


## Research Article

# Change Detection of Remote Sensing Images Based on Attention Mechanism

Long Chen,<sup>1,2</sup> Dezheng Zhang,<sup>1,2</sup> Peng Li,<sup>1,2</sup> and Peng Lv<sup>1,2</sup> 

<sup>1</sup>*School of Computer and Communication Engineering, University of Science and Technology Beijing (USTB), No. 30 Xueyuan Road, Haidian District, Beijing 100083, China*

<sup>2</sup>*Beijing Key Laboratory of Knowledge Engineering for Materials Science, Beijing, China*

Correspondence should be addressed to Peng Lv; [b1901775@ustb.edu.cn](mailto:b1901775@ustb.edu.cn)

Received 13 December 2019; Revised 13 February 2020; Accepted 2 June 2020; Published 25 August 2020

Academic Editor: Luca Manzoni

Copyright © 2020 Long Chen et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

In recent years, image processing methods based on convolutional neural networks (CNNs) have achieved very good results. At the same time, many branch techniques have been proposed to improve accuracy. Aiming at the change detection task of remote sensing images, we propose a new network based on U-Net in this paper. The attention mechanism is cleverly applied in the change detection task, and the data-dependent upsampling (DUpsampling) method is used at the same time, so that the network shows improvement in accuracy, and the calculation amount is greatly reduced. The experimental results show that, in the two-phase images of Yinchuan City, the proposed network has a better antinoise ability and can avoid false detection to a certain extent.

## 1. Introduction

Change detection in remote sensing images is a critical and challenging task, and its specific work refers to the quantitative analysis of multiple temporal remote sensing images for the same target area, determining the features and scope of surface changes and detecting the changed and unchanged parts [1]. Remote sensing image change detection is utilized to detect illegal buildings, water area supervision, natural disaster assessment, urban planning expansion research, and military reconnaissance [2].

Because of the increasing amount of data from remote sensing images and the increasing demand in this direction, manual comparison and analysis of the change area appear time-consuming and laborious. Due to factors such as seasons and solar illumination, imaging styles of different phases have huge differences [3], which make it difficult to solve the change detection task by computer vision.

Change detection is a unique task for remote sensing image processing, which can be regarded as a dichotomy problem of a region changing or not, as shown in Figure 1. Figure 1(a) shows the remote sensing images of a certain

region of Yinchuan City in 2015, Figure 1(b) shows the remote sensing images of this region in 2017, and Figure 1(c) shows the change label of this region, where black indicates that the location has not changed and white indicates that the location has changed. The task of change detection is to identify the changing areas in different phases.

General change detection methods are mainly based on autoencoders and feature extraction through the full connection between neurons [4]. But in fact, change detection can flexibly apply the method of semantic segmentation and extract features by convolution. CNNs have led the field in image processing since AlexNet [5] won the championship in 2012. With the advent of networks such as FCN [6], U-Net [7], and SegNet [8], the baseline effect in this field is getting better and better. However, due to the characteristics of the change detection task, the above-mentioned excellent networks often cannot exert the best results.

In this paper, our task is to solve the change detection in three districts of Yinchuan City. After a detailed analysis of this task, we found that the volume of the changed part is much smaller than that of the unchanged part. There is a greater degree of positive and negative sample imbalance

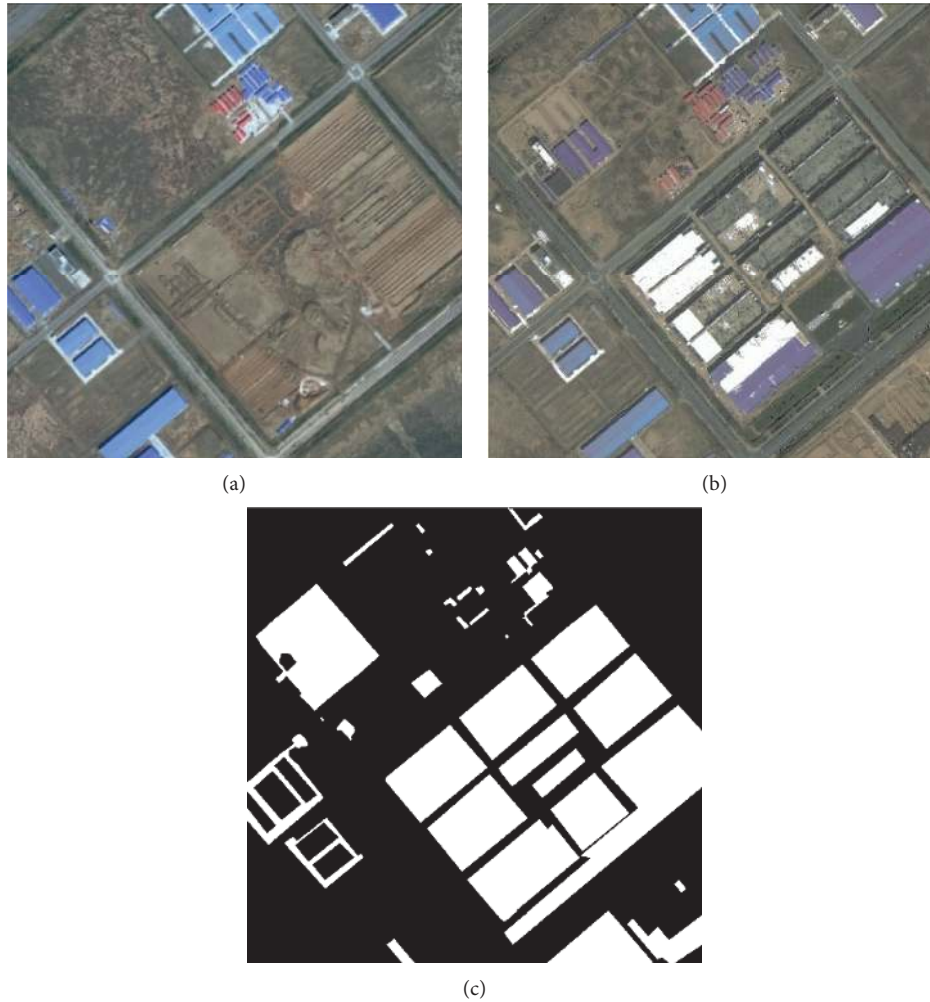


FIGURE 1: Example of remote sensing image change detection.

problem. To this end, we propose a new network based on U-Net. The network consists of encoder and decoder. Based on the residual attention model [9], we proposed a new attention mask structure for feature extraction, and a new encoder structure is also proposed in order to better perceive the changing area. By generating an attention mask, the model can pay more attention to the regions with obvious changes and improve the antinoise ability of the model. In the decoder stage, a data-dependent upsampling method (DUpsampling) [10] is used to replace the general upsampling method. The new upsampling method can be applied to smaller-resolution feature maps, which greatly reduces the computational complexity. At the same time, we propose a new loss function for the network, and the initial values of different loss functions are used to balance its impact on network training and reduce the impact of sample imbalance.

## 2. Related Work

**2.1. Convolutional Neural Network.** The achievement of today's success in image processing largely depends on the CNNs [11]. The essence of CNNs is a multilayer perceptron. The network structure includes a convolutional layer, a

downsampling layer, and a fully connected layer. The reason for its success is local connection and weight sharing method [12]. Reducing the number of weight makes the network easy to optimize and reduces the complexity of the model, which reduces the risk of overfitting. The earliest CNNs are time-delayed neural networks [13] and Lenet-5 [14]. After AlexNet won the champion of ILSVRC [15] in 2012, thanks to the support of GPU computing cluster, deep CNNs such as ZFNet [16], VGGNet [17], and GoogLeNet [18] became the winning algorithm of ILSVRC for many times. But at the same time, CNNs fail to converge with the deepening of network layers. ResNet [19] proposes the mechanism of residual learning, making the network easy to converge while getting deeper. However, the original CNN receptive field is small, and it cannot sense the neighborhood information well. Enlarging the receptive field will lead to a large increase in computing resources. At the same time, CNN's fully connected mode is too redundant and inefficient.

**2.2. Attention Mechanism.** Mnih et al. [20] confirmed the effectiveness of attention mechanism. Attention is generally classified into two types: one is top-down conscious

attention, called focus attention. The other is bottom-up unconscious attention, called saliency-based attention. Focus attention refers to the attention that has a predetermined purpose and focuses on a certain object actively and consciously [21, 22]. Saliency-based attention is also called stimulation-based attention [23]. Wang et al. [9] proposed a method to solve the problem of image classification by using attention residual learning.

*2.3. Encoder-Decoder Architectures.* Since FCN was proposed, people have been trying to use FCN to improve the accuracy of pixel-level prediction. On the one hand, people start with the atrous convolution [24, 25], which needs more complex operations, and on the other hand, people use encoder-decoder architectures. The most significant feature of encoder-decoder architectures is the ability to complete end-to-end learning. U-Net improves on the FCN framework by connecting codecs with skip connections to improve the effect. SegNet records the location of the maximum value during the maximum pooling operation of the encoder part and then realizes nonlinear upsampling through the corresponding pooling index in the decoder. DeepLab V3 [26] uses the ASPP structure to expand the receptive field, mining context information, and the improved Xception module to reduce the number of parameters and achieve the best effect of the current semantic segmentation network. Tian et al. [27] proposed a data-dependent upsampling method, which enables the encoder to sample down to the bottom layer and improve the accuracy by fusing features of different layers.

*2.4. Change Detection.* Remote sensing image change detection can greatly improve land utilization and contribute to urban planning and expansion. In the first decade of this century, CNN was rarely used in the field of change detection. Carincotte et al. [28] used a fuzzy hidden Markov chain algorithm to avoid a large number of false changes and missed detections caused by threshold segmentation. Liu et al. [29] used the stacked restricted Boltzmann machine to analyze the differential images between multiphase SAR images and classified the neighborhood features of the two-phase images. By using the deep learning algorithm, the images were classified pixel by pixel to achieve the purpose of change detection. In recent years, CNNs began to be implemented in this field. Desclee et al. [30] proposed an object-oriented forest vegetation change detection method, which firstly segments multitemporal high-resolution remote sensing images and then, based on the hypothesis chi-square test, identifies outliers of statistical differences in reflectivity and marks corresponding objects as changes. Qing et al. [31] applied Faster R-CNN to this field, greatly reducing the false changes of detection results. Ma et al. proposed a network based on multigrained cascade forest and multiscale fusion, so that the network can select image blocks of different sizes as input, thereby learning more image features [32]. Dong et al. designed a “Siamese samples” convolutional neural network to learn the semantic difference between changed and unchanged pixels [33].

The change detection in Yinchuan area includes many different landforms. Considering the particularity of this region, our method adds the attention module in the feature extraction stage and uses U-Net’s skip connection to further reduce the loss of information from upsampling and downsampling, it also uses DUpsampling to accelerate the upsampling process, while facilitating subsequent feature fusion. As a result, our model can avoid many false detections, while ensuring accuracy without occupying too many computing resources.

### 3. Our Approach

In this section, we first introduce the network we proposed and then elaborate on each functional module of the network, and we also propose a new loss function based on the problem we meet to improve the accuracy of the model. It is noteworthy that, different from the current mainstream of change detection method based on the convolution neural network, we improve the residual attention mechanism, and we proposed a new way to generate attention mask and apply the mask to change detection task. At the same time, the DUpsampling method is used to reduce the loss of computing resources. Finally, we noticed the impact of the initial size of the loss function on the network performance and proposed a new loss function, which achieved good results.

*3.1. Network Architecture.* In this paper, we propose a new network based on U-Net, as shown in Figure 2. Like U-Net, the network is divided into encoder and decoder. The encoder downsamples the input to extract the features, while the decoder upsamples the input to restore the resolution. The network uses a six-channel matrix superimposed on two pictures of different phases in the same area as input to the encoder. The encoder takes ResNet50 as the trunk and adds the attention modules. The attention module consists of trunk branch and mask branch, which perform feature extraction and mask generation on the input, respectively. The attention mask subtracts the three channels of two RGB images as the input and outputs a feature map with attention weight to highlight the key areas of feature extraction. Then residuals are performed on feature maps of mask branch and trunk branch. The DUpsampling method is used to replace the conventional bilinear upsampling on the decoder, which avoids the computation and memory footprint caused by reducing step size (such as DeepLab V3) of the decoder. The decoder maintains the original network structure of U-Net. There are skip connections between the encoder and the decoder to combine the features of the corresponding encoding during decoding.

#### 3.2. Functional Modules

*3.2.1. Encoder.* In order to improve the accuracy of the network, different from the traditional autoencoder network, we choose to deepen the network depth, thus

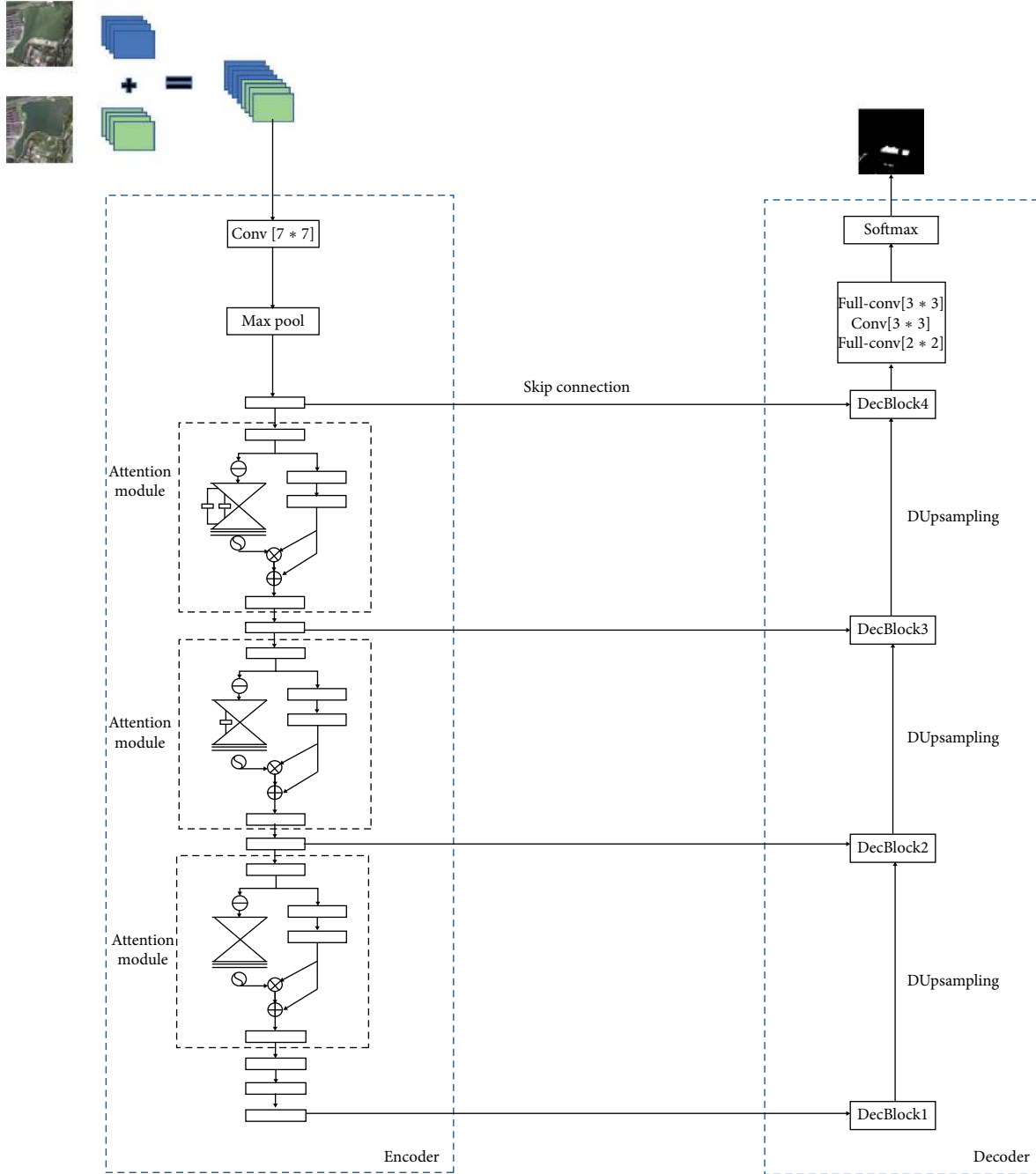


FIGURE 2: The framework of our proposed network.

introducing ResNet50. Its unique residual learning method can avoid the problem of unable to converge due to gradient explosion or gradient disappearance in deep network. In the encoder stage, ResNet50 as the backbone, and the residual attention mechanism is added to form a new encoder structure to guide the network to focus on areas with significant changes and improve the network's antinoise ability. The encoder includes three attention modules as shown in Figure 3. Each attention module is divided into mask branch and trunk branch. Trunk branch performs feature extraction just like normal convolutional neural networks. Mask branch is responsible for generating attention weights for

input features, and finally, residuals are performed on feature maps of mask branch and trunk branch. The formula is as follows:

$$H_{i,c}(x) = (1 + M_{i,c}(x)) * F_{i,c}(x), \quad (1)$$

where  $F_{i,c}(x)$  represents the features generated by the deep convolutional network,  $M_{i,c}(x)$  represents the output of the mask branch, and the value range of  $M_{i,c}(x)$  is  $[0, 1]$ . When it approaches 0,  $H_{i,c}(x)$  is approximately the original feature  $F_{i,c}(x)$ . The detailed structure is illustrated in Table 1.

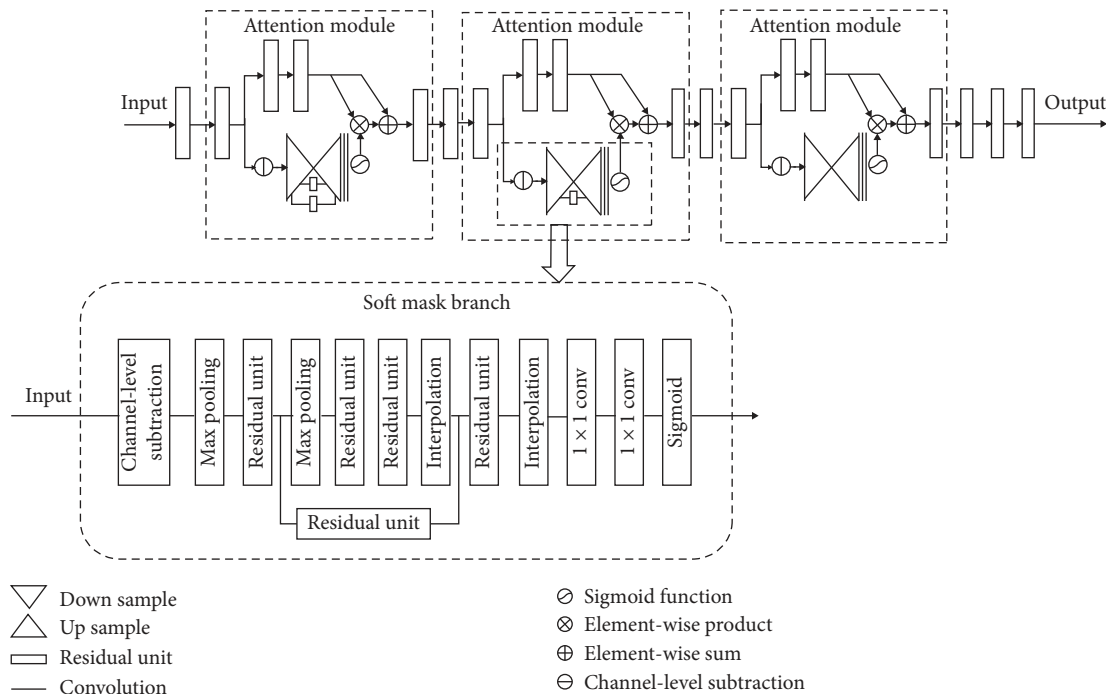


FIGURE 3: The framework with our proposed encoder which includes three attention modules and some residual units.

**3.2.2. Attention Module.** Because change detection requires comparing information of different phases in the same region, we propose a new architecture of the attention modules, which is different from semantic segmentation task when generating attention weight. In the mask branch, we subtract each channel of the original two feature maps to generate a new three-channel feature map. This step is called the channel-level subtraction. The new feature map is used to generate the attention mask and then put it into the mask branch for upsampling and downsampling and convolution, and the purpose of this step is to downsample to low resolution so that we can get strong semantic information. Each mask branch has a different number of residual units between upsampling and downsampling as skip connections to capture attention information at different scales. Each time pass the residual unit connected to the attention modules, and the size of the feature map is reduced. The architecture is displayed in Table 1.

**3.2.3. Decoder.** We keep the decoder structure of U-Net. The proposed network uses DUpSampling in the upsampling phase to replace the original bilinear upsampling procedure. The bilinear upsampling method does not take into account the correlation between the predicted pixels. DUpSampling uses the redundancy of segmentation labels to produce accurate segmentation results through the rough features generated by the encoder. And the encoder structure does not need to continue to excessively reduce the resolution of the feature map, thereby reducing the calculation time and memory usage. An important discovery is that the label of images is not independently and uniformly distributed, and the structural information it contains is related, so the label can be compressed without causing too much loss. So we

TABLE 1: The architecture details for our encoder.

Layer	Output size	Encoder
Conv1	$512 \times 512$	$7 \times 7$ , 64, stride 2
Max pooling	$256 \times 256$	$3 \times 3$ stride 2
Residual unit	$256 \times 256$	$\begin{pmatrix} 1 \times 1, 64 \\ 3 \times 3, 64 \\ 1 \times 1, 256 \end{pmatrix} \times 1$
Attention module	$256 \times 256$	Attention $\times 1$
Residual unit	$128 \times 128$	$\begin{pmatrix} 1 \times 1, 128 \\ 3 \times 3, 128 \\ 1 \times 1, 512 \end{pmatrix} \times 1$
Attention module	$128 \times 128$	Attention $\times 1$
Residual unit	$64 \times 64$	$\begin{pmatrix} 1 \times 1, 256 \\ 3 \times 3, 256 \\ 1 \times 1, 1024 \end{pmatrix} \times 1$
Attention module	$64 \times 64$	Attention $\times 1$
Residual unit	$32 \times 32$	$\begin{pmatrix} 1 \times 1, 64 \\ 3 \times 3, 64 \\ 1 \times 1, 256 \end{pmatrix} \times 3$

compress the label first and split label into multiple grids, and each grid size is  $t * t$  ( $t$  is the image size ratio, such as 16 and 32), and then we reshape the content of each grid into a vector  $v$ , then compress  $v$  into  $x$ , and stack  $x$  to get compressed labels. Formally, we have

$$x = Pv, \quad (2)$$

$$\bar{v} = Wx. \quad (3)$$

Linearly map  $v$  to  $x$  through  $P$ , and  $W$  is the inverse mapping matrix, which is the reconstruction matrix. Through the following formula to minimize the reconstruction error and optimize through SGD iteration, PCA

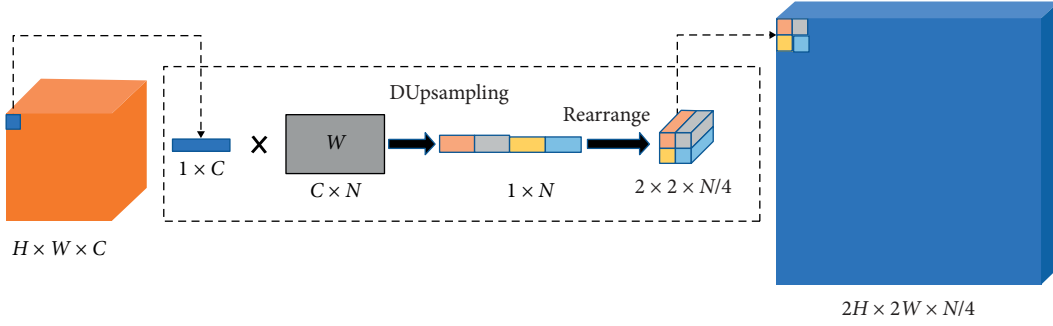


FIGURE 4: The specific process of DUpsampling.  $W$  is the reconstruction matrix, and there are three DUpsamplings in our network.

[34] can be used to obtain the closed solutions  $P$  and  $W$ . Formally,

$$P^*, W^* = \arg \min_{P, W} \sum_v v - \tilde{v}^2 = \arg \min_{P, W} \sum_v v - WPv^2. \quad (4)$$

When we get the reconstructed matrix  $W$ ,  $W$  is the parameter of the convolution kernel, which can complete the upsampling procedure, as shown in Figure 4.

At the same time, adaptive temperature softmax is introduced because the DUpsampling method may be calculated based on the one-hot label, so that the probability distribution is relatively smooth, resulting in too slow or even difficult convergence of loss in training [35], as shown in the following equation:

$$\text{softmax } z_i = \frac{\exp(z_i/T)}{\sum_j \exp(z_j/T)}. \quad (5)$$

$T$  can be learned automatically by the backpropagation algorithm without tuning. There is a skip connection between the encoder and the decoder so that the decoder can obtain the feature information in the encoder and reduce the loss of information in the decoding procedure. The decoder part contains four decoders, whose input includes the encoder map from skipping connection and the output of the previous layer. Through the upsampling procedure, the size of the feature map is doubled. Finally, two full convolutional layers and one convolutional layer are added, so that the size is enlarged again to achieve the effect of end-to-end training, as shown in Figure 2.

**3.2.4. Loss.** The two images of the change detection mission in the three districts of Yinchuan were taken in 2015 and 2017. When we look at the two phases of a total of 924 images and found that the volume of the changed part is much smaller than that of the unchanged part, change detection is essentially a binary classification problem, namely, one regional changes or not. There is a greater degree of positive and negative samples by imbalance problem. Therefore, we introduced the focal loss function [36], which can effectively solve the imbalance of positive and negative samples, as part of the loss function. Formally,

$$FL(p_t) = -\alpha_t (1 - p_t)^\gamma \log(p_t), \quad (6)$$

where  $p_t$  is the classification probability of different categories and  $\gamma$  and  $\alpha_t$  are fixed values. Here,  $\gamma = 2$  and  $\alpha_t = 0.25$ . Combined with the binary crossover entropy loss function  $L_{BCE}$ , which is commonly used in this task, combining the above two loss functions, we proposed a new loss function. First of all, we noticed that different types of loss functions have different initial values. Therefore, if we simply weighted the two loss functions, the loss function with a large initial value would dominate the loss function with a small initial value. We use the initial values of the two loss functions obtained in the first iteration to balance:

$$\text{loss} = \frac{c_1}{c_1 + c_2} FL(p_t) + \frac{c_2}{c_1 + c_2} L_{BCE}, \quad (7)$$

where  $c_1$  is the initial value of  $L_{BCE}$  and  $c_2$  is the initial value of focal loss. Experiments show that  $c_1$  is much larger than  $c_2$ , often several times larger than  $c_2$ , so that different loss functions will not distinguish the primary and secondary relationship due to different sizes. And then we add the two loss functions together to get the final loss, and with  $\alpha$  approximating 0.3, our network will achieve the best results. The loss function can be formulated as follows:

$$\text{loss} = \alpha \frac{c_1}{c_1 + c_2} FL(p_t) + (1 - \alpha) \frac{c_2}{c_1 + c_2} L_{BCE}. \quad (8)$$

## 4. Experiments

In this section, we first explain the datasets we used and the criteria we used to evaluate the effect of the method and then introduce the relevant experimental details. At the same time, we performed ablation experiments on the proposed loss function, verified the effectiveness of our loss function, and found the optimal value of  $\alpha$ . Finally, we compare the effect of the current mainstream network with our network.

**4.1. Datasets.** In our experiment, the datasets came from Gaofen-2 (Gf-2) satellite including Xixia District, Xingqing District, and Jinfeng District of Yinchuan City, known as Yinchuan three districts. The ground resolution of Gf-2 is 1 meter, indicating that per pixel represents one square meter. The image has four channels, namely, RGB channels and near-infrared channel. Since the year 2015 image has only RGB channels, we take RGB channels in both phases. The images we used have been irradiated and registered in

absolute terms, but not in relative terms. Therefore, we first performed histogram matching on the image to reduce the radiation difference between the two images, as shown in Figure 5.

Due to the particularity of change detection, we need to select the original datasets and eliminate the regions with no change or those with little change. Finally, 924 images from Gf-2 satellite were selected, covering different areas such as towns, agriculture, and industrial areas. Each image size was  $512 * 512$ . The datasets were divided into training set, verification set, and test set at 7:2:1, with 647, 184, and 92 pictures, respectively. Because of the small amount of dataset, U-Net, which has relatively loose requirements on dataset size, was chosen as the basic framework at the beginning of designing the network.

**4.2. Evaluation Criteria.** In our experiment, accuracy and  $F1$  value are used as the accuracy evaluation standard, and FLOPS is used as the efficiency evaluation standard. The calculation formula of accuracy is as follows:

$$\text{accuracy} = \frac{(\text{TP} + \text{TN})}{(\text{P} + \text{N})}, \quad (9)$$

where TP is the number of pixels with a positive detection number representing actual changes that are correctly detected and FP is the number of pixels with false detection number representing actual unchanged that are erroneously detected as changes. TP and FP correspond to TN and FN. TN is the number of pixels that have not actually changed but detected as changed and FN is the number of pixels that have actually changed but detected as not changed.  $\text{P} = \text{TP} + \text{FN}$  and  $\text{N} = \text{FP} + \text{TN}$ . The calculation formulas of precision and recall are as follows:

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \quad (10)$$

$$\text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \quad (11)$$

The precision represents the proportion of correctly detected change areas in the predicted results. The higher the precision, the less the false changes and noise, and recall represents the proportion of correctly detected change areas in the actual change areas. The higher the recall, the better the coverage of change detection results to the true change results. On the basis of these two important indicators, we calculated the weighted harmonic average of precision and recall. As a comprehensive indicator,  $F1$  value is used here:

$$F1 = \frac{2 \times \text{precision} \times \text{recall}}{\text{precision} + \text{recall}}. \quad (12)$$

FLOPS is an abbreviation of floating-point operations per second. It is often used to estimate the performance of a network.

**4.3. Experimental Details.** There are a number of areas in our datasets that have not changed, so we filter the datasets. The size of each picture is  $512 * 512$ , which is about 250,000

pixels. We consider pictures with a total number of changed pixels less than 1000 as unchanged and removed. In order to enhance the generalization of the network, after histogram matching of the images and reducing the imaging difference, we carry out random geometric transformation and color adjustment of the images in a small range and normalize each channel value of the processed remote sensing images, so that the network could rapidly converge. We use the Adam optimizer [37] to train 1000 epochs on the network, and the initial learning rate is designed to be  $1e-3$ . After each epoch, the learning rate dropped and  $\alpha$  in the loss function is set to 0.3. We observe that our networks generally converged after 200 epochs.

**4.4. Loss Function Ablation Experiments.** To prove the effectiveness of our loss function, we performed two experiments in this section. First, use formulas (7) and (13) as loss functions, respectively, to compare the effect of initial value balance on experimental results:

$$\text{loss} = FL(p_t) + L_{BCE}, \quad (13)$$

As mentioned above, the initial value of  $L_{BCE}$  is much larger than the focal loss, so during the training process,  $L_{BCE}$  will occupy the dominant position. The datasets are mentioned in Section 4.1. The experimental details are mentioned in Section 4.3. The initial learning rate is set to  $1e-3$ . The Adam optimizer is used to train 1000 epochs on multiple networks. The experimental results are shown in Table 2. At the same time, the network has significantly improved the convergence speed after the initial value balance.

After confirming the effectiveness of the initial value balance, we began to explore the optimal value of  $\alpha$ . When  $\alpha = 0$ , it means that only  $L_{BCE}$  works, and when  $\alpha = 1$ , only focal loss works. The results of the comparative experiments are as follows.

It can be seen from Figure 6 and Table 3 that, for our problem, when  $\alpha = 0.3$ , the network has the best performance. The reason may be that the imbalance of positive and negative samples was improved after the introduction of residual attention mechanism, so the effect of focal loss was not optimal. In the following experiments, we set  $\alpha$  to 0.3 by default and compared it with other networks.

**4.5. Experimental Results.** We select three mainstream networks to compare with our network on our datasets (mentioned in Section 4.1), including the stack autoencoder based on deep confidence network [38], U-Net, and PSPNet [39]. All networks are trained and tested on our datasets without using pretrained models. As shown in Figure 7, from left to right are the year 2015 images, the year 2017 images, attention mask, labels, and the results of different networks. The images on display include different scenes of cities, towns, fields, and bare grounds. As shown in the figure, our model is superior to other models in most cases. Table 4 shows the performance of the above networks and ours on  $F1$  value, recall, and accuracy. As shown in the table,

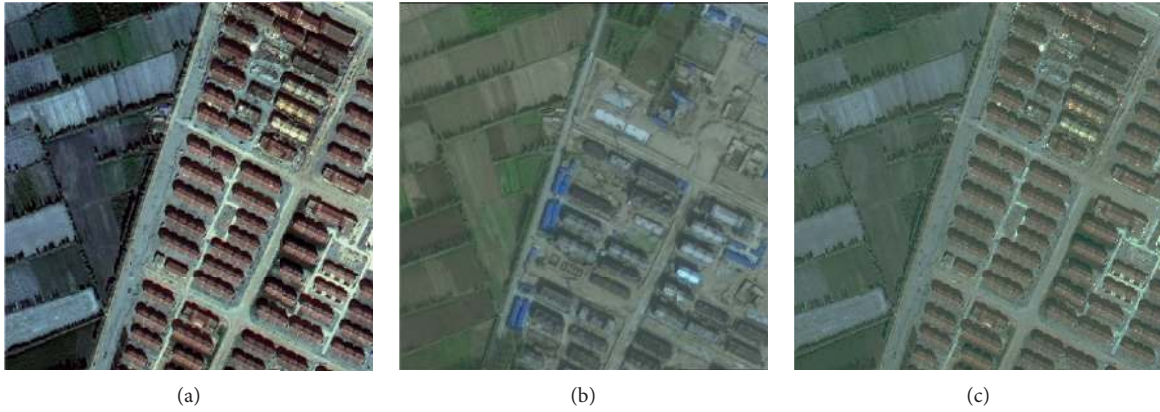


FIGURE 5: Effect before and after histogram matching. (a) The original images of 2017, (b) the original images of 2015, and (c) the processed image of 2017.

TABLE 2: Comparison of the effects of two loss functions on the network.

Method	<i>F1</i>	Recall	Pixel accuracy	Convergence epoch
Formula (7)	<b>0.615</b>	<b>0.608</b>	<b>0.693</b>	<b>340</b>
Formula (13)	0.575	0.554	0.574	570

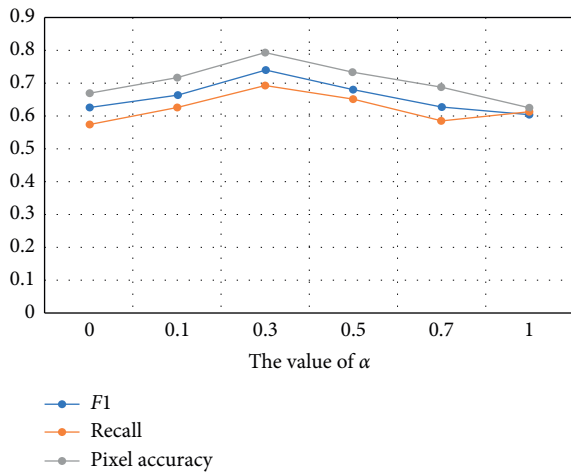


FIGURE 6: The effect of different values of  $\alpha$  on the experimental results. It proves that our method has an optimal solution when  $\alpha = 0.3$ .

TABLE 3: Comparison of *F1*, recall values, and pixel accuracy values with different values of  $\alpha$ .

$\alpha$	<i>F1</i>	Recall	Pixel accuracy
0	0.628	0.576	0.672
0.1	0.666	0.629	0.717
<b>0.3</b>	<b>0.743</b>	<b>0.697</b>	<b>0.796</b>
0.5	0.686	0.653	0.735
0.7	0.626	0.584	0.694
1	0.607	0.617	0.627

the networks based on the convolution are superior to the stack autoencoder in our datasets, and PSPNet has a strong ability of context information so that it performs better than

ordinary U-Net, and our network is 1.9% higher than PSPNet in *F1* and 5.1% higher in accuracy. The performance of *F1* and accuracy verifies that our network and the proposed loss functions are effective. At the same time, the FLOPS value of our network is  $4.5 \times 10^9$ , which is far lower than PSPNet and U-Net. It shows that our method has high efficiency while ensuring accuracy and reduces waste of computing resources. It is worth mentioning that in the attention modules, our network will subtract the corresponding channels of input to generate the attention mask. In the figure, we will visualize the attention mask, and the attention mask will optimize the network training process, thus greatly improving the antinoise ability of our network. Thanks to the attention mask, our network can avoid many false detections, such as the third set of pictures. We analyzed in detail that our task for two-phase imaging style difference is huge, so we need to do histogram match in the pretreatment stage, but after the match, color distortion will happen, for example, the third group pictures had dark green fields into a dark purple, which greatly affected the judgment of the network, excellent network like PSPNet is also unable to avoid this problem. In the fifth group pictures, we can see that the attention mask divides the left variable area into two parts, and our network can make correct judgments, while most other networks are unable to. We also observe some disadvantages from the experiment, such as the fourth group experiment, mask only focuses on the middle of the road section and ignores the rest of the change, so the module cannot detect the change of the upper left corner buildings like PSPNet; hence, generating accurate attention mask is the key to the success of our network. At the same time, because we adopt the DUpsampling method, different from the ordinary method, the training speed will be very fast. This experiment only took about 30 hours to train on the latest GPU. At the same time, it is also convenient for feature fusion of different output layers [26] to improve the accuracy. Experiments show that the accuracy of our proposed network is greatly improved compared with the baseline. In complex scenes such as densely built towns, thanks to the attention mechanism, our network can reduce false detection and missed detection and show a better effect.



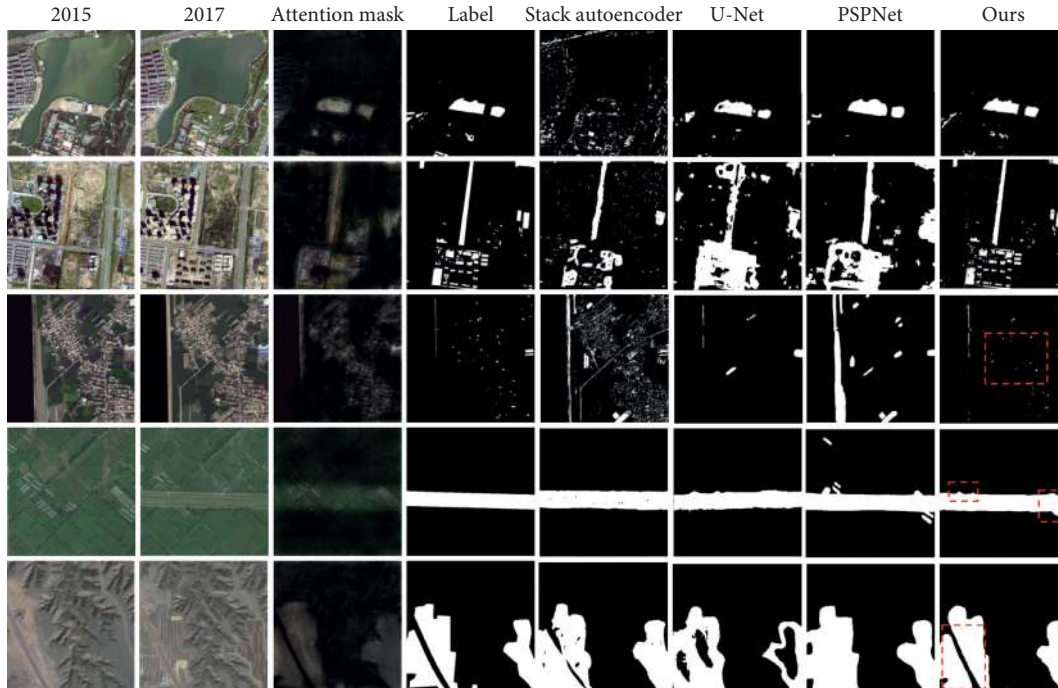


FIGURE 7: The attention mask and extraction results of the stack autoencoder, U-Net, PSPNet, and ours.

TABLE 4: Comparison of  $F1$ , recall values, and FLOPS values with the stack autoencoder, U-Net, PSPNet, and ours on the test datasets.

Module	$F1$	Recall	Pixel accuracy	FLOPS $\times 10^9$
Stack autoencoder	0.463	0.391	0.568	3.2
U-Net	0.689	0.691	0.687	5.6
PSPNet	0.724	0.704	0.745	11.2
Ours	0.743	0.697	0.796	4.5

## 5. Conclusion

In this paper, we propose a new network based on U-Net by combining the skip connection structure with the advanced residual attention mechanism and the DUpsampling method, and we also propose a new loss function suitable for application scenarios. Our network is applied to Yinchuan change detection task, and experiments show that compared with the current change detection methods, our network and loss function have improved accuracy and  $F1$  value without wasting excessive computing resources. At the same time, our method has a strong antinoise ability and certain robustness which can better solve the change detection task of Yinchuan City.

## Data Availability

The data used to support the findings of this study are available from the corresponding author upon request, but for study only, not for commercial use.

## Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

## Authors' Contributions

This work was completed in cooperation with Long Chen, Peng Li, Dezheng Zhang, and Peng Lv. Among them, Long Chen and Peng Li designed and proposed a new type of network structure. Dezheng Zhang provided innovative thinking for the attention module so that the method can be used in the direction of change detection. Peng Lv proposed constructive opinions on the final loss function to achieve better results, while optimizing the upsampling structure makes it possible to quickly achieve end-to-end output.

## Acknowledgments

This work was supported by Ningxia Hui Autonomous Region Key Research and Development Program of Ningxia Hui Autonomous Region under Grant 2019BFG02009 (Key Technologies for Intelligent Monitoring of Spatial Planning Based on High-Resolution Remote Sensing).

## References

- [1] L. I. Deren, *Change Detection from Remote Sensing Images*, Editorial Board of Geomatics & Information Science of Wuhan University, Wuhan, China, 2003.
- [2] C. Zhang, S. Wei, S. Ji, and M. Lu, "Detecting large-scale urban land cover changes from very high resolution remote sensing images using CNN-based classification," *ISPRS International Journal of Geo-Information*, vol. 8, no. 4, p. 189, 2019.
- [3] W. Ma, Y. Xiong, Y. Wu, H. Yang, X. Zhang, and L. Jiao, "Change detection in remote sensing images based on image mapping and a deep capsule network," *Remote Sensing*, vol. 11, no. 6, p. 626, 2019.

- [4] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural Computation*, vol. 18, no. 7, pp. 1527–1554, 2006.
- [5] A. Krizhevsky, I. Sutskever, and G. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proceedings of the Conference on Neural Information Processing Systems NIPS*, Curran Associates Inc, Red Hook, NY, USA, December 2012.
- [6] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 39, no. 4, pp. 640–651, 2014.
- [7] O. Ronneberger, P. Fischer, and T. Brox, "U-net: convolutional networks for biomedical image segmentation," in *Proceedings of International Conference on Medical Image Computing and Computer-Assisted Intervention*, Springer, Munich, Germany, October 2015.
- [8] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 12, pp. 2495–2481, 2015.
- [9] F. Wang, M. Jiang, C. Qian et al., "Residual attention network for image classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3156–3164, Honolulu, HI, USA, July 2017.
- [10] Z. Tian, T. He, C. Shen, and Y. Yan, "Decoders Matter for Semantic Segmentation: Data-dependent Decoding Enables Flexible Feature Aggregation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, CA, USA, January 2019.
- [11] Y. Lecun, J. S. B. Boser, R. E. HowardHenderson, W. Hubbard, and L. D. Jackel, "Backpropagation applied to handwritten zip code recognition," *Neural Computation*, vol. 1, no. 4, pp. 541–551, 1989.
- [12] Z. Fei-Yan, J. Lin-Peng, and D. Jun, "Review of convolutional neural network," *Chinese Journal of Computers*, vol. 40, no. 1, pp. 1–23, 2017, in Chinese.
- [13] A. Waibel, G. T. Hanazawa, and K. J. LangShikano, "Phoneme recognition using time-delay neural networks," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 37, no. 3, pp. 328–339, 1989.
- [14] Y. Lecun, Y. L. Bottou, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.
- [15] O. Russakovsky, H. J. Deng, H. Su et al., "ImageNet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.
- [16] M. D. Ma and R. Fergus, "Visualizing and Understanding Convolutional Networks," in *Computer Vision – ECCV 2014*, Springer, Berlin Germany, 2013.
- [17] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proceedings of the International Conference on Learning Representations (ICLR)*, pp. 1409–1556, San Diego, CA, USA, May 2015.
- [18] C. Szegedy, W. Liu, Y. Jia et al., "Going deeper with convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, MA, USA, June 2014.
- [19] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, June 2015.
- [20] V. Mnih, N. Heess, A. Graves et al., "Recurrent models of visual attention," in *Proceedings of the Conference on Neural Information Processing Systems NIPS*, Vancouver, Canada, December 2014.
- [21] K. Xu, "Show, attend and tell: neural image caption generation with visual attention," vol. 3, p. 5, 2015, <https://arxiv.org/abs/1502.03044>.
- [22] J. Lu, C. Xiong, D. Parikh, and R. Socher, "Knowing when to look: adaptive attention via a visual sentinel for image captioning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, November 2017.
- [23] P. Anderson, X. He, C. Buehler et al., "Bottom-up and top-down attention for image captioning and visual question answering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6077–6086, Salt Lake City, UT, USA, June 2018.
- [24] L. C. Chen, G. Papandreou, I. Kokkinos et al., "Deeplab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 4, pp. 834–848, 2017.
- [25] C. Peng, X. Zhang, G. Yu et al., "Large kernel matters--improve semantic segmentation by global convolutional network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4353–4361, Honolulu, HI, USA, June 2017.
- [26] L. C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," 2017, <https://arxiv.org/abs/1706.05587>.
- [27] Z. Tian, T. He, C. Shen, and Y. Yin, "Decoders matter for semantic segmentation: data-dependent decoding enables flexible feature aggregation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3126–3135, Long Beach, CA, USA, June 2019.
- [28] C. Carincotte, S. Derrode, and S. Bourennane, "Unsupervised change detection on SAR images using fuzzy hidden Markov chains," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 44, no. 2, pp. 432–441, 2006.
- [29] J. Liu, M. Gong, J. Zhao, H. Li, and L. Jiao, "Difference representation learning using stacked restricted Boltzmann machines for change detection in SAR images," *Soft Computing*, vol. 20, no. 12, pp. 4645–4657, 2016.
- [30] B. Desclée, P. Bogaert, and P. Defourny, "Forest change detection by statistical object-based method," *Remote Sensing of Environment*, vol. 102, no. 1-2, pp. 1–11, 2006.
- [31] W. Qing, Z. Xiaodong, C. Guanzhou, F. Dai, Y. Gong, and K. Zhu, "Change detection based on Faster R-CNN for high-resolution remote sensing images," *Remote Sensing Letters*, vol. 9, no. 10, pp. 923–932, 2018.
- [32] W. Ma, H. Yang, Y. Wu et al., "Change detection based on multi-grained cascade forest and multi-scale fusion for SAR images," *Remote Sensing*, vol. 11, no. 2, p. 142, 2019.
- [33] H. Dong, W. Ma, Y. Wu et al., "Local descriptor learning for change detection in synthetic aperture radar images via convolutional neural networks," *IEEE Access*, vol. 7, pp. 15389–15403, 2019.
- [34] S. Wold, K. Esbensen, and P. Geladi, "Principal component analysis," *Chemometrics and Intelligent Laboratory Systems*, vol. 2, no. 1-3, pp. 37–52, 1987.
- [35] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," 2015, <https://arxiv.org/abs/1503.02531>.
- [36] T. Y. Lin, P. Goyal, R. Girshick et al., "Focal loss for dense object detection," in *Proceedings of the IEEE International*

- Conference on Computer Vision*, pp. 2980–2988, Venice, Italy, December 2017.
- [37] D. P. Kingma and J. Ba, “Adam: a method for stochastic optimization,” 2014, <https://arxiv.org/abs/1412.6980>.
- [38] G. E. Hinton, “Reducing the dimensionality of data with neural networks,” *Science*, vol. 313, no. 5786, pp. 504–507, 2006.
- [39] H. Zhao, J. Shi, X. Qi et al., “Pyramid scene parsing network,” in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2881–2890, Honolulu, HI, USA, November 2017.