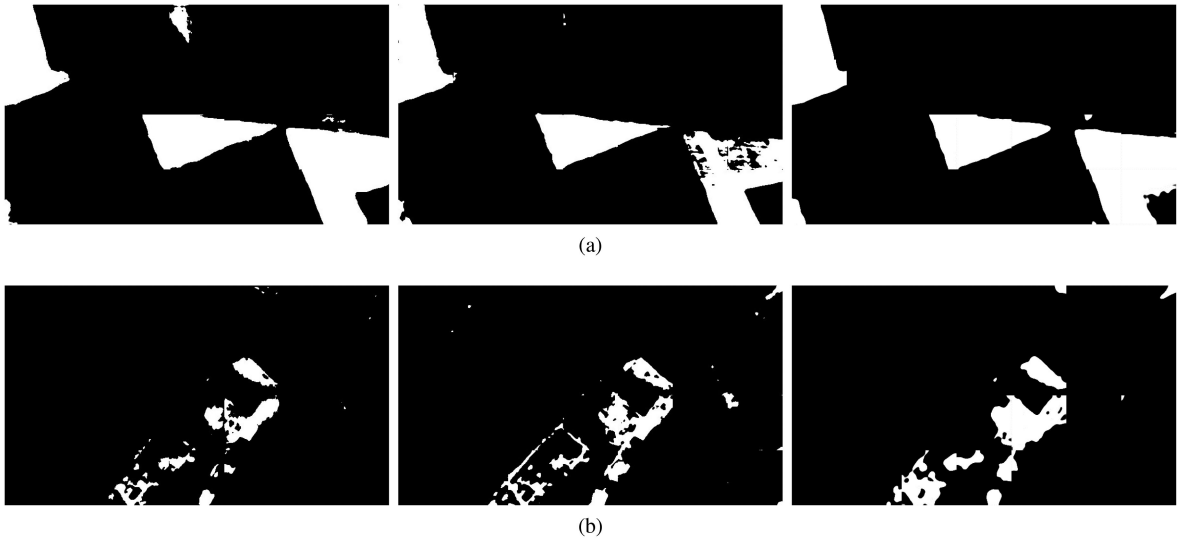


Fig. 9. Results by different loss. The leftmost is the result of BCE. The middle is the result of DICE. The rightmost is the result of joint loss. (a) TISZADOB/3 dataset. (b) SZADA/2 dataset.

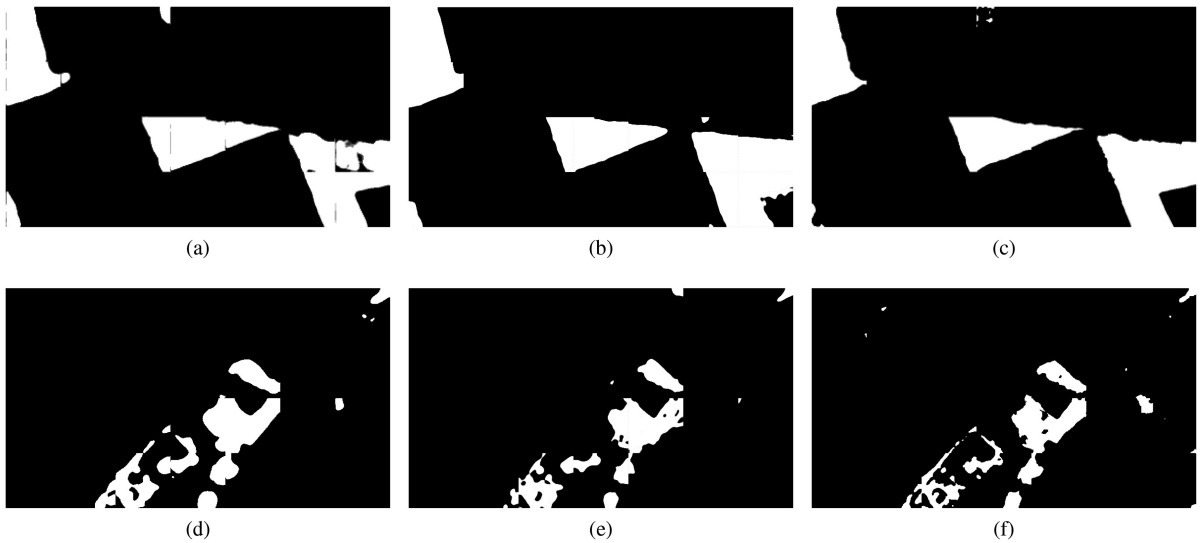


Fig. 10. Detection results by different ω on TISZADOB/3 and SZADA/2 datasets. (a) and (d) $\omega = 0.5$. (b) and (e) $\omega = 1.0$. (c) and (f) $\omega = 2.0$.

TABLE VI
QUANTITATIVE COMPARISON AMONG DIFFERENT SETTING G OF ω

Datasets	Metrics	0.5	1.0	2.0
TISZADOB/3	Precision(%)	92.6	93.4	97.5
	Recall(%)	91.9	91.5	87.0
	$F_{1-score}(\%)$	92.3	92.4	92.0
SZADA/2	Precision(%)	63.4	71.1	72.0
	Recall(%)	75.5	70.3	66.3
	$F_{1-score}(\%)$	68.9	70.7	69.0

recall rate is the maximum on both datasets compared to the other ω settings. But with the increase in ω, the recall rate is decreasing, the opposite of the case of precision. We can also see from Fig. 10, as the precision increases, that the image details are retained better, and ω equals 1 is a tradeoff between recall rate and precision.

IV. CONCLUSION

In this article, we proposed a supervised change detection method based on U-Net and depthwise separable convolution for optical aerial images, which requires very few parameters and less computation. Our proposed model is a more efficient approach, which tends to learn better feature representations with less data, but produces better-performing models. In addition, a loss function is designed that considers the change detection as pixel-level classification and segmentation simultaneously. With trained network, an input image pair is stacked into a

2c-channel image first, and then it was fed into the network, through U-Net model with depthwise separable convolution, a binary image is obtained. This method avoids the generation of difference map and postprocessing procedure, reduces the time consumption, and can handle large-scale change detection tasks. We then compared our methods with several latest algorithms and explored the influence of the traditional convolution and the depth separable convolution as well as the different depths on the detection results. The experimental results show that the proposed method can compete with the state-of-the-art methods and even performs better. We verified the effectiveness of our model in computational performance and speed. Our proposed approach can be further improved by considering pretrained encoders, such as VGG16 or other advanced pretrained networks instead of U-Net.

REFERENCES

[1] L. Miao, J. Chen, H. Tang, Y. Rao, Y. Peng, and W. Wu, "Land cover change detection by integrating object-based data blending model of Landsat and MODIS," *Remote Sens. Environ.*, vol. 184, pp. 374–386, 2016.

[2] J. Tian, S. Cui, and P. Reinartz, "Building change detection based on satellite stereo imagery and digital surface models," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 1, pp. 406–417, Jan. 2014.

[3] F. Bovolo, "A multilevel parcel-based approach to change detection in very high resolution multitemporal images," *IEEE Geosci. Remote Sens. Lett.*, vol. 6, no. 1, pp. 33–37, Jan. 2009.

[4] M. Gong, J. Zhao, J. Liu, Q. Miao, and L. Jiao, "Change detection in synthetic aperture radar images based on deep neural networks," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 1, pp. 125–138, Jan. 2016.

[5] P. Zhang, M. Gong, L. Su, L. Jia, and Z. Li, "Change detection based on deep feature representation and mapping transformation for multi-spatial-resolution remote sensing images," *ISPRS J. Photogrammetry Remote Sens.*, vol. 116, pp. 24–41, 2016.

[6] J. Liu, M. Gong, K. Qin, and P. Zhang, "A deep convolutional coupling network for change detection based on heterogeneous optical and radar images," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 29, no. 3, pp. 545–559, Mar. 2018.

[7] M. Gong, X. Niu, P. Zhang, and Z. Li, "Generative adversarial networks for change detection in multispectral imagery," *IEEE Geosci. Remote Sens. Lett.*, vol. 14, no. 12, pp. 2310–2314, Dec. 2017.

[8] 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.

[9] N. Lv, C. Chen, T. Qiu, and A. K. Sangaiah, "Deep learning and superpixel feature extraction based on sparse autoencoder for change detection in SAR images," *IEEE Trans. Ind. Informat.*, vol. 14, no. 12, pp. 5530–5538, Dec. 2018.

[10] 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.

[11] H. Lyu, H. Lu, and L. Mou, "Learning a transferable change rule from a recurrent neural network for land cover change detection," *Remote Sens.*, vol. 8, no. 506, pp. 1–22, 2016.

[12] L. Mou, L. Bruzzone, and X. Z. Xiao, "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.

[13] M. Gong, Y. Yang, T. Zhan, X. Niu, and S. Li, "A generative discriminatory classified network for change detection in multispectral imagery," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 12, no. 1, pp. 321–333, Jan. 2019.

[14] F. Gao, J. Dong, B. Li, and Q. Xu, "Automatic change detection in synthetic aperture radar images based on PCANet," *IEEE Geosci. Remote Sens. Lett.*, vol. 13, no. 12, pp. 1792–1796, Dec. 2016.

[15] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 1800–1807.

[16] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 640–651, Apr. 2017.

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

[18] C. Benedek and T. Szirnyi, "Change detection in optical aerial images by a multilayer conditional mixed Markov model," *IEEE Trans. Geosci. Remote Sens.*, vol. 47, no. 10, pp. 3416–3430, Oct. 2009.

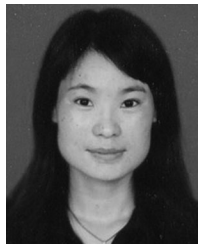
[19] R. C. Daudt, B. L. Saux, A. Boulch, and Y. Gousseau, "Urban change detection for multispectral earth observation using convolutional neural networks," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, 2018, pp. 2115–2118.

[20] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Represent.*, 2014, pp. 1–15.

[21] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," *J. Mach. Learn. Res. Proc. Track*, vol. 9, pp. 249–256, 2010.

[22] N. Chinchor and B. Sundheim, "MUC-5 evaluation metrics," in *Proc. Conf. Message Understanding*, 1993, pp. 69–78.

[23] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognit. Lett.*, vol. 27, no. 8, pp. 861–874, 2005.

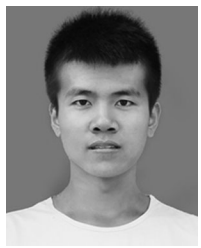


Ruochen Liu received the Ph.D. degree from Xidian University, Xi'an, China, in 2005.

She is currently a Professor with the Intelligent Information Processing Innovative Research Team of the Ministry of Education of China, Xidian University. Her research interest includes computational intelligence. Her areas of special interests include evolutionary computation, data mining, and deep learning.

Dawei Jiang received the B.S. degree from Tianjin Polytechnic University, Tianjin, China, in 2018. He is currently working toward the M.S. degree with Xidian University, Xi'an, China.

His current research focuses on deep learning and image processing.



Langlang Zhang received the B.S. degree in 2016 from Xidian University, Xi'an, China, where he is currently working toward the M.S. degree.

His current research focuses on remote sensing image change detection.

Zetong Zhang received the B.S. degree from the North University of China, Taiyuan, China, in 2017. He is currently working toward the M.S. degree with Xidian University, Xi'an, China.

His current research focuses on data mining.