Road Extraction Based On Improved Deeplabv3 Plus In Remote Sensing Image

Hanxiang Wang[1], Fan Yu [1]*, Junwei Xie[1], Haonan Wang[1], Haotian Zheng[1]

[1] Beijing University of Civil Engineering and Architecture, Beijing 102616, China, - wanghanxianghx@gmail.com(H.W.); yufan@bucea.edu.cn(F.Y.); jw_x@foxmail.com(J.X.); hn_w1997@163.com(H.W.); fuerhaotian@126.com(H.Z.)

Commission III and IV

KEY WORDS: Remote Sensing, Road Extraction, High-Resolution Remote Sensing Image, Deep Learning, Deeplabv3 Plus, ASPP.

ABSTRACT:

Urban roads in remote sensing images will be disturbed by surrounding ground features such as building shadows and tree shadows, and the extraction results are prone to problems such as incomplete road structure, poor topological connectivity, and poor accuracy. For mountain roads, there will also be problems such as hill shadow or vegetation occlusion. We propose an improved Deeplabv3+ semantic segmentation network method. This method uses ResNeSt, which introduces channel attention, as the backbone network, and combines the ASPP module to obtain multi-scale information, thereby improving the accuracy of road extraction. Analysis of the experimental results on the Deeplglobe dataset shows that the intersection ratio and accuracy of the method in this paper are 63.15% and 73.16%, respectively, which are better than other methods.

1. INTRODUCTION

As one of the classic land resources, roads will expand in all directions after people's planning. The direction of the road will have a certain impact on the distribution of surrounding features, so the complete extraction of the road structure is not only beneficial to the travel of residents, but also has a positive effect on traffic navigation, and is also important for urban planning and building. Layout also has far-reaching effects. Therefore, it has far-reaching research significance to accurately extract roads from remote sensing images. Commonly used extraction methods can be divided into traditional methods and deep learning methods.

Traditional methods are divided into pixel feature-based and object feature-based. The methods based on pixel features can be divided into spectral analysis method, edge detection method and threshold segmentation method.(Liu et al., 2017) proposed a method based on spectral features and geometric features inference combined with road material diversity. Based on object features, it can be divided into area method, knowledge model method and texture analysis method. (Alshehhi and Marpu, 2017) propose an unsupervised road extraction method based on hierarchical image segmentation.

In recent years, the deep learning method has attracted more and more attention due to the advantages of high accuracy and wide model ubiquity. The main application of road extraction is semantic segmentation technology, which is to separate the road and background segmentation in high-resolution remote sensing images. By inputting a large amount of data into the semantic segmentation neural network, the neural network can learn a certain feature from the large amount of input data, and finally the learned feature can be applied to successfully distinguish the road and the background in the photo. Semantic segmentation methods for extracting roads can be divided into three categories: methods based on FCN (Long et al., 2015), methods based on U-Net(Ronneberger et al., 2015), methods based on attention machanism, methods based on Deeplab(Chen et al., 2018, 2017, 2016).

Based on the FCN method, (Wang et al., 2018) proposed the s-FCN-loc model, which can use both RGB images and semantic contours to extract roads better and more accurately. This paper tests on the KIIT dataset and finds that s-FCN-loc is 30% faster than the FCN method, and s-FCN-loc has higher accuracy than FCN. (Zhong et al., 2016) tested the effect of hyperparameters such as learning epoch and learning rate on FCN prediction accuracy on the Massachusetts road dataset.

Based on the U-Net method, (Xie et al., 2019) proposed HsgNet, which adopts a U-shaped encoder-decoder architecture, uses ResNet as the encoder of HsgNet, and the decoder adopts the decoder of U-Net, adding a middle block to extract the context The features are tested on two public datasets, and both achieve good results. (Li et al., 2019) proposed that the Y-Net neural network combined with FCN and UNet can extract multi-scale features from high-resolution images, tested on two datasets, and has better accuracy than UNet. (Hou et al., 2021) propose a C-Unet network that uses Unet and Unet with atrous convolution to extract roads and fuse the results. Good results have been achieved on the Massachusetts road dataset.

Because the attention mechanism has a good effect on extracting features (Knudsen, 2007), more and more scholars apply it to deep learning. (Hu et al., 2019) Proposed Squeeze-and-Excitation block and combined it with ResNet to achieve the most advanced level in Scene Classification and Object Detection. (Woo et al., 2018) proposed CBAM module, which calculates the attention weights along the channel and space dimensions in turn, and then multiplies the original feature map to dynamically adjust the feature map, achieving SOTA on CIFAR-100 and ImageNet-1K. (Park et al., 2018) proposed a BAM module, calculates channel attention and spatial attention separately, fuses the two, and multiplies the fusion result with the original feature map to calibrate the original feature map. (Zhang and Yang, 2021) proposed SA-Net, first of all, group the input feature maps, each group is divided into two parts, calculate spatial attention and channel attention respectively, and concatenate the readjusted feature maps together. Finally, the subfeature maps of all groups

* Corresponding author

are aggregated together to achieve the state-of-the-art in image classification, object detection, and instance segmentation.

Based on the Deeplab method, (Chen et al., 2016) proposed the Deeplab v2 network. By adding the ASPP module to the network structure, the multi-scale extraction of image features can be achieved, which can effectively improve the accuracy of image segmentation. (Chen et al., 2018) proposed the Deeplab v3+ network, by fusing the multi-scale features extracted from ASPP with the low-level feature map, the neural network can obtain richer feature information, thereby achieving more accurate feature extraction. (Xia et al., 2018) used ResNet with atrous convolution as the backbone network of DeepLab, and achieved good results on GF-2 images, but the smoothness of the prediction curve needs to be further improved.

At present, because the Deeplab series network has achieved very good results in semantic segmentation, it has received more and more attention in the field of remote sensing. However, Deeplabv3+ does not consider the relationship between channels, resulting in poor road extraction. In view of the above shortcomings, this paper improves the Deeplab v3+ network, uses ResNeSt that can extract the relationship between channels as the backbone network, and makes full use of cross-channel information to improve the accuracy of road extraction.

2. METHOD

2.1 Deeplabv3+

The Deeplabv3+ semantic segmentation model was proposed by Chen (Chen et al., 2018). Compared with other semantic segmentation networks, it has a better extraction effect, so it is widely used in the field of remote sensing. Deeplabv3+ adds the ASPP (Atrous Spatial Pyramid Pooling) module to the encoder. Using atrous convolution can extract multi-scale feature information without reducing the resolution, which is very helpful for improving accuracy. Compared with Deeplabv3, Deeplabv3+ proposes a decoder module, which combines low-level features and high-level features to achieve better segmentation results.

The Deeplabv3+ neural network consists of an encoder and a decoder, and the encoder consists of a backbone and an ASPP module. The image is input to the backbone for feature extraction. The output of the backbone is divided into two parts. One part of the low-level semantic features is directly input to the decoder, and the other part of the high-level semantic features is output from the backbone . After multi-scale feature extraction by the ASPP module, double Double upsample to the same size as the first part, and then concatenate the output of the two parts. Then use the convolution with the kernel size of 3×3 to perform feature fusion, and then double upsampling the feature map to the original image size to get the final segmentation result. The structure of Deeplabv3+ is shown in Figure 1.

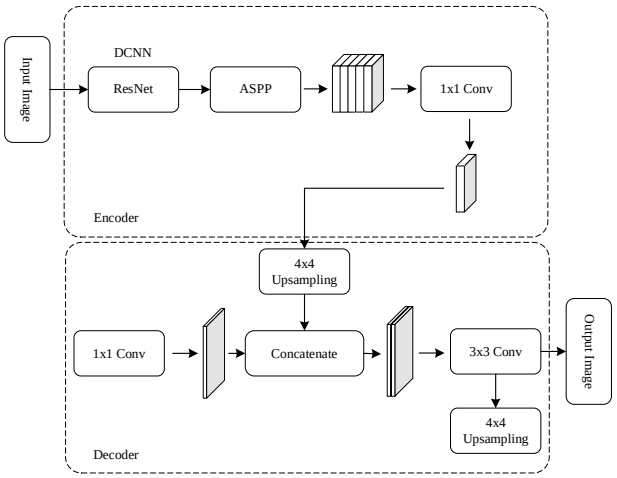

Figure 1. Deeplabv3+ architecture

2.2 Channel Attention

The attention mechanism can make the neural network pay attention to important features while ignoring unimportant noise(Hu et al., 2019), and let the neural network selectively pay attention to features while ignoring unimportant features, which can greatly improve the effect of semantic segmentation model. Channel attention can choose to focus on expressing some features and ignore some features by adjusting the weights of different channels by calculating the weight between each feature map.

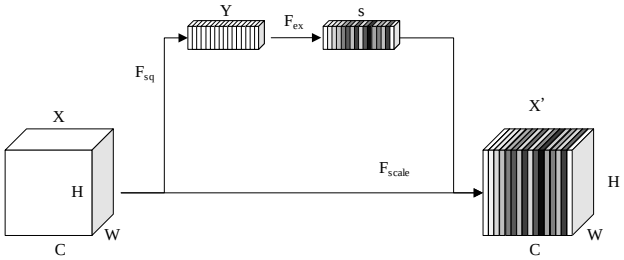

Figure 2. Channel Attention

The input feature map goes through three steps: extrusion, excitation, and scale. The weight vector of the feature map is obtained by extrusion and excitation, and the feature map is recalibrated through the scale step. The feature map of size $H \times W \times C$ gets $1 \times 1 \times C$ feature map through the squeeze function, and the attention vector of size $1 \times 1 \times C$ is obtained through the excitation function, and the value of the attention vector reflects the value of the channel. The degree of importance. The larger the value, the more important the channel is. Multiply the attention vector and input to get a recalibrated feature map of size $H \times W \times C$. The formula is as follows.

$$y = F_{sq}(X) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} X(i,j) \quad (1)$$

$$s = F_{ex}(y, W) = \sigma\big(W_2 \delta(W_1 z)\big) \quad (2)$$

$$X' = F_{scale}(X_c, s_c) = X_c \cdot s_c \quad (3)$$

Equation (1) is the squeezing function, and squeezing is achieved by global pooling. H and W are the height and width of the image, respectively. Equation (2) is the excitation function, implemented by two fully connected operations. W_1 and W_2 are the weight matrices of two fully connected operations, $W_1 \in R^{\frac{C}{r} \times C}$ and $W_2 \in R^{C \times \frac{C}{r}}$. δ refers to the ReLU function. σ refers to the sigmoid function. The input goes through a fully connected layer, through the ReLU function, through a fully connected layer, through the sigmoid function, and through two nonlinear functions to make the attention vector a richer representation. Equation (3) is the scale function, where s is the attention vector, X is the input feature map, $s \in 1 \times 1 \times C$ and $X \in R^{H \times W}$. s_c is the value of the attention vector on channel c, X_c is the c-th channel of X, and the input feature map is re-corrected by multiplying the attention vector and the input feature map.

2.3 ResNeSt

ResNeSt is a backbone based on ResNet (He et al., 2015). By adding a split-attention module while retaining the residual structure, the model can make better use of cross-channel information while retaining the advantages of ResNet, so that it can have better results. The core idea of the method in this paper can extract richer channel information by grouping and splitting the feature map into more branches. As illustrated by Figure 3.

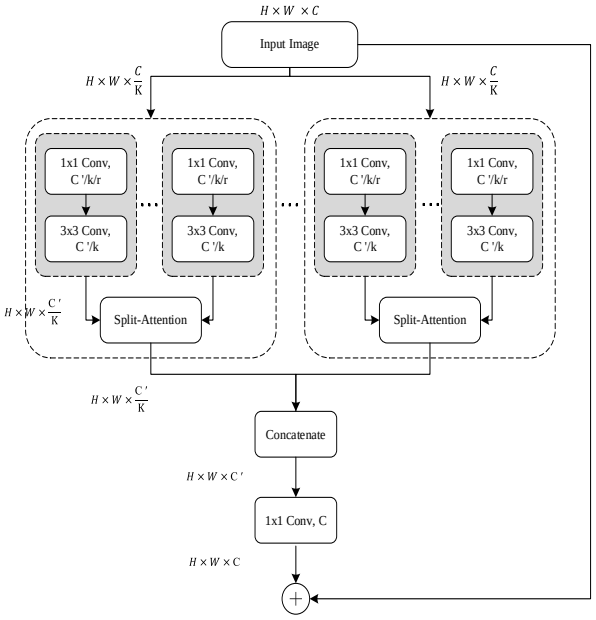

Figure 3. ResNeSt architecture

ResNeSt first divides the input feature map into K groups in the channel dimensions, and each group is divided into R splits. Now there are a total of $G = K \times R$ branches, and the sub-feature map size of each branch is $H \times W \times \frac{C}{K \times R}$, then through two convolutional networks, the size of the sub-feature map is expanded to the original R times as $H \times W \times \frac{C}{K}$, and the R splits of the same group are input into split-attention according to the extracted channel-wise attention vector re-calibration feature map. Concatenate the K groups of re-calibrated feature maps to obtain feature maps of size $H \times W \times \frac{C}{K}$ containing more diverse representations. Finally, the residual connection is added to prevent the problem of gradient explosion during the training process, so that the model can be trained better.

Split-Attention module, first add R splits, and then perform global average pooling to obtain a channel information vector of size $1 \times 1 \times C$. Adjust the size of the channel dimensions through two MLPs to obtain R channel information vectors of size $1 \times 1 \times C$. In order to ensure the independent distribution of the weights of the R splits, r-softmax is used to calculate the softmax of the weights of each split to obtain the corresponding channel attention weights. Finally, each split is multiplied by the channel attention weight and then summed. The Split-Attention module adds the feature maps of the same group and calculates the attention weight of the channel dimension. By readjusting each split channel, the important features are strengthened, and the unimportant noise is ignored, so as to improve the feature extraction accuracy. Effect.As illustrated by Figure 4.

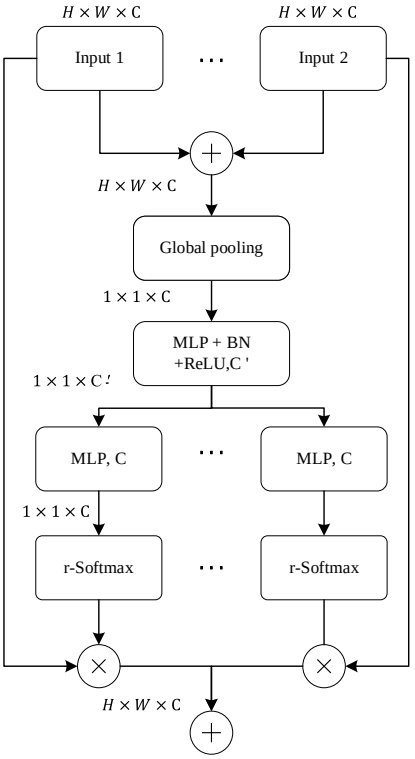

Figure 4. Split-Attention architecture

2.4 Deeplabv3+ with ResNest

In this paper, we use ResNeSt as the backbone network to extract the input features, which can better utilize the information of channel dimension than DCNN, and can improve the accuracy of road extraction without increasing the computation. The ASPP module can increase the perceptual field, obtain more feature information, and improve the accuracy of road extraction without decreasing the resolution. The network structure of this paper is shown in Figure 5.

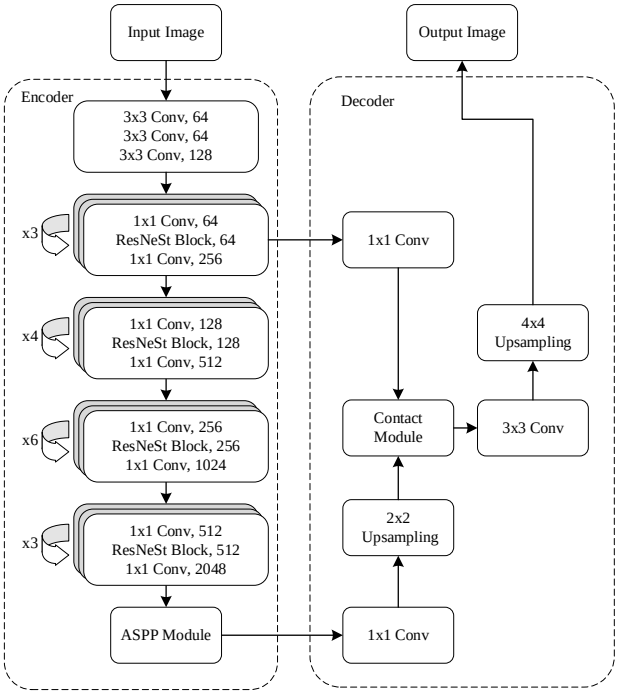

Figure 5. Deeplabv3+ with ResNeSt

First, the image is fed into the backbone network, and after the stem block, the channel of the image becomes 128, and the size of the feature map becomes $\frac{1}{4}$ of the original one. after the second block, the channel of the image becomes 256, and the size of the feature map is still $\frac{1}{4}$ of the original one, and this feature map is divided into two parts, the first part goes to the decoder and waits for feature fusion, and the second part continues The first part goes to the decoder to wait for feature fusion, and the second part continues feature extraction. The feature map output from backbone is fed into ASPP module, which extracts the multi-scale features of the feature map. The feature map from the ASPP module is double upsampled and stitched with the output of the second block, and the convolution with kernel size 3×3 is used for feature fusion, and the prediction map of the same size as the original map is obtained by quadruple upsampling.

The ASPP module includes a traditional convolution with kernel size 1×1, three atrous convolutions with kernel size 3×3, and image pooling. The dilation rates of the three atrous convolutions are 12, 24, and 36, respectively. Concatenate the five output feature maps to obtain the feature maps output by the ASPP module. A feature map with multi-scale information is obtained by concatenate the feature maps extracted from different receptive fields. As illustrated by Figure 6.

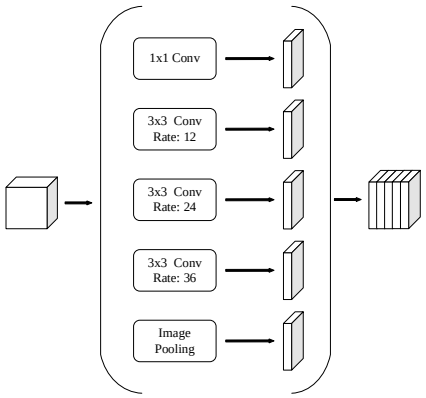

Figure 6. ASPP Module

3. EXPERIMENTS AND RESULTS

3.1 Data sets

The experiment selects the Deepglobe road extraction dataset as the experimental data. This paper uses a dataset containing 4930 training images, 1233 validation images, and 63 testing images, each of which is a three-channel RGB image of size 1024×1024 pixels. Because training the neural network requires a lot of data, at the same time, in order to prevent the neural network from overfitting and improve the robustness of the neural network, the training data is randomly cropped, scaled, image color changes, and flipped to enhance the final data set.

3.2 Implementation Details

In order to demonstrate the feasibility of Deeplabv3+ with ResNeSt as the backbone network for road extraction and the advantages of Deeplabv3+ with U-Net network and ResNet as the backbone network, this paper conducts comparative experiments on the same training, validation and test sets. The operating system of the computer used for the experiments is Windows 10, the hardware configuration is NVIDIA GeForce RTX 3060 with 6 GB of video memory and 16 GB of RAM, and the software configuration uses Pytorch version 1.11.0 and CUDA version 11.3 as the deep learning framework to build the network. The experimental parameters are shown in Table 1.

Parameter Name	Parameter Value
Dropout Ratio	0.1
Learning Rate	0.001
Loss Function	Cross Entropy Loss
Optimazer	SGD

Table 1. Parameters involved in the experiment

3.3 Metric

In this experiment, two evaluation metrics, Intersection over Union (IoU) and Accuracy (Acc) (Heipke et al., 1997), are used to evaluate the prediction results. Road feature extraction is to divide the images into two categories: road and background, and Intersection over Union is to calculate the ratio between the intersection and the concurrence of the two sets of true and predicted values. The accuracy ratio is calculated as the ratio of correct predictions among all samples. Its calculation formula is as follows.

$$\mathrm{IoU}=\frac{\mathrm{TP}}{\mathrm{FN}+\mathrm{FP}+\mathrm{TP}} \tag{3}$$

$$\mathrm{Acc}=\frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}} \quad (4)$$

In the above equation, TP (true positive) is the number of pixels where the road is correctly classified as a road, FP (false positive) is the number of pixels where the background is predicted as a road, and FN (false negative) is the number of pixels where the road is predicted as a background, and TN (true negative) is the number of pixels where the background is correctly classified as a background.

3.4 Results and Analysis

We test the proposed method and other methods on the test set and compare the predicted images. The operating environment of the experiment is the same, and the hyper parameter settings are the same. Some results are shown in Figure 6. The images from left to right are satellite images, ground truth, Deeplabv3+, and the method proposed in this paper.

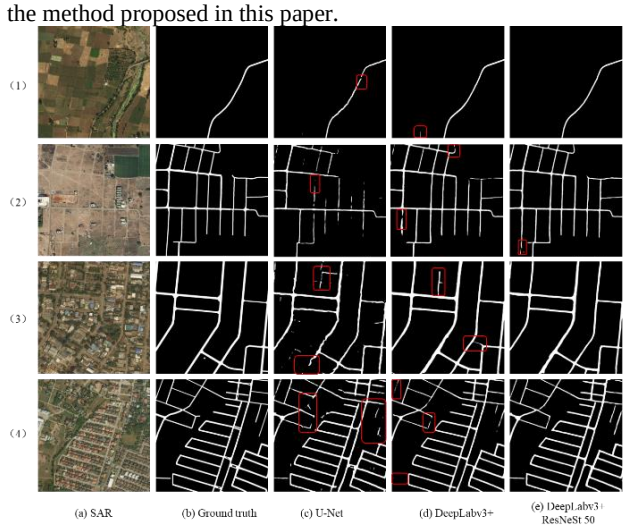

Figure 6. The visualization of results on the DeepGlobe test set

From Figure. 6 (1), we can see that for roads with simple shapes, the roads extracted by U-Net have the problem of discontinuity, and the traditional Deeplabv3+ also incorrectly identifies land as roads, while the method in this paper accurately identifies roads and land. From (2)(3), we can see that although all three methods have different degrees of identified road discontinuities, the method proposed in this paper has the highest recognition accuracy. From (4), it can be seen that there is a problem of discontinuity in the roads identified by U-Net, and Deeplabv3+ has a problem that some roads are not identified, while the method proposed in this paper has the best recognition effect.
The evaluation metrics for the three networks are shown in Table2.

method	IoU	Acc
U-Net	53.22	68.04
DeepLabv3+	61.89	71.25
ResNeSt	63.15	73.16

Table 2. A comparison of our method with other method on Deepglobe data sets.

As shown in Table 2, the method proposed in this paper improves the IoU by 1.25% and the accuracy by 1.88% compared with the traditional Deeplabv3+. Compared with U-Net, the method proposed in this paper improves the IoU by 9.93% and the accuracy by 5.12%.

4. CONCLUSIONS

Extracting roads from high-resolution remote sensing images has always been a challenging task. In order to solve the problems of incomplete structure, poor topological connectivity, and poor accuracy of road extraction, this paper proposes an improved method on the Deeplabv3+ network. network structure. And compared with the U-Net network and the traditional Deeplabv3+ network. The experimental results show that the method proposed in this paper has been significantly improved in both the extraction effect and the two metrics. However, the method in this paper needs to be further improved. For some narrow roads still cannot be identified or incompletely identified, future work will mainly study how to improve the extraction ability of narrow roads.

ACKNOWLEDGEMENTS

This work was jointly supported by Scientific research project of Beijing Education Commission (Z20006, Research on quality detection method of land cover classification based on depth confidence neural network), National key research and development program (2021YFB2600101, Tracking, monitoring and three-dimensional patrol technology of land transportation infrastructure ecological environment impact) and National key research and development program (2021YFE0194700, Common product services, promotion and application of global ten meter domestic satellite data and typical land surface elements).

REFERENCES

Alshehhi, R., Marpu, P.R., 2017. Hierarchical graph-based segmentation for extracting road networks from high-resolution satellite images. ISPRS Journal of Photogrammetry and Remote Sensing 126, 245–260. https://doi.org/10.1016/j.isprsjprs.2017.02.008

Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L., 2016. Semantic Image Segmentation with Deep Convolutional Nets and Fully Connected CRFs.

Chen, L.-C., Papandreou, G., Schroff, F., Adam, H., 2017. Rethinking Atrous Convolution for Semantic Image Segmentation. https://doi.org/10.48550/arXiv.1706.05587

Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H., 2018. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. https://doi.org/10.48550/arXiv.1802.02611

He, K., Zhang, X., Ren, S., Sun, J., 2015. Deep Residual Learning for Image Recognition. https://doi.org/10.48550/arXiv.1512.03385

Hou, Y., Liu, Z., Zhang, T., Li, Y., 2021. C-UNet: Complement UNet for Remote Sensing Road Extraction. Sensors 21, 2153. https://doi.org/10.3390/s21062153

Hu, J., Shen, L., Albanie, S., Sun, G., Wu, E., 2019. Squeeze-and-Excitation Networks.

Knudsen, E.I., 2007. Fundamental Components of Attention. Annual Review of Neuroscience 30, 57–78. https://doi.org/10.1146/annurev.neuro.30.051606.094256

Li, Y., Xu, L., Rao, J., Guo, L., Yan, Z., Jin, S., 2019. A Y-Net deep learning method for road segmentation using high-resolution visible remote sensing images. Remote Sensing Letters 10, 381–390. https://doi.org/10.1080/2150704X.2018.1557791

Liu, J., Qin, Q., Li, J., Li, Y., 2017. Rural Road Extraction from High-Resolution Remote Sensing Images Based on Geometric Feature Inference. ISPRS International Journal of Geo-Information 6, 314. https://doi.org/10.3390/ijgi6100314

Long, J., Shelhamer, E., Darrell, T., 2015. Fully Convolutional Networks for Semantic Segmentation. Presented at the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3431–3440.

Park, J., Woo, S., Lee, J.-Y., Kweon, I.S., 2018. BAM: Bottleneck Attention Module. https://doi.org/10.48550/arXiv.1807.06514

Ronneberger, O., Fischer, P., Brox, T., 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation, in: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (Eds.), Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015, Lecture Notes in Computer Science. Springer International Publishing, Cham, pp. 234–241. https://doi.org/10.1007/978-3-319-24574-4_28

Wang, Q., Gao, J., Yuan, Y., 2018. Embedding Structured Contour and Location Prior in Siamesed Fully Convolutional Networks for Road Detection. IEEE Trans. Intell. Transport. Syst. 19, 230–241. https://doi.org/10.1109/TITS.2017.2749964

Woo, S., Park, J., Lee, J.-Y., Kweon, I.S., 2018. CBAM: Convolutional Block Attention Module. Presented at the Proceedings of the European Conference on Computer Vision (ECCV), pp. 3–19.

Xia, W., Zhang, Y.-Z., Liu, J., Luo, L., Yang, K., 2018. Road Extraction from High Resolution Image with Deep Convolution Network—A Case Study of GF-2 Image. Proceedings 2, 325. https://doi.org/10.3390/ecrs-2-05138

Xie, Y., Miao, F., Zhou, K., Peng, J., 2019. HsgNet: A Road Extraction Network Based on Global Perception of High-Order Spatial Information. IJGI 8, 571. https://doi.org/10.3390/ijgi8120571

Zhang, Q.-L., Yang, Y.-B., 2021. SA-Net: Shuffle Attention for Deep Convolutional Neural Networks, in: ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Presented at the ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 2235–2239. https://doi.org/10.1109/ICASSP39728.2021.9414568

Zhong, Z., Li, J., Cui, W., Jiang, H., 2016. Fully convolutional networks for building and road extraction: Preliminary results, in: 2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS). Presented at the 2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), pp. 1591–1594. https://doi.org/10.1109/IGARSS.2016.7729406