Object Recognition of Urban Buildings in High Spatial Resolution Remote Sensing Imagery

Yuan Liu*

1 School of Geomatics and Urban Spatial Information, Beijing University of Civil Engineering and Architecture, Beijing, 100044, China
*Corresponding author’s e-mail: 2108160218007@stu.bucea.edu.cn

Abstract. High-resolution remote sensing images can finely express rich surface information. Using the macroscopic and spatial-temporal full coverage advantages of high-resolution remote sensing images for urban building objectification recognition is a current research hot spot in the field of remote sensing image analysis and application. However, the current research still lacks effective technical means to convert surface elements quickly and accurately from remote sensing image space to geographic information space. In this paper, the complementary advantages between image processing and deep learning are combined to research target-level extraction of high-resolution remote sensing urban buildings based on the building element information data. The experimental results achieved the precision of 0.9481 and the recall of 0.9733, indicating that the method proposed in this paper can be applied to the effective extraction of urban-level buildings, which expands the theoretical basis of the object recognition method of remote sensing thematic features based on the idea of object image analysis.

1. Introduction
The timely acquisition and updating of urban spatial geographic information data is an important basic work to promote urbanization and accelerate the construction of smart cities. The application of high-resolution remote sensing image data to the recognition of urban thematic feature targets can greatly improve the accuracy and output efficiency of remote sensing information data extraction applications [1]. The earliest developed remote sensing information extraction methods are mainly designed around the algorithm of local spatial feature calculation with "pixel" as the basic unit. This type of approach is difficult to integrate domain knowledge, expert experience, and geological models, resulting in insurmountable limitations in the extraction effect and application of the method [2]. With the gradual improvement of spatial resolution and the higher requirements of remote sensing applications for classification and mapping, the Object-Based Image Analysis (OBIA) method for remote sensing interpretation has been gradually developed based on image segmentation in computer vision. Based on this, GIS can be applied for mapping and analysis as well as further intelligent computing at the semantic level such as knowledge mining, representation, and inference that is more in line with human thinking [3]-[4].

All the pixel-based, object-based methods mentioned above are based on shallow architectures and manual feature descriptors that cannot capture the fine features of complex land use images for generalization. In recent years, the field of artificial intelligence has developed rapidly, especially the intensive application of deep learning algorithms in the field of image analysis has given us an effective way to learn the high-level features of data. In the field of computer vision, the CNN-based
network architecture has gradually formed a type of neural network applied to the target level, which finds an optimal bounding box through iterative regression rather than simple label classification. The position and quantity of the objects to be detected are further obtained from the four-corner coordinates of the bounding box. There are widely developed region-based models such as R-CNN [5], Fast-R-CNN [6], and Faster R-CNN [7]. Mask R-CNN [8] is the most representative instance segmentation algorithm in the field of computer vision, which builds on Faster R-CNN and extends a function to predict the mask of different objects. Mask R-CNN achieves target recognition and segmentation of foreground targets within the bounding box using the fully convolutional network.

The existing remote sensing image analysis and massive data processing technologies can hardly meet the requirements of current remote sensing big data applications, and there are still obvious limitations: (1) The recognition tasks for different elements have a different emphasis on feature selection and require sufficient industry expertise and rich prior knowledge, which leads to insufficient universality of the method. (2) Objectified recognition of targets in the high-resolution image is mainly realized by multiscale segmentation technology, but it is still difficult to complete the task of automatic information extraction for complex feature targets (e.g., buildings), and there is no systematic and efficient method and technical system to give an engineering solution. To address the above problems, this paper proposed a high-level feature enhancement method to achieve high accuracy object recognition of urban buildings from high-resolution remote sensing images, which expands the theoretical basis for object recognition of remote sensing thematic features based on the idea of object image analysis.

2. Materials and Methods

2.1. Architecture Overview

The recognition principle of the Mask R-CNN algorithm can be summarized in three parts, as shown in Figure 1. Firstly, the backbone network is used to extract features from the input image by convolution and pooling operations, and the convolution layer can extract local features from the image while maintaining the spatial continuity of the image. Secondly, the Region Proposal Network (RPN) generates the bounding boxes of the building candidates. Then the RoIAlign layer corrects the deviation of the feature map regions using the bilinear interpolation method to align the extracted features accurately with the input image features. Finally, there is the parallel processing of prediction classes, border offset refinement, and output binary mask in the entire architecture to complete the classification of buildings, boundary regression, and instance segmentation. The overall architecture of the network is shown in Figure 1.

![Figure 1. Mask R-CNN network architecture.](image)

This paper first transfers the feature weights trained in the COCO dataset to Mask R-CNN and then combines the spatial-spectral features of the high-resolution remote sensing training images, the relationship between adjacent pixels, and other feature factors to participate in the final decision classification. Since the bottom network can only learn low-level features such as spectra and edges, the middle network can learn more complex features such as textures and structures, until the top
network can learn more abstract semantic features. Therefore, we enhanced the network structure of the backbone network at the middle and high levels to improve the extraction of deep semantic information of architectural targets. After experimental comparison, ResNet116 is selected as the backbone network of the Mask R-CNN algorithm, so that the improved network can meet the requirements of WorldView-3 image building object recognition. The network structure of the enhanced features is shown in Table 1.

| Table 1. The network structure of ResNet-116. |
|--------------------------------------------|
| Conv1 | Conv2_x | Conv3_x | Conv4_x | Conv5_x |
| [7 × 7, 64] | 1 × 1,64 | 1 × 1,128 | 1 × 1,256 | 1 × 1,512 |
| 3 × 3,64 × 3 | 3 × 3,128 × 8 | 1 × 1,1024 | 1 × 1,2048 |

2.2. Evaluation Metrics
To quantify the effect of feature recognition, we use four evaluation metrics, namely: Intersection over Union (IoU), precision, recall, and F1-score (F1). F1-score considers the recall of the model classification results in addition to the precision of the integrated classification model. They are calculated as follows.

\[ IoU = \frac{TP}{TP + FP + FN} \]  
\[ Precision = \frac{TP}{TP + FP} \]  
\[ Recall = \frac{TP}{TP + FN} \]  
\[ F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \]

where the true positive (TP) indicates that the building was correctly recognized, the false positive (FP) indicates that the background was incorrectly recognized as a building, and the false negative (FN) indicates that the building was incorrectly detected as background.

3. Experiments and Results

3.1. Experiment Data
In this study, we selected the sub-center area of Tongzhou District, Beijing as the experimental area. The experimental data are high-resolution remote sensing images acquired by WorldView-3 in 2017, including a panchromatic image with spatial resolution of 0.3m and a multispectral image with spatial resolution of 1.26m. The experimental data are pre-processed by the Gram-Schmidt Pan Sharpening image fusion method, which can take full advantage of the high spatial resolution information of the panchromatic image and the rich spectral features of the multispectral image. Figure 2 shows the high-resolution remote sensing image of the experimental area after data fusion. The spatial resolution is 0.3m and contains three bands of red, green, and blue.

After image pre-processing and sample feature extraction, a high-resolution remote sensing image sample dataset is constructed for training and testing of the building recognition model. The composition of the sample dataset includes 500 training samples and 150 test samples.

3.2. Implementation Details
The model is trained on the constructed multi-feature fusion sample dataset with semantic and spatial-spectral features. Besides, the hyperparameters are evaluated and tuned experimentally, so that the final output model can achieve a comparatively excellent recognition effect and accuracy. The main contents of the model training include the input of the constructed training dataset, the tuning of hyperparameters, and the training output of the final model. The process of network model training is as follows.
The feature weights trained in the COCO dataset are first migrated to Mask R-CNN, and the network learning is made self-organized and adaptively enhanced by adding the migratory learning mechanism of the sample library. Then the image data and sample labels in the training dataset are input to the constructed network for training, and feature learning is performed by combining the spatial-spectral features of high-resolution remote sensing training images, and the relationship between adjacent pixels. Finally, the hyperparameters are evaluated and adjusted during the training process for network optimization. The hyperparameters of the experimental model are detailed in Table 2.

![Image](image.png)

Figure 2. High-resolution remote sensing image of Tongzhou New Town. The area is 170 square kilometres. The image is obtained by fusing panchromatic and multispectral images. The spatial resolution is 0.3m, including three bands of red, green, and blue.

| Parameter                  | Values | Parameter                  | Values |
|----------------------------|--------|----------------------------|--------|
| BACKBONE                  | ResNet116-FPN | GPU_COUNT | 1 |
| BACKBONE_STRIDES          | [4, 8, 16, 32, 64] | IMAGES_PER_GPU | 1 |
| RPN_ANCHOR_RATIO           | [0.5, 1, 2] | NUM_CLASSES | 2 |
| RPN_ANCHOR_SCALES         | [32, 64, 128, 256, 512] | BATCH_SIZE | 1 |
| RPN_NMS_THRESHOLD         | 0.7 | EPOCHS | 30 |
| TRAIN_ROIS_PER_IMAGE      | 150 | LEARNING_RATE | 0.0001 |
| DETECTION_MAX_INSTANCES   | 300 | LEARNING MOMENTUM | 0.9 |
| MAX_GT_INSTANCES          | 300 | WEIGHT_DECAY | 0.0001 |

3.3. Results and Discussion

As shown in Figure 3(1) - (6), we selected six representative experimental regions in the study area for building recognition experiments. To further verify the effectiveness of the method, the improved Mask R-CNN method is experimentally compared with five typical object-oriented classification methods based on hand-designed features in this paper. Several methods in the experiments use the same source data for recognizing buildings in high-resolution remote sensing images. The accuracy of each method was evaluated quantitatively for the test data based on IoU, precision, recall, and F1-Score, as shown in Table 3. The experimental results are visualized and compared in Figure 3. From left to right, the first two columns show the original image and ground truth respectively, and the last six columns show the recognition results of the Mask R-CNN method, Decision Tree, SVM, KNN, Bayes, and Random Forest methods respectively.

Precision is a measure of how accurately the classifier predicts positive samples, and recall measures the ability of the classifier to cover positive samples. From Table 3, it is known that the precision and recall of all methods exceeded 80%, and the recall of Mask R-CNN, Decision Tree,
Bayes, and Random Forest methods exceeded 90%, indicating that these classifiers have superior ability to cover positive samples. However, from a qualitative point of view, the results of each experimental region in Figure 3 show that the recognition results of all five object-oriented methods have significant noise phenomena. The experimental region (5) represents a dense arrangement of small buildings, and it is inferred from the recognition results that SVM, KNN, Bayes, and Random Forest methods have poor ability to discriminate the background. In general, the recognition results of the improved Mask R-CNN method in this paper show no noise, and the integrity of the target buildings is well maintained. Besides, the adjacent target buildings can be segmented independently. The accuracy evaluation indexes of the improved Mask R-CNN method all reach above 90%, which illustrates the effectiveness and accuracy of the method in objectified building recognition for high-resolution remote sensing.

Table 3. The building recognition accuracy of different methods.

| Ground Truth a | Prediction b | Match c | Precision | Recall | F1-Score | IoU  |
|----------------|--------------|---------|-----------|--------|----------|------|
| Decision Tree  | 150          | 174     | 143       | 0.8218 | 0.9533   | 0.8827| 0.7901|
| SVM            | 150          | 160     | 132       | 0.8250 | 0.8800   | 0.8516| 0.7416|
| KNN            | 150          | 156     | 126       | 0.8077 | 0.8400   | 0.8235| 0.7000|
| Bayes          | 150          | 179     | 147       | 0.8212 | 0.9800   | 0.9068| 0.8294|
| Random Forest  | 150          | 161     | 141       | 0.8758 | 0.9400   | 0.9068| 0.8294|
| Mask R-CNN     | 150          | 154     | 146       | 0.9481 | 0.9733   | 0.9605| 0.9241|

a The number of actual building targets on the ground.
b The number of buildings predicted by the model.
c The number of buildings correctly predicted by the model.

Figure 3. Results of building recognition by different methods in the experimental areas. Each row from top to bottom in the figure represents a different experimental area. From left to right each column indicates the original image, ground truth, the recognition results of the Mask R-CNN, Decision Tree, SVM, KNN, Bayes, and Random Forest methods, respectively.
4. Conclusions
With the development of high-resolution earth observation technology and the deepening of engineering application of deep learning, there is unprecedented urgent demand on how to mine urban building elements from high-resolution remote sensing image data for analytical applications through intelligent computing technology.

In this paper, we used object-based image analysis techniques and deep neural networks to construct the geometry of semantic image targets and gradually form a parcel analysis unit, which is fully integrated with the machine learning self-organizing optimization mechanism based on Mask R-CNN research to achieve object-based recognition of urban buildings in high-resolution remote sensing images. The experimental results show that the precision and recall of the improved Mask R-CNN method reach 94.81% and 97.33%, respectively. This study aims at the core problems in the current research and engineering application of intelligent extraction of remote sensing information and is an effort to explore the object-based recognition of urban buildings under the current conditions of remote sensing big data.

Acknowledgments
This research is funded by the National Natural Science Foundation of China (No. 41301489), Beijing Natural Science Foundation (No. 4192018, No. 4142013), Outstanding Youth Teacher Program of Beijing Municipal Education Commission (No. YETP1647, No. 21147518608), Outstanding Youth Researcher Program of Beijing University of Civil Engineering and Architecture (No. 21082716012), the Fundamental Research Funds for Beijing Universities (No. X18282, No. X20099, No. X20077), High-resolution remote sensing building information extraction project in Vietnam (No. H20173) and the BUCEA Post Graduate Innovation Project (No. PG2020078).

References
[1] Li, D., Zhang, L., Xia, G. (2014) Automatic Analysis and Mining of Remote Sensing Big Data. Acta Geodaetica et Cartographica Sinica, 43: 1211-1216.
[2] Luo, J., Wu, T., Xia, L. (2016) The Theory and Calculation of Spatial-spectral Cognition of Remote Sensing. Journal of Geo-information Science, 18: 578-589.
[3] Blaschke, T. (2010) Object based image analysis for remote sensing. ISPRS Journal of Photogrammetry and Remote Sensing, 65: 2-16.
[4] Blaschke, T., Hay, G.J., Kelly, M., Lang, S., Hofmann, P., Addink, E., Feitosa, R.Q., Van der Meer, F., Van der Werff, H., Van Coillie, F., Tiede, D. (2014) Geographic Object-Based Image Analysis – Towards a new paradigm. ISPRS Journal of Photogrammetry and Remote Sensing, 87: 180-191.
[5] Girshick, R., Donahue, J., Darrell, T., Malik, J. (2014) Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. In: 2014 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Columbus. pp. 580-587.
[6] Girshick, R. (2015) Fast R-CNN. In: 2015 IEEE International Conference on Computer Vision (ICCV). Santiago. pp. 1440-1448.
[7] Ren, S., He, K., Girshick, R., Sun, J. (2017) Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 39: 1137-1149.
[8] He, K., Gkioxari, G., Dollar, P., Girshick, R. (2017) Mask R-CNN. In: 2017 IEEE International Conference on Computer Vision (ICCV). Venice. pp. 2980-2988.