An Improved Semantic Segmentation Model for Remote Sensing Images based on HRNet

Xiaolan Zhou
University of Chinese Academy of Science, China
460 Main Street, Saco, Maine, 04072
*Corresponding author’s e-mail: Xiaolan.zhou22@thorntonacademy.org

Abstract. Deep learning has been greatly improved recently, and natural image processing based on deep learning has also been greatly improved. However, there are still great differences between natural images and remote sensing images, among which the biggest is that the size of the target in remote sensing images is greatly different, which requires the model to have a strong multi-scale processing ability. In order to meet this goal, we use HRNet with full multi-scale fusion capability to replace ResNet to process remote sensing images. HRNet fully integrates low-level detail features, middle-level structure features and high-level semantic features, which is very suitable for remote sensing images. The experimental results show that our method has been greatly improved.

1. Introduction
Semantic segmentation of remote sensing image refers to the semantic classification of pixels in each region of the image, which plays a very important role in the extraction of surface spatial information, urban land resource management, environmental monitoring and natural resource protection. With the development of remote sensing technology, the acquisition of high-resolution remote sensing images has been widely used [1], which provides a good resource base for the extraction of urban land information. The traditional method uses manual method to annotate remote sensing image information, which takes a lot of time and manpower. Therefore, it is of great importance to construct an automatic semantic segmentation method for high resolution remote sensing image. Different from traditional computer vision images, remote sensing images generally have a relatively small quantity, and an image will contain a lot of objects, such as roads, buildings, vegetation, tall trees, cars and so on. In addition, buildings have different sizes, cars are small compared to other objects, and vegetation and tall trees are only different in height. These problems add great difficulties to the semantic segmentation of label-based images. In recent years, with the development of deep learning technology, many scholars have done a lot of research on image semantic segmentation in order to solve the difficulties of image semantic segmentation. In 2015, Fully Convolutional neural Networks (FCN) [2] provides a new basic model for image semantic segmentation. It adopts the structure of "encoding and decoding" and realizes the end-to-end segmentation method, which is a great improvement compared with other models.

Aiming at the characteristics of high resolution remote sensing image, this paper designs an improved end-to-end network model. The main improvement is to use HRNET instead of ResNeT and make some targeted improvements. HRNet [3] can maintain high-resolution feature maps throughout the whole process to obtain more accurate spatial information. Its multi-scale fusion strategy can also
obtain richer high-resolution representations, making the heat-map forecast more accurate. Using HRNet to replace the global branch with RerNet, we can get the side output feature map with higher resolution and more abundant feature information, which can improve the accuracy of segmentation.

To sum up, the main work of this paper is as follows:

1. We analyze the difference between remote sensing image and natural image, and use HRNet as feature extraction backbone to replace ResNeT to meet the segmentation requirements of objects with various sizes.
2. We set up complete experiments and compare our method with various methods. Experimental results on high resolution remote sensing image data sets show that the proposed network is superior to the previous methods.

2. Method

2.1. The network structure

The basic network design is an important part of the design of a segmented network. Due to the limitation of conditions, it is difficult to obtain massive data sets of remote sensing images during the training of high-resolution remote sensing images. Generally, there are only a few or dozens of images. It is a good choice to extract features from such a small number of data sets and adopt image classification model with pre-training. HRNet is selected as the image feature extraction model in this paper, as shown in Figure 1. The network is mainly composed of HRNet and FPN [8]. We focus on how HRNet fits into remote sensing images. HRNET consists of parallel multi-resolution subnet and repeated scale fusion, which are described below.

2.2. Parallel multi-resolution subnet

Figure 1. Overview of our proposed network
The basic network design is an important part of the design of a segmented network. Due to the limitation of conditions, it is difficult to obtain massive data sets of remote sensing images during the training of high-resolution remote sensing images. Generally, there are only a few or dozens of images. It is a good choice to extract features from such a small number of data sets and adopt image classification model with pre-training. HRNet is selected as the image feature extraction model in this paper, as shown in Figure 1. The network is mainly composed of HRNet and FPN [8]. We focus on how HRNet fits into remote sensing images. HRNET consists of parallel multi-resolution subnet and repeated scale fusion, which are described below.

Figure 2. Parallel multi-resolution subnet
The parallel multi-resolution subnet is constructed by parallel connection from high resolution to low resolution subnets. Each subnet contains multiple convolution sequences, and there is a down sampling layer between adjacent subnets to halve the resolution of the feature map. With high resolution subnet as the first stage, gradually add from high to low resolution subnets to form a new stage, and then connect multiple resolution subnets in parallel. The resolution of the latter phase of the parallel subnet consists of the resolution of the previous phase and the resolution of the next phase. The network structure composed of four parallel subnets is shown in Fig. 2(a). In the figure: \( N_{SR} \) represents the subnet in phase S, and its resolution is \( 1/(2r-1) \) of the image in the initial phase.

### 2.3. Repetitive scale fusion

HRNet introduces switching units across parallel subnets so that each subnet receives information from other subnets multiple times. An example of an information exchange unit is shown in Figure 2(b). Phase 3 is divided into multiple swap blocks, each of which consists of three parallel convolution units and one swap unit. In the figure: CSR\(_{RB}\) represents the B switching block in the S stage, and the resolution of the switching unit is \( 1/(2r-1) \) of the initial stage, while SB represents the corresponding switching unit. The specific realization of HRNet exchange unit aggregating feature information of different resolutions is shown in Figure 3.

The exchange unit is shown as a response graph \{X1, X2... Xs\} as the input, each output is aggregated by the response graph of the input, and the corresponding output is \{Y1, Y2... Ys\}, where, the resolution and dimension of Yi and Xi are the same. The switching unit of HRNet uses a 3x3 convolution with a step size of 2 for down-sampling, while the up-sampling is realized by bilinear interpolation. It is worth noting that if \( I = k \), then \( a(x_i, k) \) represents an identity mapping, that is, \( a(x_i, k) = x_i \).

### 3. Experiments

#### 3.1. Design of experimental scheme

The selected data set is Vaihingen data set [4], which collects standard aerial remote sensing images taken by self-photographed aircraft and consists of 33 high-resolution aerial images covering an area of 1.38km\(^2\). The average size of images is 2494x 2064, and each image has 3 bands. NDSM images represent the height data of objects on the ground as supplementary data input. Among the 33 images, 16 of them were manually labeled. We selected 12 of them as the training set and 4 as the verification set.

In general, the size of a single image of a high-resolution remote sensing image is too large to be directly input into the deep learning network, and most high-resolution remote sensing images only provide a very limited amount of data. For example, the Vaihingen dataset only provides 16 complete images with a size of 2494x2064 and a label. Although there are many deep learning can enter any size of the image semantic segmentation model, but due to the restrictions of the GPU memory and image number of reasons, an input for such a big image is obviously not appropriate, we need to
random cropping of the image, we in the original training images, on the basis of random cropping the image size is 256 x 256, And the image into 0°, 90°, 180°, 270°, horizontal and vertical 6 directions of random rotation, in the realization of the process of image clipping and training is not separate, so that you can ensure that each random image can be different.

During the design, our program was designed using the PyTorch framework. The image workstation used in the experiment was configured as: 8-core CPU, 32G internal memory, Teslav100 GPU, 16G video memory, Ubuntu 16.04 operating system. The optimizer adopted the stochastic gradient method, and the parameters were set as follows: Lr = 0.01, Momentum = 0.9, weight _ decay = 1e-4, number of iterations 50000, batch size 16

3.2. Analysis of experimental results

Four different semantic segmentation networks, FCN-8S, UNET [5], SEGNET [6], and DEEPLABV3 + [7], are adopted to carry out the score analysis. As can be seen from the data in Table 3, the network designed by us has improved to a certain extent in the equalization and parallel ratio (MIOU) and accuracy (OA). Compared with the basic network FCN-8S, our network improves by 5.1% on MIOU and 3% on OA, which proves that our network is effective. Our model also has some improvement in the recognition of small objects. For example, the IOU of the car category has reached 73.06%, and it is better than other models in the recognition of similar objects. For example, when buildings are large and have different colors, it is difficult to recognize them, and there is often the phenomenon of missing pixels in the middle. The proportion of IOU of building category in our model reached 90.78%, and the recognition was relatively complete from the forecast map.

To evaluate the performance of the network, we use global accuracy (OA) and equal-and-parallel ratio (MIOU) for comparison. The evaluation function is as follows: OA = TP/N (3) IOU= TP TP+FP+FN (4) TP represents the pixels of "positive sample is classified as positive sample", FP table shows the pixels of "negative sample is classified as positive sample", FN represents the pixels of "positive sample is classified as negative sample", and N represents the total pixel value.

4. Conclusion

We propose a segmentation network for remote sensing images, and change the feature extraction network into HRNET according to the characteristics of remote sensing images. Our work is more effective on the data set than the previous methods.

Acknowledgments

Authors wishing to acknowledge assistance or encouragement from colleagues, special work by technical staff or financial support from organizations should do so in an unnumbered Acknowledgments section immediately following the last numbered section of the paper.

References

[1] Qiu C,Mou L,Schmitt M,et al. Local Climate Zone-Based Urban Land Cover Classification from Multi-Seasonal Sentinel-2 Images with a Recurrent Residual Network[J]. ISPRS J Photogramm Remote Sens,2019,154,151-162.

[2] Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation[C]/Proceedings of the IEEE conference on computer vision and pattern recognition. 2015: 3431-3440.

| Network  | Road  | building | vegetation | tree  | car   | Overall (mIOU) | OA    |
|----------|-------|----------|------------|-------|-------|----------------|-------|
| FCN-8s [1] | 80.32 | 83.03    | 67.44      | 73.44 | 58.31 | 72.51          | 85.91 |
| U-Net [5]   | 79.9  | 86.07    | 59.8       | 75.68 | 71.31 | 74.554         | 87.52 |
| SegNet [6]   | 81.2  | 86.6     | 68.6       | 74.9  | 68.2  | 75.92          | 87.82 |
| Deeplabv3+ [7] | 80.81 | 88.65    | 62.13      | 78.24 | 72.14 | 76.39          | 87.90 |
| Ours        | 82.96 | 91.78    | 62.45      | 78.49 | 73.05 | 77.57          | 88.92 |
[3] Wang J, Sun K, Cheng T, et al. Deep high-resolution representation learning for visual recognition[J]. IEEE transactions on pattern analysis and machine intelligence, 2020.

[4] 2D Semantic Labeling Vaihingen Data. Available online: http://www2.isprs.org (accessed on 5 November 2019).

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

[6] 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): 2481-2495.

[7] L C, Zhu Y, Papandreou G, et al. Encoder-decoder with atrous separable convolution for semantic image segmentation[C]//Proceedings of the European conference on computer vision (ECCV). 2018: 801-818.

[8] Lin T Y, Dollár P, Girshick R, et al. Feature pyramid networks for object detection[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 2117-2125.