Building extraction based on improved message passing network

Manlin Zhang¹, Xiaoming Guo²* and Ying Bao³

¹ Liaoning Petrochemical University, Fushun, Liaoning, 113000, China
*Corresponding author’s e-mail: panbin@lnpu.edu.cn

Abstract. In this paper, an improved building edge extraction method based on message passing networks (MPNNs) is proposed. In the field of high-resolution remote sensing image processing and analysis, the accuracy of building edge extraction is studied from three aspects: data set construction, network training and post-processing of edge probability map. Based on the WHU building data set, the proposed method can effectively extract the building edge, and can maintain high extraction accuracy. PPG Net, L-CNN and UNet++ algorithm is used to compare with the results of this paper. Through the analysis of experiments, the results show that the method proposed in this paper has advantages over PPG Net, L-CNN and UNet++ algorithm in the accuracy, recall and F1 score of region extraction.

1. Introduction

Building edge extraction is an important research branch in the field of three-dimensional reconstruction. It combines the latest research results of image processing, pattern recognition and other disciplines. It mainly analysis the high-resolution remote sensing image to realize the building edge extraction.

In recent years, convolutional neural network has shown good performance in computer vision target classification, image segmentation and other aspects. Many scholars have convolutional neural network to remote sensing field and gradually improved it.

First of all, the VGGNet[1] proposed by Simonyan et al. The feature map generated by the convolution module into a fixed-length feature vector, which is used for image-level classification tasks. But it cannot be directly applied to the extraction of buildings from remote sensing images. On this basis, Badrinarayanan et al. proposed a code-decoding structure, SegNet, for semantic segmentation. The up-sampling part of the decoder integrates the indexes in the maximum pooling of the corresponding encoder, which is conducive to feature recovery.

Later, Ronneberger et al. proposed a deep convolutional neural network (UNet) architecture for medical image segmentation in the paper. The same code-decoding structure was used. The encoding part was used to obtain the feature information of different scales, and the decoding part gradually restored the image resolution to achieve per-pixel classification. On the basis of UNet, Zhou et al. proposed a better medical image deep supervised segmentation network UNet++[2], which improved the extraction efficiency.

The above research has improved the accuracy of building extraction from remote sensing images, but there are still some shortcomings. To further improve the effect of building edge extraction, this paper combined convolutional Message Passing network and extended Residual network, and proposed a building extraction method based on Message Passing network called DR-MPNNs (Dilated Residual Networks -- Message Passing Neural Networks).
2. Materials and Methods

2.1. Building extraction method based on message passing network

2.1.1. Network structure
This chapter combines convolutional message passing network and extended residual network, and proposes a method of building extraction based on message passing network, which is called DR-MPNNs (differentiated residual networks message passing neural networks). The whole framework is a standard convolutional messaging network framework, as shown in Figure 1. The network is divided into two parts:

1) Feature extraction module
   In the first part, the Residual Network [3][4] Rest Net 50 is used to replace VGG-16 as the backbone network to generate candidate regions;
   In the second part, the first three convolution layers of the extended Residual Network DRN-C-26 are used as the basic network to extract features and get the feature map in the last layer;
   In the third part, MLP is used as the coding network to encode the message, encode it with the feature extracted in the second part, and then update it as the feature vector;

2) Edge verification module
   In the fourth part, the CNN is used as the decoder to align the feature channels, and then the candidate regions are classified by the maximum pooling layer, and the classification information is mapped to the candidate regions by the full connection layer, and the final confidence score is output.

2.2. Feature extraction module

2.2.1. Feature candidate region generation based on improved faster-RCNN
In this paper, faster-RCNN backbone network is replaced from vgg16 to rest net50 to generate candidate regions. The specific process is as follows:

1) Input RGB image;
2) VGG-16, whose backbone is replaced by RestNet-50, is used to generate a certain number of candidate regions with corners and edges, and each corner is regarded as an 8 × 8 candidate region.
   As can be seen in Figure 2, in the improved model, the backbone network residual network block can naturally replace the original fast RCNN[5] and combine with the RPN network based on the candidate region. At the same time, the parameters of the whole network are calculated and shared based on the initial input image in the training process, which improves the overall detection performance of the network. The upper part of the graph is the backbone network rest net of feature
extraction and the RPN network generated by candidate regions, and the lower part is the interest pooling layer and the classification and border regression detection layer.

![Figure 2 Building target detection based on Faster-RCNN generates candidate regions](image)

2.2.2. Feature extraction of deep residual network with extended convolution

Based on the deep residual network, this network model constructs the feature extraction module by introducing the extended convolution. In the process of feature extraction of building remote sensing image, the large receptive field with extended convolution can be used to further extract the image detail features, which greatly improves the learning efficiency and learning quality of residual network feature learning, and improves the learning efficiency the feature extraction quality of building remote sensing image is analysed. In this paper, the first three layers of DRN-C-26 model are used as feature learning module[6], and a network structure suitable for building extraction is designed to extract feature information.

2.2.3. Feature fusion coding based on message passing network

In DR-MPNNs, the update function in the standard message passing network[7] is replaced by MLP, so that the standard form of feature vector update is to encode the message and mix it with the current feature.

\[
fv \leftarrow MLP(f_v; \sum_{w \in N(v)} MLP(f_v; f_w))
\] (1)

\(f_v\) represents the feature vector related to the node, \(N(v)\) represents the set of adjacent nodes, and ";" represents the feature series connection.

In DR-MPNNs, CNN is used to encode messages. Edge pixels need to be clipped for feature fusion operation at decoder end. So, before inputting data into decoder, we use bilinear difference method to zoom the picture to \(256 \times 256 \times 4\) and then input data into the model.

By max pooling, all information is kept in the message. The features are merged across all adjacent nodes to encode the message, and then the feature vector is updated by the encoder

\[
f_v \leftarrow CNN\left[ f_v; \max\text{-}Pool f_w \right]
\] (2)

2.3. Edge verification module

The edge verification module uses CNN structure as the decoder, samples the output of the feature extraction module, and finally outputs the corresponding confidence score of the feature node, which is used to judge whether the node is a real building edge.

In the decoding process, the low-resolution feature map is deconvoluted; Then, it decodes by feature fusion; Finally, a convolution operation is used at the end of the decoder to convert the number
of channels to 1, and then the ReLU activation function is used to output the confidence scores of candidate edges and corners.

2.4. Edge loss function
The task that indicates the authenticity of the predicted edge belongs to the classification task, which is divided into two categories. If you want to judge whether the feature of a predicted edge or corner is true or not, you can find out whether the feature of a predicted edge or corner is true. With the classifier, whether the output is a real edge is represented by or 1, where 1 means yes and 0 means no.

The loss function of the training phase adopts the two-classification cross entropy loss function, and the calculation formula is as follows:

$$\text{Loss} = -\sum H \log \hat{H} - \lambda (1 - H) \log(1 - \hat{H})$$

(3)

Where, $H$ it is expressed as real value and $\hat{H}$ as predicted value. $\lambda$ used to increase weight.

3. Results & Discussion

3.1. Experimental environment and data pre-processing

The experiment is carried out under 64-bit Ubuntu operating system, and the network is constructed and trained by using the python framework. The hardware configuration is Intel Xeon (R) platinum 8160, 250g memory, and the GPU is four NVIDIA P100.

The improved res net parameters are pre trained on ImageNet, and the pre trained parameters are used to initialize the network model. The training uses the Space Net data set and the WHU building data set.

In the process of training, the batch size is set to 1, and the initial learning rate is set to 1. When the test loss is not reduced in four cycles, the training rate is reduced to 0.8. When the test loss does not decrease in 20 cycles, the training process will be terminated.

3.2. The results were compared and analysed

In the training process, due to the limitation of GPU memory, this paper only uses 815 buildings for training. The time of one, two and three iterations is about 26, 34 and 40 hours. Due to the limitation of computer hardware, only three iterations are carried out. Figure 3 show the comparison of the results of this algorithm on the test data set.

Figure 3 Partial image of extraction results from WHU dataset
It can be seen from the extraction results of WHU building data set in Fig. 3 that this method can extract the geometric structure of the roof in the building remote sensing image very well. After three iterations, this method can extract the buildings in the image more accurately. From the building extraction results of the left one, it shows that there is almost no missing edge and corner, the output structure is close to the real annotation structure;

But in the extraction results in Figure 4, we can see that the recognition accuracy of the algorithm in this paper is slightly reduced due to the relatively complex building structure. Although the corners of the buildings in the image are extracted, the edges connected with them in the upper right corner of the image are not completely extracted. However, from the extraction results of most of the building data in the data set, the extraction results of this method are still good, and the correct geometric structure of the building can be extracted after three iterations.

3.3. Contrast experiment

The results of this paper are compared with those of PPG Net[8], L-CNN and UNet++, and the results are shown in Figure 5.

It can be seen from Figure 5 that the extraction results of PPG Net and the algorithm in this paper on the simple structure building (c) are more accurate and basically consistent with the real structure; L-CNN and our algorithm are consistent in the extraction of building (b); In the building extraction task shown in figure (d), the effect of UNet++ is the best of all the comparison algorithms. However, on the building image with simple structure, UNet++ produces some wrong nodes and too many redundant building edges. To sum up, in the visual comparison results, compared with other algorithms, the reconstruction effect of this paper is better than the extraction effect of previous algorithms, and the extraction of circular building structure is slightly lacking.
The algorithm in this paper is compared with the contrast algorithm according to the accuracy, recall and F1 index of corner, edge and region, and the results are shown in Table 1.

The data in Table 1 show that this paper achieves the best high-order (region) index in the no priori solution. For the angle and edge parameters, the effect of this algorithm is not always the best. Especially, on the edge data, the performance of L-CNN is slightly better than that of this algorithm, which is reflected in the recall rate and F1 value; In other results, this algorithm is better than the comparison algorithm.

|                | Precis. | Recall | F1-score | Precis. | Recall | F1-score | Precis. | Recall | F1-score |
|----------------|---------|--------|----------|---------|--------|----------|---------|--------|----------|
| PPG-Net        | 76.5    | 71.4   | 73.3     | 55.1    | 50.6   | 52.8     | 32.4    | 31.8   | 31.6     |
| UNet++         | 52.6    | 57.8   | 58.1     | 25.4    | 24.6   | 23.8     | 51.0    | 36.7   | 42.7     |
| L-CNN          | 66.7    | 88.2   | 75.2     | 51.0    | 71.2   | 59.4     | 25.9    | 41.5   | 34.9     |
| D-MPNs         | 78.9    | 80.2   | 80.1     | 55.6    | 60.7   | 58.7     | 53.1    | 57.6   | 56.2     |

4. Conclusions
Building target extraction has always been a hot and difficult problem in the field of remote sensing signal processing and pattern recognition. In this paper, aiming at the problems of low precision of large building extraction, imprecise building contour extraction and underutilization of high-resolution features in automatic building extraction of remote sensing image, a coding and decoding network based on message passing network is proposed. The convolution module for feature extraction introduces extended residual packet convolution to alleviate the problem of gradient disappearance of deep network. The multi-layer perceptron is used to connect the sub networks of different scales in parallel and continuously fuse the context information of the adjacent sub networks. The high-resolution features of the input image are maintained in the whole process, which ensures the effectiveness of the subsequent features. Experiments show that, compared with the existing full convolution neural network PPG Net, L-CNN and UNet++, the model has a significant improvement in the effect of building extraction.

Acknowledgments
The National Natural Science Foundation of China under Grant Nos. 61602228, 61572290; the Revitalization Talents Program of Liaoning Province under Grant No. XLYC1807266.

References
[1] Simonyan K, Zisserman A. Very deep convolutional networks for largescale image recognition [J]. arXiv preprint arXiv: 1409. 1556, 2014.
[2] David Acuna, Huan Ling, Amlan Kar, and Sanja Fidler. Efficient interactive annotation of segmentation datasets with polygon-rnn++. 2018.
[3] He K, Zhang X, Ren S, et al. Deep Residual Learning for Image Recognition[C]//IEEE Conference on Computer Vision and Pattern Recognition. Las Vegas: IEEE Computer Society, 2016: 770-778.
[4] Girshick. Fast R-CNN[C]. IEEE International Conference on Computer Vision. Santiago, Chile,2015.
[5] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with regionpropositional networks. In Advances in neural information processing systems, pages 91–99, 2015.
[6] Fisher Yu, Vladlen Koltun, and Thomas A. Funkhouser. Dilated residual networks. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 636–644, 2017.
[7] Daniel Ritchie, Anna Thomas, Pat Hanrahan, and Noah Goodman. Neurally-guided procedural models: Amortized inference for procedural graphics programs using neural networks. In Advances in neural information processing systems, pages 622–630, 2016.

[8] Ziheng Zhang, Zhengxin Li, Ning Bi, Jia Zheng, Jinlei Wang, Kun Huang, Weixin Luo, Yanyu Xu, and Shenghua Gao. PPGNet: Learning point-pair graph for line segment detection. arXiv preprint arXiv:1905.03415, 2019.