# Multi-scale Spatial Aggregation Network for Remote Sensing Image Segmentation

Xinkai Sun, Jing Gu, Jie Feng, Shuyuan Yang, Licheng Jiao

# Multi-scale Spatial Aggregation Network for Remote Sensing Image Segmentation

Xinkai Sun[1], Jing Gu[1], Jie Feng[1], Shuyuan Yang[1], Licheng Jiao[1]

[1] School of Artificial Intelligence, Xidian University, Xian, China
xuer6126@126.com

**Abstract.** Semantic segmentation of remote sensing images is of great significance to the interpretation of remote sensing images. Recently, convolutional neural networks have been increasingly used in this task since it can effectively learn the features in the image. In this paper, an end-to-end semantic segmentation framework, Multi-scale Spatial Aggregation Network (MSAN), is proposed for the remote sensing image segmentation. At first, a classical SegNet is employed as the backbone of the network because its simple structure is suitable for the remote sensing images that have a small quantity of samples. Then several skip connections and a densely connected block are utilized to enhance the usage of the low-level feature and reduce the loss of the detail information in the original image. Moreover, multi-scale spatial information fusion module and a spatial path are added between the encoder and decoder of SegNet, which can effectively extract the features of objects with different sizes in the remote sensing images. Finally, a smoothing algorithm is presented to improve the blocking effect of the remote sensing image segmentation results. The proposed MSAN is tested on the ISPRS Vaihingen dataset and the dataset of a city in southern China, which obtains the satisfactory results.

**Keywords:** image segmentation, remote sensing, feature aggregation.

## 1 Introduction

Remote sensing image is captured by the imaging equipment carried on the aircraft or space shuttle, which has the advantage of all-time imaging. Semantic segmentation [1,2] is an important basis in the field of remote sensing image processing [3]. Its purpose is to distinguish different types of land covers in remote sensing image, and then provide useful information for many fields, such as geography [4], surveying [5] and mapping, military [6], and so on. However, the remote sensing images have the characteristics of the large scale scene, many objects with the small size, the blurred boundaries between different categories, and being easily interfered by factors such as seasons and shadows, which bring certain difficulties to the image segmentation. In recent years, more and more scholars are beginning to utilize the deep learning to achieve the image semantic segmentation [7,8], especially the Convolutional Neural Network (CNN), because it can learn the high-level feature representation and train the network model from a large number of data by an end-to-end way. At present, the mainstream semantic segmentation methods [9] can be divided into two categories: the region based semantic segmentation method [10,11] and the full convolution network semantic segmentation method. The region based semantic segmentation method first divided the original image into numerous free-form regions, and then classified these regions by the extracted features of regions. Typical methods include EPLS [12], SPMK [13]. However, the extracted feature in this kind of method did not contain global spatial information, so the region based semantic segmentation method cannot accurately generate the boundary, which will affect the final segmentation effect greatly. On the other hand, the full convolution network based semantic segmentation methods [14,15] replace the full connection layer with the convolution layer to realize the prediction of the input image with any size. Since FCN can perform end-to-end training on the entire input image and accurately restore the image through upsampling, the FCN based semantic segmentation methods have attracted the attention of many scholars. In this paper, the FCN is adopted as the basic frame work for the semantic segmentation.

In this paper, a novel multi-scale spatial aggregation network (MSAN) is proposed for remote sensing image segmentation. Considering that there is a small quantity of the labeled samples for the remote sensing images, MSAN employs a shallow network SegNet [16] as the backbone network, which includes an encoder and a decoder. A dense connection block [17] is added to the encoder; meanwhile several skip connections [18], multi-scale spatial information fusion module and a spatial path [19] are utilized to connect the features with the same scales in the encoding and decoding modules to enhance the perfor-

mance of MSAN. Specifically, the contributions of this paper are described in detail as following.

1. We propose a Multi-scale Spatial Aggregation Network (MSAN) for the remote sensing image segmentation, which adopts skip connection structure and dense connection block to take full advantage of the low-level feature, thus reducing the loss of effective information in the downsampling process without increasing extra computational cost. Meanwhile, multi-scale spatial information fusion module and spatial path are utilized to aggregate the features with different receptive fields that can comprehensively describe the targets with different scales in the remote sensing image, thus improving the recognition performance of different land covers.

2. In order to avoid losing the semantic information of boundary, a smoothing algorithm is proposed in this paper, which alleviates the block effect of segmentation results effectively.

## 2 Related Work

The earlier semantic segmentation networks are often proposed and applied in the natural scene images, since a large number of the natural scene labeled images can more easily be obtained than the remote sensing. On this basis, some semantic segmentation methods of the remote sensing image are proposed. The two kinds of approaches will be explained in detail below.

### 2.1 Semantic Segmentation of Natural Scene Image

At present, many semantic segmentation methods have been presented to segment the natural scene image. First of all, Long et al. [20] replaced the fully connected layer in the VGG16 [21] network with a convolutional layer, and successfully applied a convolutional neural network to the natural image semantic segmentation. Subsequently, a SegNet was proposed, which used a pooling operation with locations to avoid the loss of spatial information in the down-sampling process. Then, an innovative encoder-decoder network structure was designed and named as U-Net, which could effectively restore the size of the original feature map and has been used widely. Furthermore, a multi-scale information fusion method, atrous spatial pyramid pooling (ASPP)[22], was proposed by Chen et al in DeepLab v3+[23]., which used an atrous convolution in the spatial dimension. The ASPP module can take parallel sampling with different sampling rates for a given input by the atrous convolution [24]. Similarly, Zhao et al. introduced a pyramid pooling module (PPM) in PSPNet [25], which could obtain a set of feature maps with different receiving field sizes by performing the pooling operations on the feature maps with different proportions. In addition, a brand-new Efficient Spatial Pyramid (ESP) module [26] was proposed to replace the original convolution module, which has obvious advantages in speed, reaching 112 frames per second. Zhang et al. [27] focused on the network's comprehensive understanding of contextual information, and proposed a context encoding module, which greatly improved the effect of semantic segmentation at the expense of increasing a small amount of calculation. Since ASPP and PPM can effectively extract the features of image in various scales, they are introduced into a simple framework SegNet to reserve the low-resolution features of the deep network in this paper, thus increasing the accuracy of the segmentation results.

### 2.2 Semantic Segmentation of Remote Sensing Image

In recent years, more and more semantic segmentation methods [28,29] have been applied in the remote sensing field [30], such as DST_2[31] , ONE_7, CVEO, CAS_Y1, CASIA. A multi-core convolutional layer was introduced to extract the image features by aggregating the convolutions of different scales. Chen et al. [32] were inspired by a residual module in ResNet and proposed a shortcut block to replace the conventional convolution operation, which ensures the rapid and direct transmission of the gradient information. At the same time, considering that it is difficult for the network to train the entire remote sensing image, an image cropping strategy [33] was proposed. Yu et al. [34] combined the PPM module in PSPNet with ResNet and proposed a brand new network. Liu et al. [35] replaced the original concatenation operation on all context information with fusing step by step in pairs and continuously corrected the image feature details to realize the image feature fusion. In most of the existing semantic segmentation methods, did not take advantage of the low-level features the detail and spatial information of the image were easily lost with the deepening

network. The proposed MSAN takes full advantage of the low-level features and spatial information to optimize the network.

## 3 Proposed Method

In this paper, a novel MSAN is proposed for the semantic segmentation of the remote sensing image. Since the remote sensing image has large scale and constantly changing land covers with the change of the season, the labeled remote sensing image samples are difficult to be obtained. The deeper network will cause over-fitting in the case of a small amount of samples. Therefore, as a simple and effective shallow semantic segmentation network, SegNet is selected as the backbone network of MSAN. Meanwhile, the location-based pooling method in SegNet can accurately restore the position information lost during the downsampling process, which is crucial for the remote sensing image with the blurred boundaries and complex gradients. Moreover, the skip connection and densely connected block are employed to decrease the lost detail information. In order to make full use of the effective features in different levels and enhance the recognition ability of different scale targets, multi-scale information fusion module and spatial path are added to the proposed model. The overall framework of MSAN and the structure of every module will be elaborated as following.

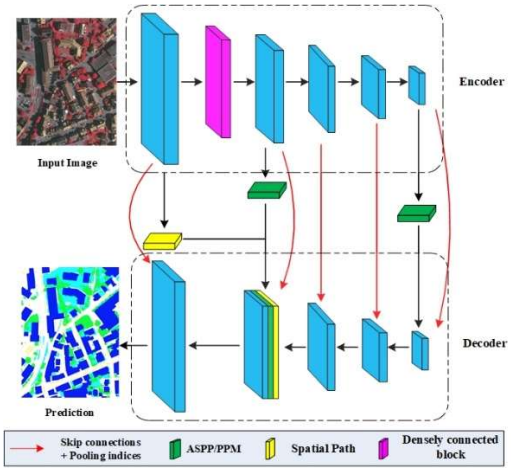



**Fig. 1.** The overall architecture of the proposed MSAN.

### 3.1 The Architecture of MSAN

The overall architecture of MSAN is shown in Figure 1, which includes an encoder and a decoder. In the encoding part, the down-sampling network is composed of an encoding sub-network and a spatial path branch network to extract the high-level and low-level information of the image respectively. The encoding sub-network is composed of five convolutional layers and a densely connected block, where each convolutional layer has a corresponding pooling operation and a skip connection structure. The densely connection block is added between the first layer and the second layer of the encoding sub-network to enhance the fusion among the features. Furthermore, two spatial information fusion modules are added between the encoder and decoder to fuse middle-level and high-level features. Additionally, the output of the spatial path branch is cascaded with the fourth upsampling feature.

The decoder of MSAN mainly consists of five up-sampling layers corresponding to the encoded convolutional layers, where each layer is cascaded with the output of the corresponding skip connection, and the channel is compressed through convolution. The prediction of the image is completed by restoring the size of the original input image layer by layer.

### 3.2 Densely Connected Structure

Inspired by ResNet and Inception network, the dense connection structure was proposed in DenseNet by Huang et al., which is different from the previous network towards a deeper and wider direction, but starts from the characteristics. The features of all layers are connected to ensure the maximum transmission of effective information between the layers. In short, the input of each layer comes from the output of all previous layers, as shown in Figure 2. The densely connected block consists of four layers, and each layer includes a

convolution (Conv), batch normalization (BN), and nonlinear transformation based on ReLU function (Conv_BN_ReLU). This connection method achieves better effect and fewer parameters, which can also avoid the gradient disappearance and strengthen the transfer between the features. The above advantages make us add it to the proposed method to improve the performance of the network.

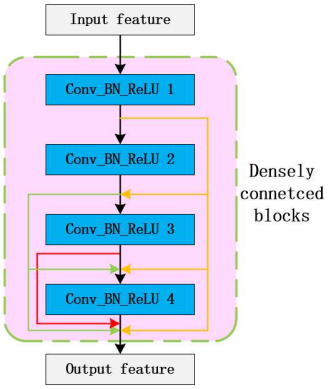



**Fig. 2.** Schematic diagram of densely connected blocks structure.

### 3.3 Multi-scale Information Fusion Module

The researches show that the performance of the semantic segmentation network can be greatly improved by combining the deep and shallow features of images, because it can get more global information and increase the ability of network to grasp the overall image at macro level. Along this idea, many scholars have made many meaningful attempts, where ASPP and PPM of PSPNet have been verified to have favorable performance. ASPP is composed of four atrous convolutions of different scales and a global average pool operation to obtain the feature maps with different receptive fields, and then they are integrated to obtain the fusion results concluding different scale spatial information. In PPM of PSPNet, four pooling operations with different sizes are used to explore the context information of different sizes. Then, the channel is compressed by convolution, and the original feature graph size is recovered by bilinear interpolation. Finally, the obtained features are cascaded to the input features, thus achieving the spatial information fusion.

### 3.4 Spatial Path

Practice has proved that the rich spatial information is essential to improve the accuracy of semantic segmentation networks, which is more prominent in the field of remote sensing image. At present, some mainstream semantic segmentation frameworks often encoded the input images through the deeper networks to obtain the feature maps with the higher resolution, or used the pyramid pooling modules to perform information fusion on feature maps of different sizes. However, a deeper network will increase the computational cost. At the same time, experimental data shows that the excessive compression of feature maps will bring about the loss of spatial information. Therefore, a too deep network is not suitable for remote sensing images with rich semantic information. Yu et al. creatively proposed a spatial path module, which is composed of three convolutional layers with a step size of 2. Each convolutional layer is followed by batch normalization [36] and ReLU [37], so the output of this module is 1/8 of the original input image size. This module can retain rich spatial features and has small computation complexity, which inspired us to apply it to the proposed method. Figure 3 shows the overall structure of the spatial path.

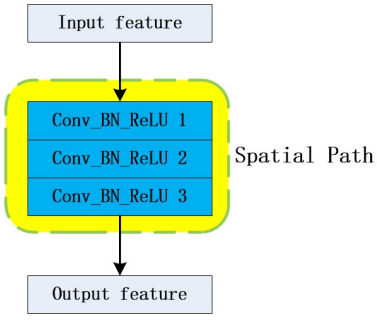



**Fig. 3.** The overall structure of the spatial path.

### 3.5 Smoothing Algorithm

It is well known that the size of remote sensing images is large. Therefore, it is difficult for the network to train the entire image. The original image needs to be cropped into small image blocks, and then these cropped image blocks are fed to the network. Finally, the network's prediction results are stitched into a complete image. The traditional cropping method generally crops the original image into some non-overlapping image blocks according to a prescribed step. In this process, the clipping operation will cause the loss of the semantic information of the boundary of the small image blocks, which makes the spliced image produce block effect and causes a sense of visual incoherence. In order to solve this problem, a smoothing algorithm is proposed in this paper. Specifically, the cropping interval is set to 1/2 of the required size so that the boundary in the large remote sensing images is repeated in different cropped image blocks to avoid losing the semantic information of the boundary. During the splicing process, the majority voting method is adopted for the overlapping part, and the voting result is used as the final prediction result. The presented smoothing method alleviates the block effect of the segmentation results effectively.

## 4 Experimental Results and Analysis

The performance of the proposed MSAN is verified by a series of the experiments. At first, the test dataset and the experiment setting are described respectively. Moreover, the experimental results of MSAN and other typical methods are shown and analyzed, where two common metrics overall accuracy (OA) and mean intersection over union (mIoU) are used to evaluate the performance of these semantic segmentation methods. Finally, several ablation experiments are carried out to demonstrate the effectiveness of the proposed network.

### 4.1 Dataset

The proposed MSAN is test on the ISPRS Vaihingen dataset [38] and the remote sensing image dataset of a city in southern China. The Vaihingen dataset shows a small town in Germany, which is composed of 33 pieces of different sizes cropped from a remote sensing image, where each image has its corresponding ground truth and digital surface model (DSM). The data set is manually divided into six different categories, including the impervious surfaces, building, low vegetation, tree, car, and clutter/background. There are five remote sensing images in the remote sensing image dataset of a city in southern China, showing scenes of rural and urban areas. All ground truths are manually marked, including the vegetation, buildings, water bodies, roads, and background.

### 4.2 Experimental Setting

On the ISPRS Vaihingen dataset, all the results of this dataset are fairly compared by using the conventional ground truth. Among them, 16 pictures are used as the training set and the rest are used as the test set, and the input size of the network is $256\times256$. At the same time, in order to overcome the overfitting of the data, all input images are randomly flipped. Multiple pooling sizes in PPM are set to $1\times1$, $2\times2$, $4\times4$, $8\times8$ pooling respectively. The sampling rates of the atrous convolution in ASPP are set to 6, 12, and 18 respectively.

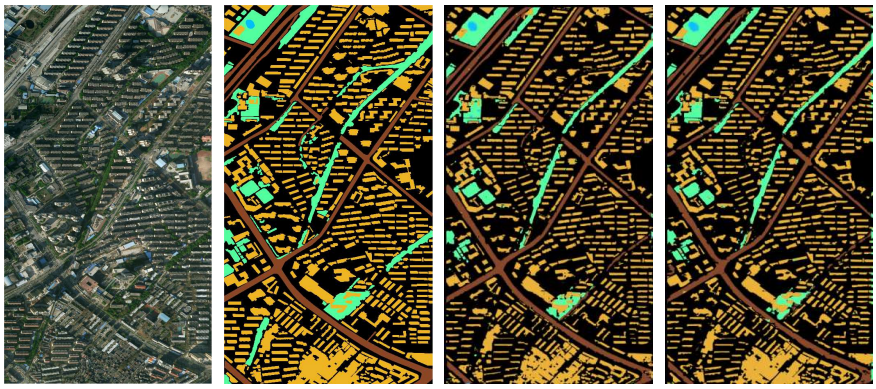
The proposed method was compared with the mainstream semantic segmentation networks SegNet, U-Net, PSPNet and DeepLab v3+, and the segmentation results will be evaluated in terms of the running speed and segmentation accuracy. The experiment platform is Intel Core i7 9700K, 64GB RAM and Nvidia GeForce GTX2080 Ti (11264MB memory). The deep learning framework is Python 3.6, Tensorflow 1.4.0 and Keras 2.1.3.

### 4.3 Experimental Results

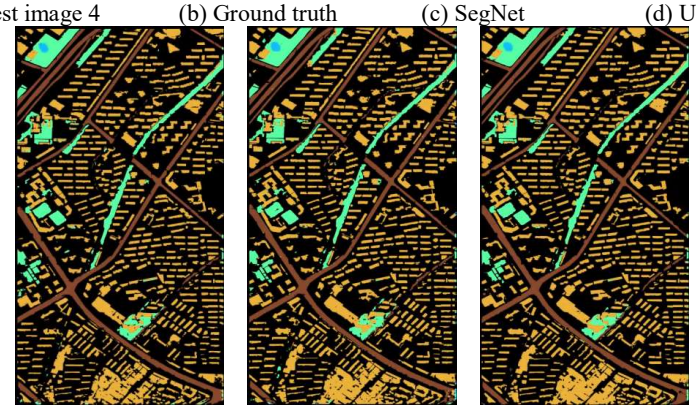
In order to verify the performance of the proposed MSAN, the prediction results of different remote sensing image semantic segmentation methods on a test image (ID 27) of the Vaihingen dataset are shown in Table 1. As shown in the Table 1, the accuracy of DST_2 has reached 86.1, but the amount of parameters is as high as 184M. The parameter quantity of ONE_7 is the most competitive, and the accuracy is relatively higher. By contrast, the performance of CVEO and CAS_Y1 is relatively poor. MSAN is similar with CASIA in the segmentation accuracy and better than other methods, but the amount of the parameters in MSAN is much smaller than CASIA, which shows that the propose method balances the segmentation accuracy and the running speed greatly.

**Table 1.** Overall accuracy and the number of parameters of different methods on the Vaihingen dataset (ID 27).

| Method | DST_2 | ONE_7 | CVEO | CAS_Y1 | CASIA | MSAN |
|---|---|---|---|---|---|---|
| OA | 86.1 | 86.9 | 85.5 | 84.7 | 87.5 | 87.3 |
| Parameters | 184M | 24M | 52M | 85M | 151M | 37M |



(a) test image 4 (b) Ground truth (c) SegNet (d) UNet



(e) PSPNet (f) DeepLab v3+ (g) MSAN

**Fig. 4.** Semantic segmentation results of different methods on test image 4 of a city in southern China.

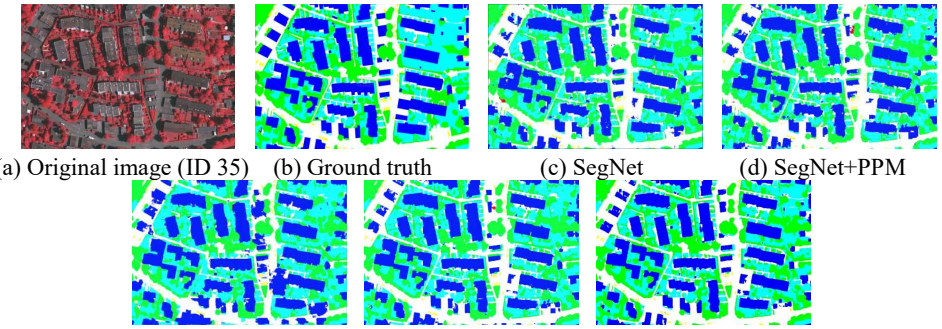
Figure 4 shows the comparison between the proposed method and mainstream semantic segmentation methods. Figure 4(a) and (b) shows an optical remote sensing image (ID 4) of a city in southern China and its corresponding ground truth. The results of SegNet and U-Net are shown in Figure 4(c) and (d). As we can see, in Figure 4(c) and (d), some dense buildings are not recognized due to without the information fusion module. Compared with Figure 4(e) and (f) respectively corresponding to the results of PSPNet and DeepLab v3+, MSAN has better regional consistency, as shown in Figure 4(g). Table 2 shows the OA and mIoU of the mentioned methods on 4 test images. Since the test images 3 and 4 contain many dense targets, the OA obtained by various methods is lower than the first two images. In general, the segmentation accuracy of MSAN is highest in Table 2.

**Table 2.** OA and mIoU of the results of different methods on test images of a city in southern China.

| Method | test image 1 | | test image 2 | | test image 3 | | test image 4 | |
|---|---|---|---|---|---|---|---|---|
| | OA(%) | mIoU(%) | OA | mIoU | OA | mIoU | OA | mIoU |
| SegNet | 85.3 | 72.5 | 85.8 | 72.5 | 73.3 | 49.9 | 74.3 | 50.0 |
| U-Net | 88.1 | 73.4 | 86.8 | 73.4 | 71.5 | 50.5 | 76.2 | 52.1 |
| PSPNet | 89.6 | 73.8 | 88.8 | 76.2 | 77.4 | 52.5 | 78.8 | 55.5 |
| DeepLab v3+ | 90.1 | 74.0 | 87.4 | 77.0 | 77.6 | 53.8 | 78.2 | 55.3 |
| MSAN | 93.1 | 79.2 | 92.6 | 80.2 | 80.8 | 58.3 | 79.8 | 56.9 |

### 4.4 Ablation Experiments

In order to prove the effectiveness of the added modules, the segmentation results of adding different modules is shown in Figure 5. Figure 5(a) and (b) shows a test image (ID 35) of the Vaihingen dataset and its corresponding ground truth, and Figure 5(c) shows the prediction results of the basic backbone network SegNet. Figure 5(d) and (e) shows the prediction results of adding ASPP and PPM to SegNet respectively. Figure 5(f) and (g) shows the final prediction results under different spatial fusion modules. It can be seen that the spatial regional consistency has been significantly improved in the final prediction results.



(a) Original image (ID 35) (b) Ground truth (c) SegNet (d) SegNet+PPM

(e) SegNet+ASPP (f) MSAN(PPM) (g) MSAN(ASPP)

**Fig. 5.** Semantic segmentation results of different network structures.

**Table 3.** OA and mIoU of the results of different network structures.

| Method | OA | mIoU |
|---|---|---|
| SegNet | 75.4 | 49.7 |
| SegNet+PPM | 77.1 | 50.2 |
| SegNet+ASPP | 77.9 | 50.4 |
| MSAN(PPM) | 83.2 | 56.9 |
| MSAN(ASPP) | 83.9 | 57.8 |

Table 3 shows the OA and mIoU of different methods. It can be seen from Table 3 that the recognition accuracy of targets with different scales is enhanced gradually with the increase of the modules, which shows that the adding modules are effective.

## 5 Conclusions

In this paper, a novel Multi-scale Spatial Aggregation Network (MSAN) is proposed for the semantic segmentation of the remote sensing images. First of all, MSAN adopts SegNet as the backbone network. Then, the skip connections and dense connection block are aggregated into the backbone network to reduce the loss of the detail information in the downsampling process. Moreover, in response to the large gap between different target sizes in the remote sensing images, multi-scale information fusion modules and spatial path are employed. Finally, a smoothing algorithm is given to improve blocking effect and increase the accuracy of the final image segmentation result.

## References


1. Saurabh Gupta, Pablo Arbelaez, and Jitendra Malik. Perceptual organization and recognition of indoor scenes from rgb-d images. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 564–571, 2013.
2. Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In European Conference on Computer Vision, pages 746–760. Springer, 2012. 1, 2
3. Yu C, Wang J, Gao C, Yu G, Shen C, Sang N. Context prior for scene segmentation. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition ,2020
4. Lin G, Milan A, Shen C, Reid I. Refinenet: Multi-path refinement networks with identity mappings for high-resolution semantic segmentation. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition, 2017
5. Fu J, Liu J, Tian H, Fang Z, Lu H. Dual attention network for scene segmentation. In Proc. IEEE Conference on Computer Vision and Pattern Recognition,2019
6. Julio Silva-Rodríguez a, Adrián Colomer b, B V N. WeGleNet: A Weakly-Supervised Convolutional Neural Network for the Semantic Segmentation of Gleason Grades in Prostate Histology Images[J]. Computerized Medical Imaging and Graphics, 2021.
7. Zhang Z , Huang J , Jiang T , et al. Semantic segmentation of very high-resolution remote sensing image based on multiple band combinations and patchwise scene analysis. Journal of Applied Remote Sensing, 2020, 14(1):1.
8. Xiaofeng Ren, Liefeng Bo, and Dieter Fox. Rgb-(d) scene labeling: Features and algorithms. In 2012 IEEE Conference on Computer Vision and Pattern Recognition, pages 2759–2766. IEEE, 2012. 1
9. Yu C, Wang J, Peng C, Gao C, Yu G, Sang N. Learning a discriminative feature network for semantic segmentation. In. Proc. IEEE Conference on Computer Vision and Pattern Recognition,2018
10. Badrinarayanan V, Kendall A, Cipolla R (2017) SegNet: A deep convolutional encoder-decoder architecture for image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence 39(12):2481–2495

11. Paszke A, Chaurasia A, Kim S, Culurciello E (2016) Enet: A deep neural network architecture for real-time semantic segmentation. arXiv
12. Liu J , Geng Y , Zhao J , et al. Image Semantic Segmentation Use Multiple-Threshold Probabilistic R-CNN with Feature Fusion[J]. Symmetry, 2021, 13(2):207.
13. Bergum S , Saad A , Stahl A . Automatic in-situ instance and semantic segmentation of planktonic organisms using Mask R-CNN[C]// IEEE Oceanic Engineering Society & Marine Technology Society. IEEE, 2020.
14. Long, Jonathan, Shelhamer, Evan, Darrell, Trevor, "Fully Convolutional Networks for Semantic Segmentation," IEEE Trans. Pattern Anal. Mach. Int., vol. 39, no. 4, pp. 640-651, 2014.
15. Chen G , Zhang X , Wang Q, "Symmetrical Dense-Shortcut Deep Fully Convolutional Networks for Semantic Segmentation of Very-High-Resolution Remote Sensing Images," IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens., vol. 11, no. 5, pp. 1633-1644, 2018.
16. V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," IEEE Trans. Pattern Anal. Mach. Int., vol. 39, no.12, pp. 2481–2495, 2017.
17. Huang G , Liu Z , Laurens V D M , et al. Densely Connected Convolutional Networks.In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,2017
18. O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in Proc. Int. Conf. Med. Image Comput., pp. 234–241,2015.
19. Yu C , Wang J , Peng C , et al. BiSeNet: Bilateral Segmentation Network for Real-time Semantic Segmentation. In European Conference on Computer Vision. Springer, Cham, 2018.
20. Long, Jonathan, Shelhamer, Evan, Darrell, Trevor, "Fully Convolutional Networks for Semantic Segmentation," IEEE Trans. Pattern Anal. Mach. Int., vol. 39, no. 4, pp. 640-651, 2014.
21. Simonyan, Karen, and A. Zisserman. "Very Deep Convolutional Networks for Large-Scale Image Recognition." In Proc. Int. Conf. Learn. Represent., 2014.
22. Chen LC, Papandreou G, Kokkinos I, 2018. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. IEEE Trans. Pattern Anal. Mach. Int. vol. 40, no. 4, pp. 834-848.
23. Chen, Liang Chieh, 2017. Rethinking Atrous Convolution for Semantic Image Segmentation. in Proc. Int. Conf. Comput. Vis. Pattern Recognit.
24. Chen, Liang-Chieh, et al., 2018. Encoder-decoder with atrous separable convolution for semantic image segmentation. in Proc. Eur. Conf. Comput. Vis.
25. H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., pp. 6230–6239, 2017.
26. Mehta S, Rastegari M, Caspi A, Shapiro L, Hajishirzi H (2018) Espnet: Efficient spatial pyramid of dilated convolutions for semantic segmentation. In Proc. European Conference on Computer Vision, pp 552–568.
27. Zhang H, Dana K, Shi J, Zhang Z, Wang X, Tyagi A, Agrawal A (2018a) Context encoding for semantic segmentation. In Proc. IEEE Conference on Computer Vision and Pattern Recognition, pp 7151–7160.
28. Li X, Liu Z, Luo P, Loy CC, Tang X (2017) Not all pixels are equal: difficulty-aware semantic segmentation via deep layer cascade. In Proc. IEEE Conference on Computer Vision and Pattern Recognition.
29. Mehta S, Rastegari M, Shapiro LG, Hajishirzi H (2019) Espnetv2: A light-weight, power efficient, and general purpose convolutional neural network. In Proc. IEEE Conference on Computer Vision and Pattern Recognition.
30. Audebert N , Saux B L , Sébastien Lefèvre., "Semantic Segmentation of Earth Observation Data Using Multimodal and Multi-scale Deep Networks," in Proc. Lect. Notes Comput. Sci., pp.180-196, 2017.
31. Y. He , X. Dong , G. Kang , et al., "Asymptotic Soft Filter Pruning for Deep Convolutional Neural Networks," IEEE Transactions on Cybernetics, vol. 50, no. 8, pp. 3594-3604, 2020.
32. Chen G , Zhang X , Wang Q, 2018. Symmetrical Dense-Shortcut Deep Fully Convolutional Networks for Semantic Segmentation of Very-High-Resolution Remote Sensing Images. IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens., vol. 11, no. 5, pp. 1633-1644.
33. He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, 2016. Deep Residual Learning for Image Recognition. in Proc. Int. Conf. Comput. Vis. Pattern Recognit.
34. Bo Y , Lu Y , Fang C., "Semantic Segmentation for High Spatial Resolution Remote Sensing Images Based on Convolution Neural Network and Pyramid Pooling Module," IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens., pp. 1-10,2018.
35. Liu Y , Fan B , Wang L, "Semantic labeling in very high resolution images via a self-cascaded convolutional neural network," ISPRS Journal of Photogrammetry and Remote Sensing, vol. 145, pp. 78-95, 2018.
36. Ioffe S, Szegedy C, "Batch normalization: accelerating deep network training by reducing internal covariate shift," in Proc. Int. Conf. Mach. Learn., ICML., pp. 448-456, 2015.
37. V. Nair and G. E. Hinton, "Rectifified Linear Units Improve Restricted Boltzmann Machines," in Proc. 27th Int. Conf. Mach. Learn., 2010.
38. M. Everingham, S. A. Eslami, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes challenge: A retrospective," in Proc. Int. J. Comput. Vis., vol. 111, no. 1, 98–136, 2015.