Semantic segmentation network of uav image based on improved U-net

Ziyi Liu¹, Jin Huang¹,*

¹School of electrical engineering, southwest jiaotong university, chengdu, China
*Corresponding author e-mail: jhuang@swjtu.edu.cn

Abstract. The u-net, which is popular in the field of deep learning, is improved on the basis of the Fully Convolutional Network (FCN) [1]. U-net [2] was first published in the field of medical image, because the network has many advantages, so many scholars put their migration to use on the other computer vision problems, finding its on different problems still show the excellent performance. In this paper, U-net is applied to the problem of uav image processing, and the semantic segmentation of uav images at pixel level is carried out to identify the types of ground objects and the pixels contained in them. In order to increase the recognition accuracy of neural network, this paper proposes an improved network model based on u-net, and applies it to this problem. By training, verifying and testing the same small sample data set, it is found that the improved network has better recognition accuracy than U-net.

1. Introduction
Convolutional Neural Network (CNN) can classify images, but no one can recognize the image at pixel-level before 2015. The fully convolutional network proposed by Jonathan at that time became the pioneer of this problem. FCN classifies images at pixel-level and solves the problem of semantic segmentation. FCN can accept any size of the input image and use the deconvolution on feature map to up-sample, making feature map back to the same size of input image, which can produces a class prediction for each pixel. After that, many semantic segmentation networks inherited the idea of FCN and improved on the basis of its structural framework. The best one is U-net, which has excellent segmentation performance and can adapt to segmentation tasks in various scenarios. Until now, it has been widely used in different segmentation problems.

Drozdal et al. [3] imported residual structure in U-net to make the network converge faster during training and support the use of deeper network structure. Jegou et al. [4] imported dense connection structure in U-net and realized multi-scale supervision. Badrinarayanan et al. [5] improved the up-sampling in the decoder. Instead of using deconvolution operation, they used the maximum pooling index of the decoder in corresponding scale to conduct sparse up-sampling, in order to reduce the training cost. Chen et al. [6] used atrous convolution in the convolution after the last second maximum pooling, and used conditional random field (CRF) to recover boundary details and achieve accurate positioning performance. Zhao et al. [7] imported atrous convolution into the encoder of ResNet structure and imported auxiliary loss in the mid-layer to optimize the learning. Lin et al. [8] used fine-grained features in early convolution to refine and capture deeper network layers, and realized high-resolution prediction by using remote residual connection. Islam et al. [9] solved the problem of ambiguity in the decoder by using gating to refine the feedback unit and control the information flow from the encoder to the decoder. Zhou et al. [10] improved the U-net structure by using a series of
nested and dense jump connections in the codec subnetwork, which effectively improved the network accuracy.

2. FCN and U-net

2.1. FCN
The power of CNN is that its multi-layer structure can automatically learn features, and it can learn features of multiple levels. The shallow convolution layer has a smaller perceptual domain and learns some local features, while deep convolution layer has a larger perceptual domain and can learn more abstract features. These abstract features are very helpful for classification, and can be used to determine what classes of objects are included in an image. However, due to the loss of some details of objects, the specific outline of objects cannot be given well and cannot define which object the pixel belongs to. Therefore, it is very difficult to achieve accurate segmentation.

The difference between FCN and CNN is that the fully connected layers at the end of CNN are replaced by the convolution layer, and the output is a predicted image. The process is equivalent to recovering the category of each pixel from the abstract features sampled from the convolution network. That is, from the classification of image level to the classification of pixel level. FCN also has some inherent defects. For example, the result is not precise enough. The result of up-sampling is fuzzy and smooth, and it is insensitive to the details in the image. Secondly, the relationship between pixels is not fully considered and the spatial consistency is lacking.

2.2. U-net
Many segmentation networks are based on FCN, including U-net. Like FCN, U-net is divided into feature extraction by down-sampling and prediction by up-sampling. The two parts are put together to form a U-shaped network structure, so it is named U-net. The improvement of U-net over FCN lies in the multi-scale network structure, where more detailed pixel information can be seen and features extracted from different scales can be combined to enable the network to not only conduct pixel segmentation at the micro level, but also see more neighbor information to correctly classify pixels. In addition, the network has no full connection layer, which can process input image of different scales and return output segmentation results of the same size, including support for large images.

![Figure 1. U-net structure](image)

2.3. An improved semantic partitioning network atrous-unet based on U-net
After the emergence of U-net, many scholars have improved on its foundation, such as UNet ++. Inspired by some improvement work, this paper improves U-net by combining with other improvement schemes. First, image scale of input and output is not completely equal in U-net. The network structure can be seen in figure 1, the output predictive label image is somewhat smaller than the input original image, the reason is that no pixel padded before each convolution, leading each
convolution operation make the feature graph smaller. Therefore, in order to make the network input and output images of the same size, zero pixels will be padded before each convolution operation to prevent the image from shrinking. Secondly, the intermediate step of atrous convolution is added on the basis of U-net multi-scale convolution, and the convolution layer with large kernel is added before the final prediction of output label. The purpose of the above two methods is to improve the accuracy of network prediction by combining the broader neighbor feature information in pixel prediction, as shown in figure 2.

![Atrous-unet structure](image)

**Figure 2. atrous-unet structure**

### 3. Predict land classification through uav image

In this chapter, U-net is firstly built in the environment of Python3.7 and Tensorflow. Uav image data set is used for training on a single K80 calculating video card, with 12GB memory and 3.7 calculating force of single video card. Latter, atrous-unet network was built and measured by using the same hardware and data set.

#### 3.1. Introduction of small sample data set and prediction task

There are 280 oversized uav photos with marked labels, each of which is a 5472X3648 high-definition image, mainly including road, house, wasteland, weed field, forest and different kinds of farmland at the top view angle. The prediction problem is to find out 6 kinds of ground objects and the distribution of pixel of each kind of ground objects from uav images. The six geographical features in order are rice, corn, taro, house, tree and house. The entire 280 image data sets are randomly divided into 250 training sets and 30 test sets, as shown in figure 3 and figure 4.

![Train set sample diagram](image)

**Figure 3. Train set sample diagram**
When two kinds of neural networks are used for prediction, they actually train on a single K80 GPU, save the training results of 10 and 15 times of traversing the training set respectively, and then put the training results on the general CPU and use the test set for prediction. Since the original uav image is very large, and the video memory of a single GPU is limited to 12GB, it is not suitable for neural network to make one-time prediction. Therefore, each picture should be divided into 64 small pictures according to the two network video memory occupation. After the training, the pictures with same shape were used in the test, and finally, the small pictures were spliced into the prediction results of the size of the original image, and the average prediction accuracy of each small picture was used as the prediction accuracy of each original image.

3.2. Implementation and test of neural network
Firstly, the U-net network is implemented by code, and the training results of traversing the train set for 10 and 15 times are obtained by using the small sample data set of uav image. The highest accuracy of 71.64% was obtained by testing each result, as shown in figure 5 and figure 6. For the prediction of test set diagram below, the black area represents the pixels predicted as background, red pixels representing prediction for rice, green pixels representing prediction for corn, yellow pixel represents prediction for taro, blue pixel represents prediction for house, purple pixel represents prediction for soybean, light blue pixels on behalf of the prediction for the tree, as shown in the following table 1.

| Label class | Label color | Label legend |
|-------------|-------------|--------------|
| Background  | black (R0, G0, B0) |  |
| Rice        | red (R255, G0, B0) |  |
| Corn        | green (R0, G255, B0) |  |
| Taro        | yellow (R255, G255, B0) |  |
| House       | blue (R0, G0, B255) |  |
| Soybean     | purple (R255, G0, B255) |  |
| Tree        | light blue (R0, G255, B255) |  |
After that atrous-unet is implemented, and 10 and 15 times of training results are obtained by using small sample data set. The two results were higher than the highest accuracy of U-net, 73.66% and 72.67% respectively, as shown in figure 7 and figure 8. For the small graph after cutting (684x456), the network prediction time was only 1.2s. By comparing samples from all the test sets, atrous-unet's segmentation of image details is superior because shallow networks also incorporate context information in empty convolution.

| Network     | Highest accuracy |
|-------------|------------------|
| U-net       | 71.64%           |
| Atrous-unet | 73.66%           |

Figure 5. performance in test set with u-net traversing 10 times of train set

Figure 6. performance in test set with u-net traversing 15 times of train set

Figure 7. performance in test set with atrous-unet traversing 10 times of train set
4. Conclusion

In this paper, U-net is applied to the problem of uav image processing, and the semantic segmentation of images at pixel level is carried out to identify the types of ground objects and the pixels contained in them. In order to increase the accuracy of neural network identification, atrous-unet, an improved network model based on U-net, is proposed in this paper. Using the same small data set for training, validating and testing, we found that the improved network has better recognition accuracy than U-net, which can reach more than 73%. However, due to some data set limitations, such as some classes have too few samples, making the predictions are partial to the classes with large sample number, and the recognition rate of some categories needs to be improved.

Acknowledgments

This work was financially supported by Chengdu Science and Technology Bureau(2018-YF05-01424-GX), Information Center of Sichuan Land and Resources Department (KJ-2018-16).

References

[1] Long, Jonathan, E. Shelhamer, and T. Darrell. "Fully Convolutional Networks for Semantic Segmentation." IEEE Transactions on Pattern Analysis & Machine Intelligence 39.4(2014):640-651.
[2] Ronneberger, Olaf , P. Fischer , and T. Brox . "U-Net: Convolutional Networks for Biomedical Image Segmentation." (2015).
[3] Drozdzal, Michal , et al. "The Importance of Skip Connections in Biomedical Image Segmentation." International Workshop on Large-Scale Annotation of Biomedical Data and Expert Label Synthesis International Workshop on Deep Learning in Medical Image Analysis Springer International Publishing, 2016.
[4] Jégou, Simon, et al. "The One Hundred Layers Tiramisu: Fully Convolutional DenseNets for Semantic Segmentation." (2016).
[5] Badrinarayanan, Vijay , A. Kendall , and R. Cipolla . "SegNet: A Deep Convolutional Encoder-Decoder Architecture for Scene Segmentation." IEEE Transactions on Pattern Analysis and Machine Intelligence (2017):1-1.
[6] Chen, Liang Chieh , et al. "Semantic Image Segmentation with Deep Convolutional Nets and Fully Connected CRFs." Computer Science (2014).
[7] Zhao, Henghuang, et al. "Pyramid Scene Parsing Network." (2016).
[8] Lin, Guosheng , et al. "RefineNet: Multi-Path Refinement Networks for High-Resolution Semantic Segmentation." (2016).
[9] Islam, Md Amirul , et al. "Gated Feedback Refinement Network for Dense Image Labeling." 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) IEEE, 2017.
[10] Zhou, Zongwei , et al. "UNet++: A Nested U-Net Architecture for Medical Image Segmentation." (2018).