WEP Warwick Evans Publishing

International Journal of Computer Science and Information Technology

ISSN: 3005-9682 (Print), ISSN: 3005-7140 (Online) | Volume 2, Number 1, Year 2024
DOI: https://doi.org/10.62051/ijcsit.v2n1.24
Journal homepage: https://wepub.org/index.php/IJCSIT/index

Building Extraction from UAV Images Based on Attention Enhancement

Chao Fang*, Yunmao Liao

School of Geomatics and Urban Spatial Information, Beijing University of Civil Engineering and Architecture, Beijing, 102616, China.

ABSTRACT

Automatic extraction of buildings from remote sensing images using deep learning methods is crucial for urban and rural construction and management. However, the existing models are affected by the background noise and the complexity of building types during the extraction process, resulting in poor building extraction. In order to solve this problem, this paper proposes an improved model based on Mask R-CNN, which establishes a building instance extraction model for UAV imagery by adding a CBAM attention mechanism module to the backbone residual neural network ResNet. The traditional villages around Beijing are selected as the study area, and the comprehensive experimental results on the homemade building dataset show that the mAP value of the improved model is 86.4%, which is 2.1% higher than that of the original Mask R-CNN model, indicating that the improved model is more effective in building extraction.

KEYWORDS

Building Extraction; UAV Images; Mask R-CNN; CBAM

1. INTRODUCTION

High-resolution remote sensing images are capable of providing massive building information on the ground surface and have been widely used in such fields as urban and rural planning, change monitoring and disaster assessment. The breakthroughs in remote sensing technology in recent years have provided the availability of more and more high-resolution remote sensing images. Low-altitude remote sensing technology represented by unmanned aerial vehicles (UAVs) can provide centimeter-level ultra-high-resolution remote sensing images due to its advantages of high flexibility, high timeliness, low cost, and freedom from geographic and environmental constraints, making the spatial structure of the features on the images, the surface texture characteristics, and the edge feature information clearer. In view of the challenges of complex scene structure distribution and diverse building sizes and appearances in remote sensing images, how to accurately and efficiently extract buildings has been a hot research topic.

Traditional building extraction methods mostly utilize spectral, texture, geometric and other features of remote sensing images [1,2], however, due to many complex and variable factors such as the presence of a large number of shadows, spectral features and other factors affecting the accuracy of the building extraction, the traditional methods can not effectively, correctly, and completely extract the buildings, and it is difficult to satisfy the needs of different tasks and practical applications, so there are many scholars who utilize machine learning methods to extract the buildings. Such as object-oriented[3], support vector machine[4], random forest[5] etc. However, the extraction results of these machine learning methods often require a large amount of sample data, which can not be generalized, which makes the machine learning methods face a huge challenge in terms of reliability and generality

in accurate building extraction. There are also researches on the use of airborne LiDAR point cloud data to realize building extraction[6], as well as the combination of LiDAR point cloud data and highresolution imagery to realize building extraction by using machine learning methods[7,8], which have achieved certain results, but the high cost of obtaining airborne LiDAR point cloud data has impeded the effective realization of large-scale building extraction.

With the rapid development of deep learning technology in recent years, its application in the field of image recognition has been expanding. The deep learning model represented by convolutional neural network has powerful feature extraction capability, which can be well used to extract buildings in high-resolution remote sensing images. Based on this, many researchers are currently using semantic segmentation frameworks such as FCN[9], U-Net[10], DeepLabV3+[11], SegNet[12] to extract buildings. When confronted with buildings in close proximity, these semantic segmentation methods are unable to distinguish individual buildings, which is not conducive to the application and research of building instance extraction. Compared with image semantic segmentation, instance segmentation can recognize and segment all object instances in an image at the same time, which is very suitable for complex countryside building extraction and can be well applied to extract each building instance.

Among the existing instance segmentation methods, Mask R-CNN [13] has been proved to be a powerful and adaptable deep learning model in many fields, which is a classical and standard two-stage instance segmentation framework with a simple network structure including two parallel tasks of bounding box regression and pixel-level mask prediction for each object, and it is currently the most widely used detection-segmentation method. A large number of scholars have applied it to building extraction, He et al. [14] added path aggregation network and feature enhancement to Mask R-CNN, which can extract buildings efficiently and accurately on different building datasets. Lin et al. [15] achieved good results on homemade high resolution image building dataset by optimizing the FPN layer by adding more lateral connections as well as bottom-up and top-down paths and optimizing the NMS with Soft-NMS. Hu et al. [16] were able to accurately predict each building without adhesion by adding a layer of convolution operation after the feature extraction part of the feature map at each level, as well as adding a branch to the original mask prediction structure.

In summary, the purpose of this paper is to establish a better automatic building extraction model for UAV images using the well-structured Mask R-CNN instance segmentation model, by incorporating the CBAM attention module into ResNet, the backbone network of the Mask R-CNN, so that it can better extract the information of buildings, and comprehensive experiments based on the UAV-collected dataset of buildings in the surrounding areas of Beijing show that that compared with the original Mask R-CNN method, our method has higher extraction accuracy and better instance extraction effect.

2. METHODS

2.1. Mask R-CNN

Mask R-CNN network is an instance segmentation model proposed by He et al. It can be used as an algorithm for multi-tasks such as target detection and segmentation , Mask R-CNN adds a branch using Full Convolutional Network (FCN) to Faster R-CNN for predicting segmentation masks and simultaneously realizing classification, localization, and segmentation, it also combines Feature Pyramid Network (FPN) and Residual Network (ResNet) for feature extraction to better utilize the multi-scale information. The Mask R-CNN model can be divided into four main parts, which are Feature Extraction Network part, Region Proposition Network RPN part, Alignment Layer of Interest RoI Align part, and Header Network part as shown in Figure1. The Mask RCNN network adopts a two-stage structure, the first stage is to extract the RPN of the candidate target bounding box to generate the candidate object region where the target may be present, and the second stage is to extract the features from the candidate region of the RPN by utilizing the region of interest alignment RoI

Align, and to perform the category classification, bounding box regression, and binary mask generation.

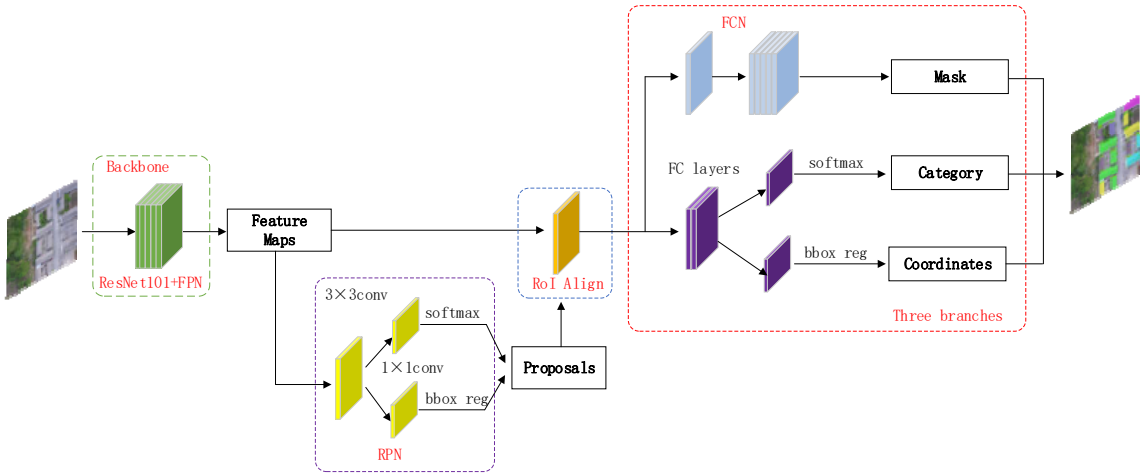


Figure 1. Mask R-CNN

The workflow based on Mask R-CNN is as follows: (1) the preprocessed image is fed into the residual network to extract features and generate a multi-scale feature map; (2) a fixed number of RoIs are assigned to each point on the feature map to obtain multiple RoIs; (3) the candidate RoIs are fed into the RPN network for foreground-background binarization classification and bounding regression, and at the same time, the non-maximal value suppression method to filter out the RoIs with low classification scores, and for the remaining RoIs in the RoI Align operation is performed to obtain the bounding box; (4) finally each pixel on these RoIs is predicted by classification and regression to generate a high-quality instance segmentation mask of the detected object.

2.2. Improved Mask R-CNN

Although Mask R-CNN is an advanced pixel-level instance segmentation model with high instance segmentation capability, its direct use for building extraction in this study area will have the problem of low extraction accuracy. Due to the interference of the complex background of remote sensing images and the large differences in the distribution, shape, size and texture of the buildings, it leads to incomplete building extraction and missed detection misdetection. The core task of the attention mechanism is to assign different weights to many pieces of information with certain strategies, and then filter out the most critical and important information for the task objectives. In the UAV image building extraction scale, the model needs to focus on the building's roof feature information, therefore, this paper proposes an improved Mask R-CNN model, in order to better extract the building features, the CBAM attention module is added to the feature extraction network ResNet, so as to focus more on the target features and suppress the complex background information in the feature extraction.

2.2.1. CBAM

CBAM [17] is a lightweight attention module (structure shown in Figure 2), which consists of two independent sub-modules, including the channel attention mechanism and the spatial attention mechanism, so that the feature maps will pass through the channel and the spatial attention modules sequentially to achieve double conditioning, which achieves better results in practical applications. It combines the attention with the input feature map to optimize the adaptive features. In the process of image feature extraction, the correlation between channel and space should be utilized to enhance the expression of target features, thus suppressing the expression of invalid features.

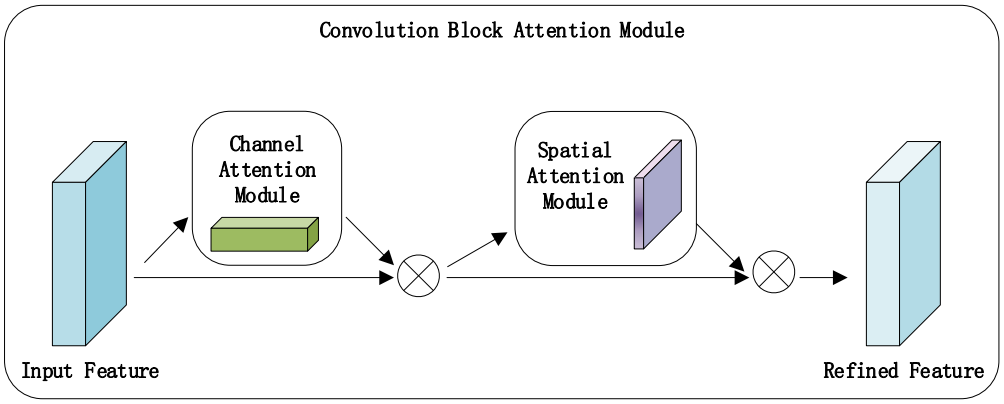


Figure 2. convolutional block attention module

The structure of the channel attention module is shown in Figure 3. In CAM, the input feature mapping F is extracted from the global features by two parallel branches of global maximum pooling and global average pooling, and then the number of channels is compressed to 1/r and then expanded back to the original number of channels by the multilayer perceptron (MLP) module, respectively, and the outputs of the two branches are summed up cell-by-cell, and the weighting coefficients MC of CAM are obtained by the Sigmoid activation function, and finally, the weighting coefficients MC is multiplied with the input feature mapping F to obtain the input feature F' of the SAM module. The calculation formula is shown in (1):

$$M_C(F) = \sigma\{W_1[W_0(F_{avg}^C)] + W_1[W_0(F_{max}^C)]\} \quad (1)$$

In equation (1), F is the input feature mapping, MLP is the multilayer perceptual layer, W0 and W1 are are the single fully connected layers in the MLP, avg denotes the average pooling operation, max denotes the global maximum pooling operation, F_{avg}^Cdenotes the channel description features after average pooling, and F_{max}^C denotes the channel description features after maximal pooling. the difference between the CAM and the SE is that a parallel maximal pooling layer is added. a parallel maximum pooling layer, and the extracted high-level features are more comprehensive and richer.

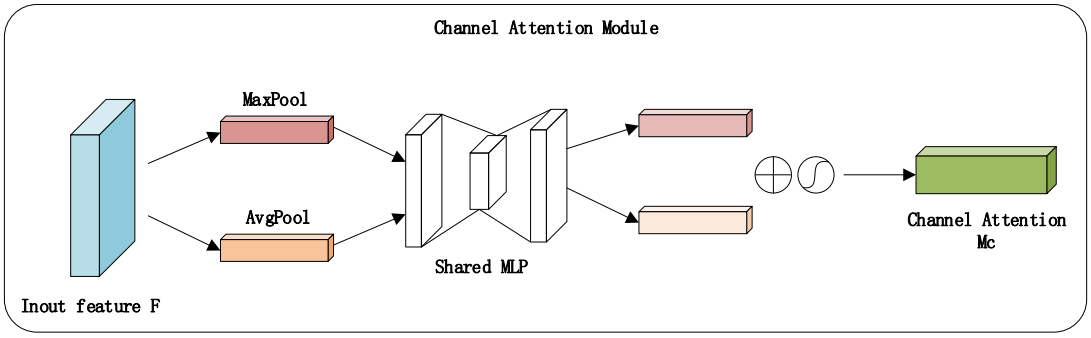


Figure 3. channel attention module

The structure of the spatial attention module is shown in Figure 4. SAM firstly performs the maximum pooling operation and average pooling operation on the channel for the input feature map F' with the size of H×W×C to get the two feature maps of H×W×1, and then splices these two feature maps based on the channel. It is changed into 1-channel feature map by 7×7 convolution, and after Sigmoid activation operation, the weight system MS of the feature map is obtained, and finally the final features are obtained by multiplying and scaling the weight coefficients MS with the feature map F'. The calculation formula is shown in (2):

$$M_S(F') = \sigma[f^{7\times7}(F_{avg}^S; F_{max}^S)] \quad (2)$$

In equation (2), σ is the sigmoid function, $f^{7\times7}$ is the convolution operation, and the convolution kernel size is 7×7.

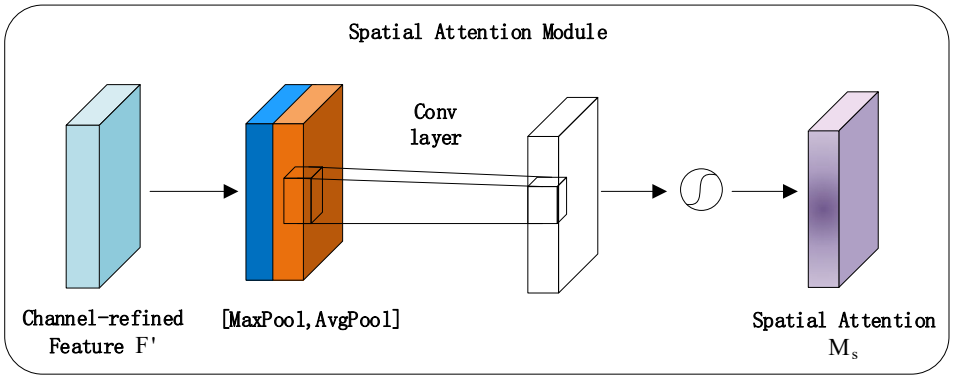


Figure 4. spatial attention module

2.2.2. Improved Model

The building extraction model for UAV images is established by introducing the CBAM attention mechanism module into the ResNet backbone network. The improved network module structure is shown in Figure 5. In this study, the CBAM attention mechanism is embedded in the residual module, so that the model can pay better attention to the most important feature channels and spatial locations in the image during the training and prediction process, and the attention of the feature extraction is focused on the information of the building at all times, which improves the quality of the feature extraction, and thus improves the accuracy of the building extraction model.

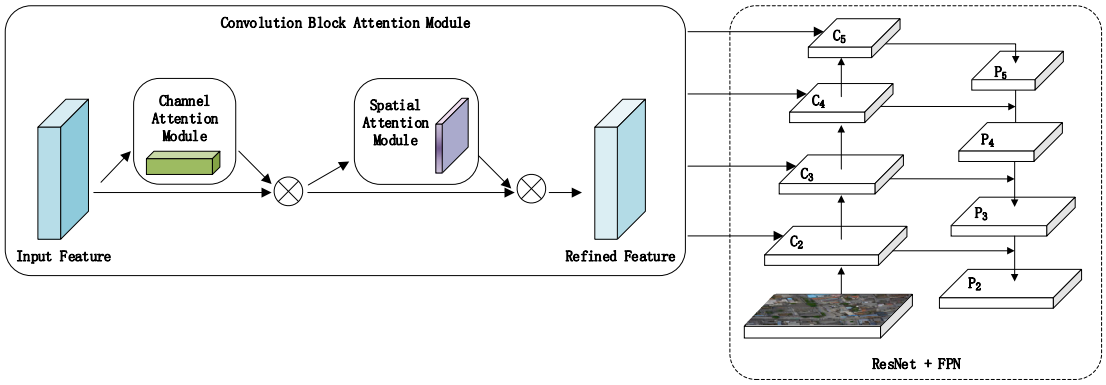


Figure 5. improved model

2.2.3. Evaluating Indicator

In this experiment, the average precision mAP is used as the evaluation index. mAP represents the average value of all categories of AP (Average Precision, AP), and AP is the average value of all categories of each category by the precision P (Precision) as the vertical coordinate, the recall R (Recall) as the horizontal coordinate, the P-R (Precision-Recall) curve is plotted, and the area under the curve is the AP. The formula is shown below:

$$\text{mAP} = \frac{\sum_{i=1}^{k} AP_i}{K} \tag{3}$$

$$\text{AP} = \int_0^1 p(r)dr \tag{4}$$

Where k denotes the number of all categories, p denotes Precision, r denotes Recall, and both are calculated as follows:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{5}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{6}$$

In the formula, where TP indicates that the original is a positive sample and the model predicts a positive class, FP indicates that the original is a negative sample and the model predicts a positive class, TN indicates that the original is a positive sample and the model predicts a negative class, and FN indicates that the original is a negative sample and the model predicts a negative class.

3. EXPERIMENTS AND ANALYSIS

3.1. Dataset

In order to test the performance of the improved model, we obtained ultra-high resolution remote sensing images of the villages around Beijing by UAV, and the selected study area contains relatively abundant rural buildings. Since the acquired single image is too large to be directly inputted into the network for training, we cropped the image to a size of 1024 pixels × 1024 pixels, and annotated the image using labelme software. The labelme software was used to label the images, and a corresponding .json file was generated for each image, and the labeled images were converted into the COCO format data required by the network for the experiments. Finally, the data samples are divided into training, validation and test sets based on the ratio of 6:2:2. In order to ensure the credibility of the test accuracy, there is no duplication of the image samples in the test set and validation set.

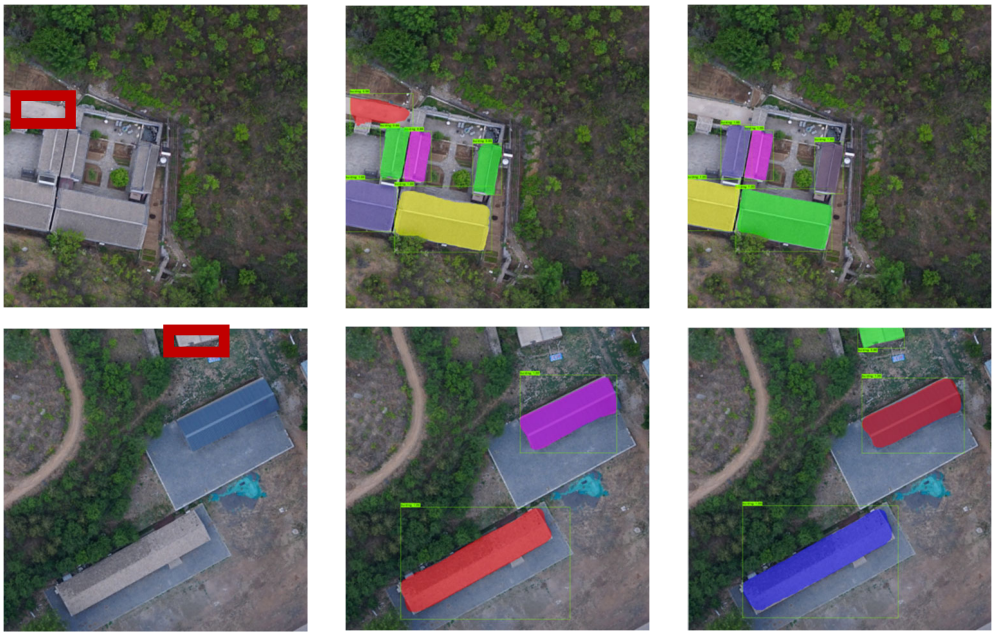
3.2. Implementation Settings

The environment of this experiment is Windows 10 operating system by Python programming language and PyTorch deep learning framework. And the model was trained and tested under the hardware environment of NVIDIA GeForce RTX 3060 12GB. In order to accelerate the model convergence, the pre-training weights of the COCO 2017 dataset are initialized to the model, and the training of the homemade building dataset is continued on the basis of this pre-trained model.

3.3. Results Analysis

In order to verify the performance of the model, the qualitative evaluation results of the Mask R-CNN and the improved model on the homemade building dataset are compared, respectively, and two experimental results are selected from the results for illustration, as shown in Figure 6. The first column in the figure is the original image, the second column is the Mask R-CNN extraction result, and the third column is the improved Mask R-CNN extraction result. We mark the extraction failures (misdetection and omission) with red boxes. From the first test result, it can be seen that the Mask R-CNN is prone to misdetect the concrete floor, which is similar to the feature information of the roof, as a building when the background noise interferes, indicating that the addition of the CBAM attention module can suppress the background interference very well so as to extract the building correctly. The second test result shows that Mask R-CNN easily ignores some edge building information when the building type is complex, which leads to missed detection, indicating that adding the CBAM attention module pays more attention to the target feature information, which is also helpful for the detection of edge targets. By comparing the experimental extraction results, we find that after adding the CBAM attention module, the model can pay more attention to the features

of the buildings in the samples as well as suppress the background interference to achieve better extraction results.



Original Image Mask R-CNN Improved Mask R-CNN

Figure 6. comparison of experimental results

In order to quantitatively analyze the building extraction accuracy of this paper's method, the experimental results of the two models Mask R-CNN (ResNet + FPN) and the improved Mask R-CNN (ResNet + FPN + CBAM) on the homemade building dataset are given in Table 1, from which it can be seen that the average accuracy of the improved model is 86.4% at an IoU threshold of 0.5 , which is improved by 2.1% compared with the original Mask R-CNN, indicating that the improved Mask R-CNN model in this paper has a significant advantage in building instance extraction, which shows that the model proposed in this paper is feasible.

Table 1. Precision Comparison

Methods	Backbone	mAP
Mask R-CNN	ResNet + FPN	84.3
Ours	ResNet + FPN + CBAM	86.4

4. CONCLUSIONS

In this paper, an improved fully automated building instance extraction model for UAV remote sensing images based on Mask R-CNN and CBAM attention mechanism is proposed. In order to improve the accuracy of building extraction, a CBAM attention module is added to the ResNet residual module of the Mask RCNN model as a way to help the network focus on the important feature information while suppressing the noise during the training process. Comprehensive experiments on homemade building data show that the improved model is able to extract building instances effectively in the presence of complex backgrounds and the influence of the buildings themselves. In addition, our improved Mask R-CNN model can achieve better results compared to the original Mask R-CNN method, further indicating that the improved model has potential applications in the field of urban and rural planning. Although the improved model has achieved satisfactory results in building instance extraction, there are still some shortcomings, and there are still cases of omission or misdetection. In our future work, the above problems can be the focus of

future research, and there is also a need to study and develop a more accurate and faster building instance extraction model.

ACKNOWLEDGEMENTS

We thank all authors for their contributions to this paper and we would like to thank the anonymous reviewers for their constructive and valuable suggestions.

REFERENCES

[1] Y.Z. Liu, B.M. Zhang, J.F. Xu, K. Hou and X. Zhou. Building Extraction from High Resolution Remote Sensing Imagery with Multi-feature and Multi-scale[J]. Bulletin of Surveying and Mapping, 2017, (12):53-57.

[2] F.F. Zhu, Z.Q. Li, S.W. Yang and M. Yang. Urban building object-oriented extraction method based on feature component[J]. Science of Surveying and Mapping,2020,45(1):84-91.

[3] X. Li, J.N. Cao. Obiect-oriented Building Extraction Based on Feature Optimization[J]. Computer Systems & Applications,2022,31(09):360-367.

[4] Q. Liu, X.Y. Hu, X.T. Li and X.L. Qin. Building Recognition Method in Forest Districts Combining the Pixel-level and Object-level[J]. Remote Sensing Technology and Application, 2021,36(06):1350-1357.

[5] C. Fan, H. Jiang. Precise Extraction of Buildings' lnformation in WorldView2 lmages Based on Random Forests[J]. Geospatial Information,2016,14(01):58-62+5.

[6] Z.H. Pan, J. Jin, S.L. Chen, T. Su, S. Liu and N.P. Gao. Automatic Building Extraction Based on Airborne LiDAR Point Cloud Data[J]. Geospatial Information,2022,20(05):57-59+101.

[7] L.Y. Chen, H. Lin, J.H. Wu. Building extraction based on random forest and superpixel segmentation[J]. Bulletin of Surveying and Mapping,2021(02):49-53.

[8] Y.Q. Chen, Y.F. Luo, J.S. Long and X.Y. Guo. Contextual Extraction of Urban Buildings from Airborne LiDAR Point Cloud[J]. Journal of Spatio-temporal Information,2019,26(05): 58-63.

[9] Y.M. Zhang, H.T. Fu, H.Y. Sun, R. Zhang, L.C. Chen and L.H. Pan. A Building Recognition Method for Multispectral lmage Based on lmproved FCN[J]. Computer Engineering,2019, 45(01):239-245.

[10] S. Jin, M. Guan, Y.C. Bian and S.L. Wang. Building Extraction from Remote Sensing lmages Based on lmproved U-Net[J]. Laser & Optoelectronics Progress,2023,60(04):59-65.

[11] Y.J. Ren, X.S. Ge.An road synthesis extraction method of remote sensing image based on improved DeeplabV3+ network[J].Bulletin of Surveying and Mapping, 2022(06):55-61.

[12] Y. Lin, Q.H. Zhao, Z.Y Shen and Y. Li. A building segmentation method for remote sensing with improved SegNet and transfer learning[J]. Science of Surveying and Mapping,2022, 47(06):78-89.

[13] K.M. He, G. Gkioxari, P. Dollár, R. Girshick. Mask R-CNN. In Proceedings of the IEEE International Conference on Computer Vision;2017; pp. 2961–2969.

[14] D.Y. He, W.Z. Shi, Z.B. Lin, X.X. Qiao, Y.X. Liu and Y.H. Lin. Building Extraction from Remote Sensing Image Based on Improved MaskR-CNN[J]. Computer Systems & Applications,2020,29(9):156−163.

[15] N. Lin, T. Huang, P.L. Sun and Y.Y. Wang. Building Extraction of High-resolution Remote Sensing lmagery on Optimized Mask-RCNN[J]. Remote Sensing Information,2022,37(3):1-6.

[16] M.J. Hu, D.J. Feng, Q. Li. Automatic extraction of buildings based on instance segmentation model[J]. Bulletin of Surveying and Mapping,2020,(4):16-20.

[17] S.Woo, J. Park, J.Y. Lee and I.S. Kweon. CBAM: Convolutional Block Attention Module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany,8–14September 2018, pp,3–19.