Improved U-Net Network Segmentation Method for Remote Sensing Image

Letian Zhong[1] ,Yong Lin[1] ,Yian Su[1] ,Xianbao Fang[1]
1. College of Electrical and Automation Engineering, Hefei University of Technology, Hefei, China
2371725573@qq.com, linyong@hfut.edu.cn, 384238329@qq.com, 1071389132@qq.com
Corresponding Author: Yong Lin Email:linyong@hfut.edu.cn

2022 IEEE 6th Advanced Information Technology, Electronic and Automation Control Conference (IAEAC) | 978-1-6654-5864-1/22/$31.00 ©2022 IEEE | DOI: 10.1109/IAEAC54830.2022.9929616

***Abstract*—Semantic segmentation and extraction based on remote sensing images has important theory and significance. Deep learning has become one of the mainstream methods to extract information from remote sensing images. In this paper, based on the improvement of U-Net network structure, we combine ASPP and skip connection. Improve the residual module to improve the information extraction method. The main improvements of this paper are:① Based on the U-Net network structure, we use the multi-scale feature detection capabilities of Pyramid to introduce. The ASPP module and the residual structure are improved, paying more attention to semantic and detail informatization, overcoming the limitations of U-Net in small target detection; ② We have improved the U-Net network, using skip connections to get more layers of information. Experiments show that the model proposed in this paper has significantly higher MPA and MIOU than the U-Net model on both the VOC dataset and the Vaihingen dataset. It means that ARU-Net can extract information better.**

Keywords—semantic segmentation; U-Net; Pyramid; ASPP; Residual block; Remote sensing image; Skip connection

I. INTRODUCTION

With the development of science and technology, the cost of obtaining remote sensing images is getting lower and lower, which leads to that each remote sensing platform can obtain more remote sensing data than before[1]. Common remote sensing data include: high, medium and low resolution data, hyperspectral data, radar data, laser data and so on. However, in remote sensing data, the road and other background feature elements in remote sensing images are usually complex and diverse, so there are still limitations such as accuracy in semantic segmentation and feature extraction of road information. However, most of them still stay at the level of artificial visual recognition, usually only regard remote sensing images as basic information, and do not extract and process information. However, with the development and deepening of deep learning technology, it becomes more important to use deep learning to extract and classify the target information in remote sensing images. Traditional image segmentation methods include threshold segmentation[2], edge segmentation[3], region segmentation[4] and so on. These methods divide the pixels into several classes according to the differences of the foreground and background in gray, shape, texture and other low-level features, so as to achieve the separation of foreground and background. However, the remote sensing image contains rich information and the target scale is complex, so it can not be segmented by traditional methods like the simple image. And the final result of the traditional image segmentation method has no semantic annotation. In recent years, with the continuous development of deep learning and the powerful feature extraction ability of neural network, the semantic segmentation model based on deep learning has gradually become the mainstream remote sensing image segmentation method. Not only the accuracy but also the speed has been significantly improved[5].

In 2015, Long J et al. proposed a fully convolutional neural network (FCN)[6], which realizes end-to-end prediction at the pixel level of an image by converting a fully connected layer into a convolutional layer, thus classifying each pixel of the image and solving the problem of image segmentation at the semantic level. FCN can accept input images of any size, and it is also the basic skeleton of most of the subsequent semantic segmentation methods. However, the result of FCN segmentation is not fine enough to segment the details of the target image, which leads to the result of semantic segmentation is not accurate enough. On the basis of FCN, SegNet[7] is more efficient than FCN by using encoding-decoding structure and skip connection similar to U-Net, and using maximum pooling. The U-Net model proposed by Ronneberger[8] and others adds more skip connections on the basis of FCN, and reuses the low-level features of the encoding stage in the decoding stage, so as to integrate the high-level and low-level semantic features of the network, thus achieving better extraction results. Based on the U-Net network, Guo[9] et al. combined the multiple loss method with the attention mechanism to improve the segmentation accuracy of small size targets in remote sensing images. On the basis of U-Net, Ibtehaz N and Rahman M S[10] use asymmetric convolution blocks and index pooling to improve the recognition ability of multi-directional targets in remote sensing images. On the basis of U-Net, Xu Jiawei and others[11] added pyramid pooling module, residual module and convolution block attention module to establish PRCU-Net model, which can pay attention to more semantic information and detail information, and make up for the lack of U-Net for small target detection. Liu Tongxin et al[12]. Proposed the ASU-Net model based on U-Net network. By adding the channel attention mechanism in the encoding part, adding the spatial pyramid pooling module behind the last convolution layer of the encoder, and finally adding the spatial attention mechanism in the decoder part, the new model designed in this way has excellent network performance, higher segmentation accuracy and is more suitable for remote sensing image segmentation.

Based on the U-Net model, this paper proposes a new network model for remote sensing image semantic segmentation. Specific contribution is in the following 3 respects: 1. In order to fully obtain the global information of multi-scale remote sensing ? Enhanced the ability to extract objects of different scales and shapes, and added an improved ASPP[13] module. 2. Adding a part of skip connection in the coding part , thus alleviating the problem of lost information in upsampling and having a good multi-scale property. 3. The improved residual structure is designed. It can retain more original features, improve the feature extraction ability of the network, and enhance the proportion of effective features.

II. Improved U-Net Network

A. *The Network of Network*

As shown in the figure 1, ARU-Net is an improved network based on the U-Net network architecture. An improved residual block (ResBlock) and spatial pyramid pooling (ASPP) are added to the original U-Net network, and the decoding part improves the network performance through skip connection. ARU-Net continues the structural characteristics of U-Net and is left-right symmetrical. The left part is the coding network, which is downsampled four times to reduce the spatial dimension of the data and obtain the deep language information. The residual block (ResBlock) is a combination of Batch Normalization, activation function (Relu), and convolutional layers. The BN layer can standardize the data, which can make the network training easier and accelerate the convergence speed. While the Relu activation function avoids gradient explosion or gradient disappearance. ASPP connects the encoding path and the decoding path, and it is also the last layer in the encoding network. The feature map after the convolution layer contains a lot of spatial semantic information through multiple convolution kernels. The feature map is input into the improved ASPP module, and different semantic information is extracted by hole convolution with different expansion rates, and then concat operation is carried out, and finally input into the 1 * 1 convolutional network, which is convenient for describing image features at different scales. Such a coding network has finer segmentation positioning. Not only the low-level network can extract the detailed features of the required image, but also the high-level network can get useful feature information, thus improving the efficiency and accuracy of the coding network. The right part is the decoding network. After four upsampling operations, the deep features obtained by the encoding network are upsampled to the required size. The first two decoding parts of the right half part respectively have two inputs: the shallow information of the corresponding network of the left half part is fused with the deep feature information obtained by deconvolution of the upper layer by using skip connection. It alleviates the problem of lost information in upsampling and has good multi-scale characteristics. The deep features obtained by the first and second decoding modules are added to the latter two decoding modules respectively by adding skip connections in the decoding network, and the features of different scales are further combined and utilized. When the feature map is upsampled to the size of an input picture, it is input to a 1 × 1 convolution and sigmoid activation function module to obtain an extraction map.

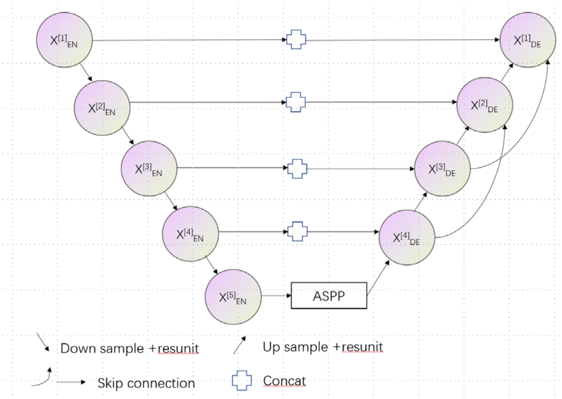

Fig. 1. Structure of the ARU-Net

B. *Improved ASPP module*

The traditional pooling operations are usually maximum pooling, average pooling or random pooling. Although the pooling operation can reduce the dimension of the image, it will inevitably lose some spatial information. As a result, the original information of the image can not be well restored in the later decoding part. However, He Kaiming[14] once pointed out that the fully connected layer needs to fix the size of the input image, resulting in unnecessary loss of accuracy, and proposed the pyramid pooling module. (SPP), so that the convolutional network of the fully connected layer can adapt to various sizes of the input picture. In other words, the meaning of spatial pyramid pooling is to transform feature maps of any size into feature vectors of demand size. However, PPM, SPP and FPN are not good at representing the global information of images, for example, SPP can not fully reflect the semantic relationship between global information and local information. However, research also has shown that hole convolution can effectively increase the visual field. That is to say, under the same computing resources, the use of hole convolution can replace the larger number or size of convolution kernels. Therefore, this paper adds the hole space pyramid pooling (Atrous Spatial Pyramid Pooling) module in the network. For a training image, at any pixel I, the input signal is X, w is the convolution filter, K is the convolution kernel size, the expansion ratio of the input image sampling step is R, and the output of the ith network on the output feature map is y [I]. Then the calculation formula is as follows:

$$y[i] = \sum_{k=1}^{K} x[i + r \cdot k]\omega[k] \qquad (1)$$

In this paper, the ASPP network is improved, and the model is shown in the figure 2. The ASPP module in this paper is composed of four parallel hole convolutions. The first is a 1*1 convolution block, and then three hole convolutions with a convolution kernel size of 3*3 are connected in parallel, where the expansion rates are set to 1, 3 and 5 respectively. After the convolution of these three holes, a 1*1 convolution block, batch normalization and activation function are added. In this way, the previous layer can be associated with a wider field of view on the local features, so as to prevent the small target features from being lost in the information transmission. The first branch of the network is a 1*1 standard convolution, which can keep the original receptive field. The second branch to the fourth branch are respectively convolution blocks with different expansion rates and can perform feature extraction to obtain local features with different scales, and the last branch directly pools the input global average to obtain the global features. Finally, the five branches are concatenated into a concat and a 1*1standard convolution is performed to fuse the information of different scales. And that final characteristic map is obtain and sent to a subsequent network part.

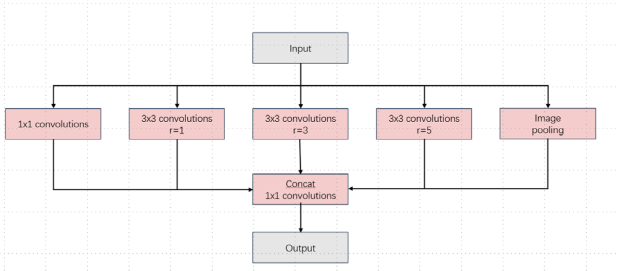

Fig. 2. Improved ASPP module structure

C. Residual module

ResNet is a deep network structure obtained by stacking multiple residual network modules. Compared with the general network model, the network model has a very deep depth. Generally speaking, different layers of convolutional neural network will extract different information and generate different feature representations, such as texture information in the lower layer and shape information in the higher layer. The more layers there are in the network, the more information can be obtained at different levels, and the more representations can be obtained by combining them, and the richer information can be obtained. In other words, the more layers and the deeper structure of the neural network model, the more parameters it can have, so that it can learn to extract more complex features and obtain better learning ability. However, before ResNet was proposed, in fact, if the depth of the network was simply superimposed, the performance of the network could not be improved, and even the performance was not as good as that of the shallow network. This is because as the depth of the network increases, the value of the loss function of the whole model is difficult to effectively transfer to each layer of the network structure, which will make the loss value of the function increase or decay exponentially with the number of layers, resulting in gradient explosion or disappearance. Forming a so-called network degradation problem. The residual network adds a direct mapping between different layers by adding a Shortcut to the convolution layer, which makes the latter layer have more abundant image information, improves the efficiency of information dissemination, and reduces the problems of gradient disappearance explosion and network degradation.

As shown in the figure3, the basic architecture of ResNet is to add direct edges, also known as shortcut branches, on one side of the basic convolutional network block.

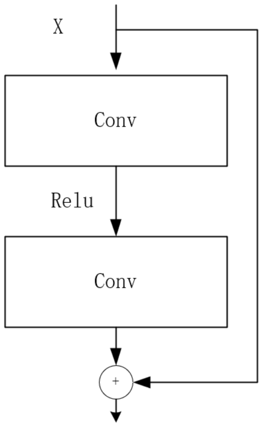

Fig. 3. Traditional residual module

In this way, the efficiency of information dissemination can be improved. By adding direct edges, the network can better complete the identity mapping. When the identity mapping needs to be completed, it can be completed directly through the direct edges. There is no need for the convolutional network to learn to complete the identity mapping. The convolutional network only needs to approach 0. In this way, it is easier to achieve the goal, thus alleviating the problem of model degradation. And because of the existence of the shortcut branch, when the error is reversely transmitted according to the chain rule, the cross-layer transmission is realized, so that the previous network parameters have a larger gradient value, and the gradient disappearance is alleviated.

The figure 4 shows the residual improvement residual network based on ResNet in this paper. In this paper, the output part of the first convolution of the convolution module is feedback connected, and the extracted features obtained in the middle are added to the input information, and then the convolution features are extracted. It can not only enhance the extracted features, but also refine the input information, which reduces the loss of effective information with the increase of convolution layers. In the actual image task, the information of the first convolution block is returned to the initial input through skip connection, which not only retains the original features, but also improves the feature extraction ability of the network and enhances the proportion of effective features.

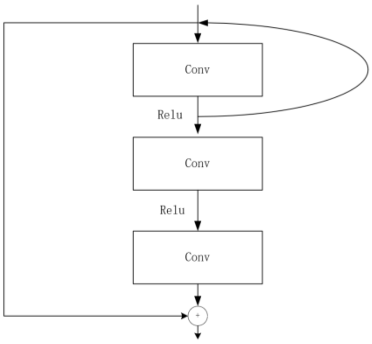

Fig. 4. Improved residual module

III. EXPERIMENTS AND RESULT ANALYSIS

A. Data set and data preprocessing

（1）PASCAL VOC2012 Dataset

The VOC2012 Dataset comes from the image competition held by PASCAL VOL, which is usually used for image classification, object detection, image segmentation and other tasks. The normal number of categories in VOC2012 data set is 20. They are usually divided into the following four categories:

People

Animals: cats, birds, cows, sheep, dogs, horses

Indoor tools: bottle, dining table, chair, plant, sofa, TV

Transportation: plane, ship, bicycle, bus, car, motorcycle, subway

The data set contains 5995 VOC training samples and their corresponding labels, which are randomly divided into 5000 training samples, 500 validation samples and 495 test samples. Because the size of some images is not consistent, it needs to be resized. Here, the image is resized with gray bars to avoid distortion. VOC can well reflect the effect of the model on the normal training target.

（2）Vaihingen Dataset

The Vaihingen dataset consists of 33 high-resolution remote sensing images with a spatial resolution of 9 cm and varying pixel sizes, with an average size of 2494 × 1064. There are five categories in the dataset: buildings, impervious water surfaces, low vegetation, trees, and cars. Due to the large size of remote sensing image and the limitation of GPU hardware, it should be processed by clipping and slicing. The experimental cutting size is 256 × 256. At the same time, in order to prevent the over-fitting phenomenon, we need to rotate, translate and other operations to enhance the data set.

B. Performance index

In this paper, we use semantic segmentation to achieve the target detection of people in VOC data. Therefore, the evaluation index commonly used in semantic segmentation can be used:

(1) Average pixel accuracy (MPA): calculate the proportion of the pixels marked correctly in each class to the total pixels, and then perform average evaluation.

(2) Mean Inter-section over Union (MIOU): MIOU is a standard metric for semantic segmentation. That is to say, the ratio of the intersection union of the true value and the predicted value of each class is calculated and finally averaged.

The evaluation formula is as follows:

$$mPA = \frac{1}{k+1}\sum_{u=0}^{k} \frac{P_{uu}}{\sum_{v=0}^{k} P_{uv}} \quad (2)$$

$$mIoU = \frac{1}{k+1}\sum_{u=0}^{k} \frac{P_{uu}}{\sum_{v=0}^{k} P_{uv}+\sum_{v=0}^{k} P_{vu}-P_{uu}} \quad (3)$$

In which K is that numb of categories contain in the training sample without background, u is the correct sample, V is the error sample, and Puu represents the number of sample that are actually positive samples and predicted to be positive samples, that is, TP (TRUE POSITIVE); Puv represents the number of samples that are actually negative but incorrectly predicted to be positive, i.e. FP (FALSE POSITIVE); Pvu represents the number of samples that are actually positive but incorrectly predicted to be negative, also FN (FALSE NEGATIVE).

C. Experimental environment and network parameter configuration

The experimental framework used in this paper is PyTorch deep learning framework, the experimental computer hardware configuration is RTX3070, the video memory is 8G, and the memory is 16G. In the process of experiment, considering the requirement of remote sensing image analysis for training time, the idea of freezing training is adopted: in the freezing stage, the backbone network is frozen, and the feature extraction network is not changed, only the network is fine-tuned. In the thawing stage, the backbone network is also added to the training, and the total number of iterations is 100.

Among them, the first 50 times are frozen for training, and the last 50 times are thawed for training, and all parameters are updated. Adam optimization algorithm is used in the experiment in this paper. Compared with SGD, it can make the model converge to the optimal performance faster. Parameter settings: LR = 0.0001, momentum = 0.9, weight _ decay = 0, batch size = 4.

D. Experimental results

Based on the semantic segmentation network evaluation indicators mentioned above, this section uses FCN, U-Net and other commonly used semantic segmentation frameworks on the VOC test set and Vaihingen data set to obtain the index values for comparison.

Table 1 showed the average result values of the model in this paper and the comparison networks constructed in the VOC data set. Compared with the original U-Net network, the algorithm proposed in this paper improves by 3.7 in MPA and 6.6 in MIOU. It can be well proved that the model has better performance than the original U-Net model. It is similar to the Deeplab v3[15] algorithm in the classification results. We also can see that the consumption time of the model proposed in this paper is faster than that of Deeplab v3. From the results, it can be concluded that ARU-Net can also get good results in daily semantic segmentation tasks.

Table 1 comparison of VOC results

Method	MPA	MIOU	TIME/ms
FCN	69.3	75.4	700
SegNet	72.8	78.7	780
U-Net	72.2	78.9	980
Deeplab v3	76.7	84.7	900
ARU-Net	75.9	85.1	870

Based on the semantic segmentation network evaluation index mentioned above, in order to reflect the model's ability to extract information features from remote sensing images, the model's generalization ability is evaluated and compared on the Vaihingen data set, and the results are shown in Table 3.

It can be seen from Table 2 that the ARU-Net model proposed in this paper is obviously superior to the previous traditional model in terms of average accuracy and average interaction ratio in remote sensing images. Compared with the SegNet network, the MPA value is increased by 3. 3, and the MIOU is increased by 6. 7, so that the corresponding target can be segmented from the remote sensing image more accurately. Compared with other segmentation models, in order to prevent the loss of low-level detail information, ARU-Net adds ASPP module and skip connection in the decoding network, and improves the residual link, which can obtain more multi-scale deep and shallow information, thus improving the generalization ability of the model and improving the performance of image segmentation.

Table 2 comparison of Vaihingen results.

Method	MPA	MIOU
FCN	66.5	54.5
SegNet	70.8	60.1
U-Net	70.1	59.5
Deeplabv3	73.9	65.2
ARU-Net	74.1	66.8

IV. Conclusions

In order to improve the segmentation accuracy of remote sensing image semantic segmentation, an improved remote sensing semantic segmentation algorithm ARUNRT based on U-Net is proposed in this paper. Through a bottom-up and top-down multi-scale structure combined with dense skip connection, the ability of U-Net network to obtain multi-scale information is improved, and the ASPP module and residual module are introduced to improve the ability of network to resolve different targets, which further improves the accuracy of remote sensing image semantic segmentation. The experimental results on VOC data set and Vaihingen data set also prove the effectiveness of the improved model for daily tasks and remote sensing tasks. In the future, more excellent backbone networks such as RESNET will be tried to further improve the performance and speed.

References

[1] Lin X G, Zhang J X. Object-based morphological building index for building extraction from high resolution remote sensing imagery[J]. Acta Geodaetica et Cartographica Sinica, 2017,46(6):724-733.]

[2] Long J W.Research on Key Techniques of Image Thresholding[D]. Jinlin: Jilin University, 2014.

[3] Jiang F, Gu Q, Hao H Z, et al. Survey on Content-based Image Segmentation Methods[J]. Journal of Software, 2017, 28(1): 160-183.

[4] Wang Y.Y. Overview and Comparison of Image Region Segmentation Algorithms[J]. Forum of Industry & Technology, 2019, 18(13): 54-55.

[5] Xu Y, Feng M R, Pi J T, et al. Remote Sensing Image Segmentation Method Based on Deep Learning Model[J]. Computer Applications, 2019, 39(10): 2905-2914.

[6] Shelhamer E,Long J,Darrel T. Fully convolutional networks for semantic segmentation[J].IEEE Transactions on Pattern Analysis and Machine Intelligence,2017,39(4):640-651.

[7] Badrinarayanan V, Kendall A, Cipolla R. SegNet: a deep convolutional encoder-decoder architecture for image segmentation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(12): 2481-249

[8] Ronneberger O, Fischer P, Brox T, et al. U-Net: Convolutional networks for biomedical image segmentation[C].International Conference on Medical image computing and computer- assisted intervention. Springer, Cham,2015:234-241

[9] Guo M Q, Liu H, Xu Y Y, et al. Building extraction based on U-Net with an attention block and multiple losses[J].Remote Sensing, 2020,12(9):1400..

[10] Ibtehaz N, Rahman M S. MultiResUNet: Rethinking the U-Net architecture for multimodal biomedical image segmentation[J]. Neural Networks, 2020,121:74-87.

[11] Xu J W, Liu W,Shan H Y, et al. High-resolution remote sensing image building extraction based on PRCUnet[J]. Journal of Geo-information Science,2021,23(10):1838-1849.

[12] SONG Tingqiang, LIU Tongxin, ZONG Da. Research on Road Extraction Method from Remote Sensing Images Based on Improved U-Net Network [J]. Computer engineering and Application,2021,57(14):209-216.

[13] Chen L C, Papandreou G, Schroff F, et al. Rethinking atrousconvolution for semantic image segmentation[EB/OL].2017 arXiv:1706.05587.https://arxiv.org/abs/1706.05587

[14] He K, Zhang X, Ren S, et al. Spatial pyramid pooling in deep convolutional networks for visual recognition[J].IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015,37(9):1904-1916.

[15] Chen L C, Papandreou G, Schroff F, et al. Rethinking atrous convolution for semantic image segmentation[EB/OL]. (2017-06-17) [2021-02-10].