Scientific Research Publishing

International Journal of Geosciences, 2019, 10, 884-905
https://www.scirp.org/journal/ijg
ISSN Online: 2156-8367
ISSN Print: 2156-8359

# Instance Segmentation of Outdoor Sports Ground from High Spatial Resolution Remote Sensing Imagery Using the Improved Mask R-CNN

**Yijia Liu[1], Jianhua Liu[1,2*], Heng Pu[1], Yuan Liu[1], Shiran Song[1]**

[1]School of Geomatics and Urban Spatial Information, Beijing University of Civil Engineering and Architecture, Beijing, China
[2]Key Laboratory for Urban Geomatics of National Administration of Surveying, Mapping and Geoinformation, Beijing, China
Email: *liujianhua@bucea.edu.cn

**How to cite this paper:** Liu, Y.J., Liu, J.H., Pu, H., Liu, Y. and Song, S.R. (2019) Instance Segmentation of Outdoor Sports Ground from High Spatial Resolution Remote Sensing Imagery Using the Improved Mask R-CNN. *International Journal of Geosciences*, **10**, 884-905.
https://doi.org/10.4236/ijg.2019.1010050

**Received:** August 26, 2019
**Accepted:** October 11, 2019
**Published:** October 14, 2019

## Abstract

Aiming at the land cover (features) recognition of outdoor sports venues (football field, basketball court, tennis court and baseball field), this paper proposed a set of object recognition methods and technical flow based on Mask R-CNN. Firstly, through the preprocessing of high spatial resolution remote sensing imagery (HSRRSI) and collecting the artificial samples of outdoor sports venues, the training data set required for object recognition of land cover features was constructed. Secondly, the Mask R-CNN was used as the basic training model to be adapted to cope with outdoor sports venues. Thirdly, the recognition results were compared with the four object-oriented machine learning classification methods in eCognition®. The experiment results of effectiveness verification show that the Mask R-CNN is superior to traditional methods not only in technical procedures but also in outdoor sports venues (football field, basketball court, tennis court and baseball field) recognition results, and it achieves the precision of 0.8927, a recall of 0.9356 and an average precision of 0.9235. Finally, from the aspect of practical engineering application, using and validating the well-trained model, an empirical application experiment was performed on the HSRRSI of Xicheng and Daxing District of Beijing respectively, and the generalization ability of the trained model of Mask R-CNN was thoroughly evaluated.

## Keywords

Instance Recognition, Urban Remote Sensing, High Spatial Resolution Remote Sensing Imagery, Deep Learning, Mask R-CNN

## 1. Introduction

With the rapid development of remote sensing science and technology, and the improvement of resolution of remote sensors, we have entered the era of sub-meters. The detailed information of spectrum, geometry and texture of outdoor stadiums can be reflected in the high spatial resolution remote sensing imagery (HSRRSI) clearly, which provides a useful data source for land features detection and identification of outdoor stadiums. This paper makes full use of the advantages of HSRRSI and integrates with automatic feature learning and target detection technology of deep convolutional neural networks to explore a new method that is more suitable for the recognition of outdoor sports venues: football field, basketball court, tennis court and baseball field.

In the remote sensing research field, high-precision identification of features in HSRRSI has always been an important research topic. In recent years, many researchers have introduced deep learning techniques to solve this problem [1]. Among the candidate region-based target detection and recognition algorithms, R-CNN, Fast R-CNN, Faster R-CNN and Mask R-CNN are representative.

The R-CNN algorithm was first proposed by Ross Girshick *et al.* [2]. R-CNN follows the traditional target detection method, and uses the four steps of generating candidate frames, extracting features in each frame, performing image classification, and outputting non-maximum suppression results. The difference is in the step of feature extraction, where R-CNN replaces the traditional feature extraction method with a deep convolution network. However, there are a large number of repetitive operations when feature extraction is performed for each candidate frame, which limits the speed of the algorithm, reduces training efficiency and requires a large amount of disk space.

In 2015, the Fast R-CNN algorithm [3] was designed. The algorithm refers to the ideas of R-CNN and SPPNet (Spatial Pyramid Pooling Convolutional Networks) [4] in the implementation process. However, SPPNet uses the Support Vector Machine (SVM) for classification, while Fast R-CNN is directly implemented using the Full Connection Layer. There are two outputs in the fully connected layer of Fast R-CNN, one for classification and the other for candidate box regression. This kind of thinking makes the whole training process more compact and greatly improves the training efficiency. Compared with R-CNN, the training speed is increased by 9 times and the target detection speed is increased by 200 times.

After the publication of Fast R-CNN, it has been found that most time consuming procedure is not the computational neural network classification, but the selective search, which provides direction for subsequent research. In 2017, the Faster R-CNN algorithm [5] was designed. Compared with Fast R-CNN, Faster R-CNN replaced selective search with RPN (Region Proposal Network). The algorithm speed and the accuracy were greatly improved [6] [7].

In Faster R-CNN, the algorithm uses the rounding operation in the calculation process. Although it has little effect on the RoI classification, it is detrimental to the pixel-level target detection and recognition accuracy. This operation makes each RoI unable to be aligned. Therefore, in 2017, the Mask R-CNN algorithm [8] was designed to improve the RoI Pooling layer and proposed RoI Align. The use of bi-linear interpolation in Mask R-CNN allows each RoI to better align the RoI on the original image, enabling accurate pixel-level target segmentation [9].

After scholars' unremitting research, deep learning in the field of target recognition shows advantages, especially for HSRRSI, which can fully extract remote sensing image features [10]. However, few studies are devoted to the use of deep learning methods for outdoor sports ground instance recognition. This paper uses the Mask R-CNN deep learning model to develop the outdoor sports ground identification method and provides a set of research ideas for reference. This paper is arranged as follows: Section 1 reviews the related work. Section 2 describes the proposed method of land cover features recognition, including image pre-processing, feature extraction, network training, comparative research and application research. Section 3 introduces the data of this paper, explains the experimental results based on the Mask R-CNN recognition method, and compared them with the results of four object-oriented methods. Then the recognition results of empirical application experiments are presented. Finally, the conclusion of our study is summarized in Section 4.

## 2. Method

### 2.1. Data Pre-Processing

The data pre-processing includes four steps: image fusion, framing, linear stretching and image filtering. While enhancing the features of the target features, the image quality is guaranteed to improve the recognition accuracy [11].

#### 2.1.1. Basic Image Pre-Processing

The main function of image fusion is to make the processed image integrate the advantages of high resolution of panchromatic image and rich spectral features of multispectral image. The HSRRSI of WorldView-3 is used in this paper, which includes panchromatic image with spatial resolution of 0.3 m and multispectral image of 1.24 m. The two images are fused by Gram-Schmidt Pan Sharpening, and finally a true color HSRRSI of 0.3 m is obtained. Compared with the original image, the spatial information and spectral information of the fused image have been greatly improved, and a better visual effect is obtained.

Because the remote sensing image's image size is relatively large and information is complex, in order to reduce the interference of other features, and considering the load capacity, training efficiency and image fidelity of the neural network model, this experiment resizes the three parts of the image, by dividing them into multiple small images of 500 * 500.

The image clipped to 500 * 500 is linearly stretched, with enhanced contrast, and more prominent, spectral information which is beneficial to improve the accuracy of subsequent object recognition.

Then the image is subjected to Laplacian filtering. The main feature of the target feature in this experiment is the internal texture information. It can be seen that the filtered image texture information is more prominent and the sample quality is improved.

#### 2.1.2. Sample Dataset Construction

This experiment uses the open source tool Labelme® [12] to manually extract the target feature samples. Labelme® is an image annotation tool that can mark any shape on the image and assign its corresponding category label. The manual process uses the technical process is shown in Figure 1. It allows multiple image objects on a single image, manually draws each target feature along the target feature contour, and then labels the semantic information of its actual object category to generate the corresponding Json file. Finally, by parsing the properties and mask information of the feature generated by the Json file, there is a one-to-one relationship between each image and the file.

### 2.2. Deep Convolutional Neural Networks of Recognition

The overall framework of Mask R-CNN is depicted in Figure 2.

The model is briefly described as follows:

Mask R-CNN consists mainly of three phases. In the first stage, the convolutional network (Residual Neural Network and Feature Pyramid Network) is used to extract the features of the outdoor sports venues; in the second stage, the candidate target bounding box containing the interested venues is extracted by the

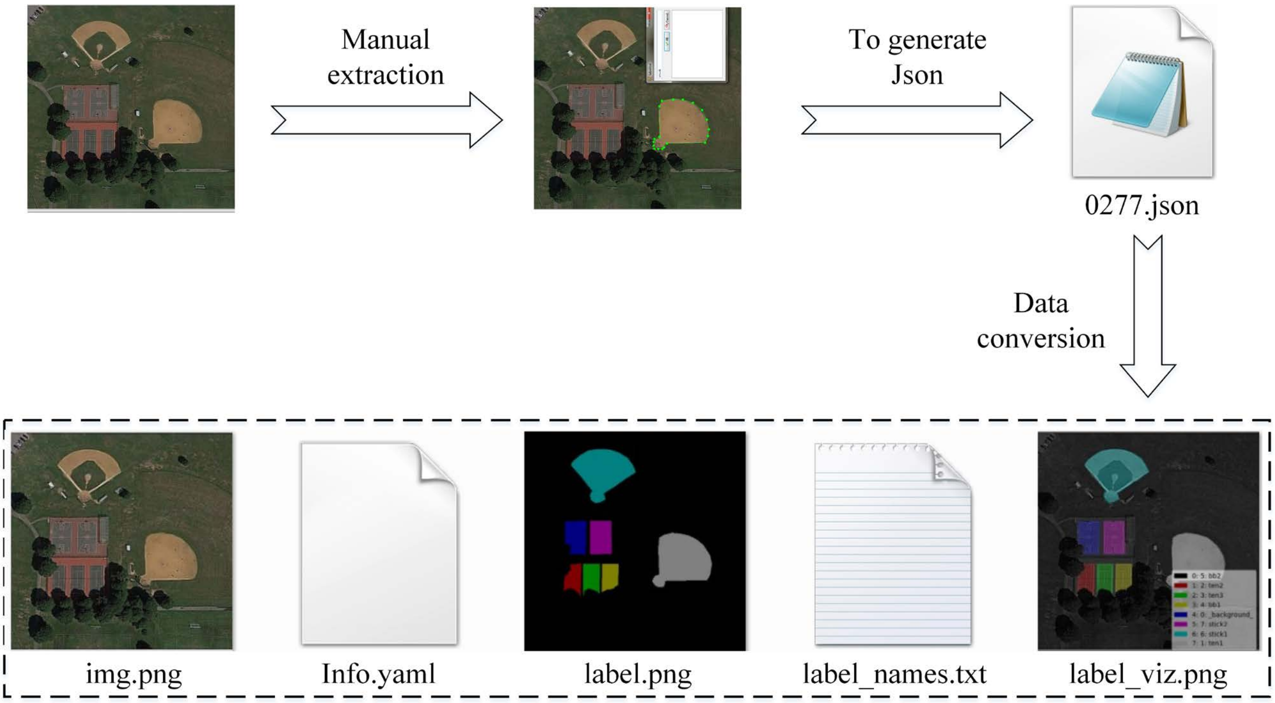


Figure 1. Technical procedure of sample construction method. The original image is manually labeled to generate the label, mask, and position information of the target sample.

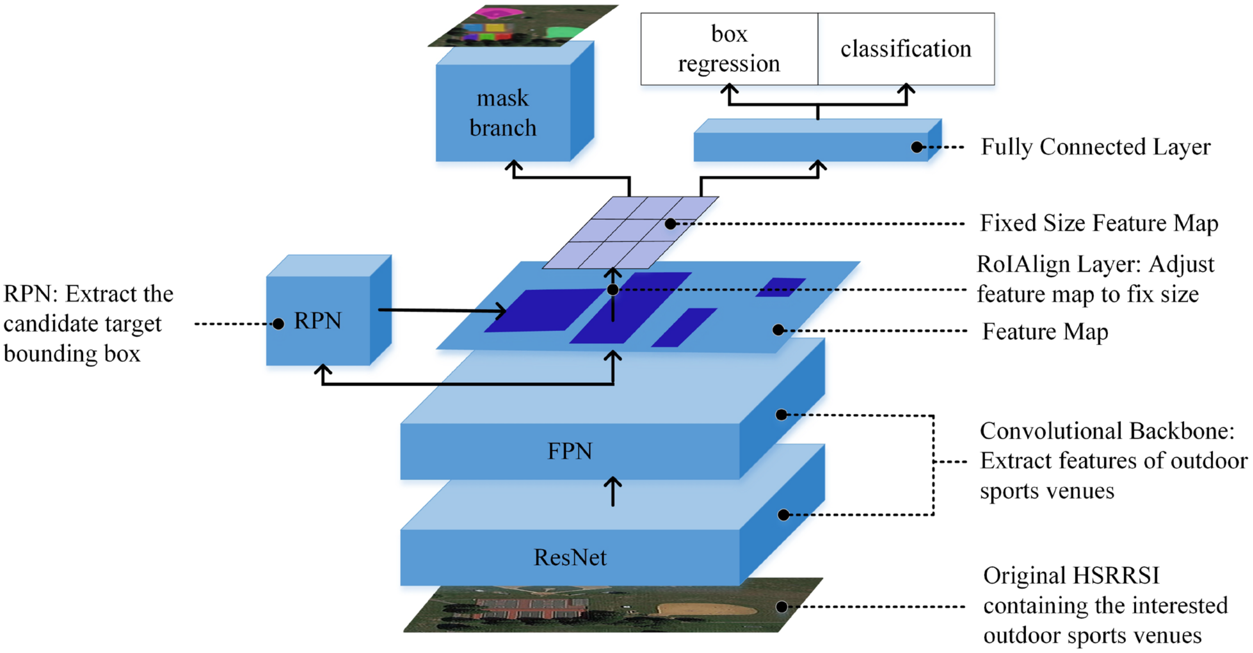

box regression
classification
mask branch
Fully Connected Layer
Fixed Size Feature Map
RoIAlign Layer: Adjust feature map to fix size
Feature Map
RPN: Extract the candidate target bounding box
RPN
FPN
ResNet
Convolutional Backbone: Extract features of outdoor sports venues
Original HSRRSI containing the interested outdoor sports venues

Figure 2. Mask R-CNN overall framework.

Regional Proposed Network (RPN); and in the third stage, the RoIAlign layer is used from each candidate box. The prediction class, frame offset refinement and output binary mask are processed in the same time to classify, regress and segment.

The HSRRSI outdoor sports ground object recognition method based on Mask R-CNN [13] deep learning can be summarized into three steps: training the Mask R-CNN model with the sample data set, using the verification data set to detect the model performance, and testing the data based on the trained Mask R-CNN model. The overall flow chart is shown in Figure 3.

## 2.3. Traditional Object-Oriented of Recognition

In eCognition®, the feature recognition process based on artificial design features can be summarized into three steps: segmentation, selection of classifiers for classification and accuracy evaluation. In this paper, the multi-scale segmentation and classification methods are chosen to compare with deep learning. In order to ensure better segmentation effect, the band weights, scale parameters and homogeneity criteria used in segmentation will be different [14]. In this experiment, the band weight is fixed to B:1G:1R:1NIR:1, the size generally between 50 - 60, the shape 0.5 - 0.6, the hue 0.4 - 0.5, the smoothness 0.5 - 0.6, and the compactness 0.4 - 0.5. Then, the Decision Tree, Bayes, KNN and Random Forest [15] [16] [17] are used to train the samples, and finally the classification is performed by using the results. If the classification result is not satisfactory, the parameters are adjusted until a satisfactory classification result is obtained. Finally, the classification results are evaluated for accuracy. The overall technical flow chart is shown in Figure 4.

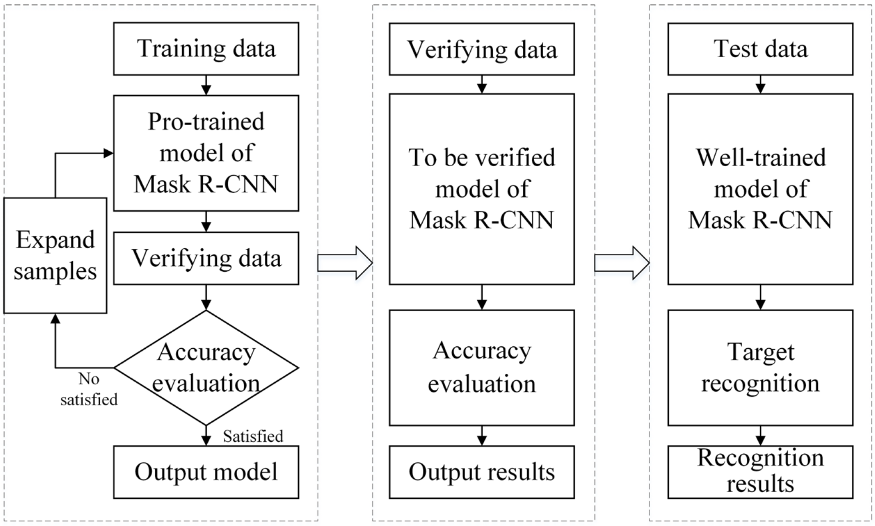


Figure 3. The overall technical procedure. The Mask R-CNN model is trained with the sample data set, and the performance of the model is verified by the verifying data. The test data is identified based on the well-trained Mask R-CNN model.

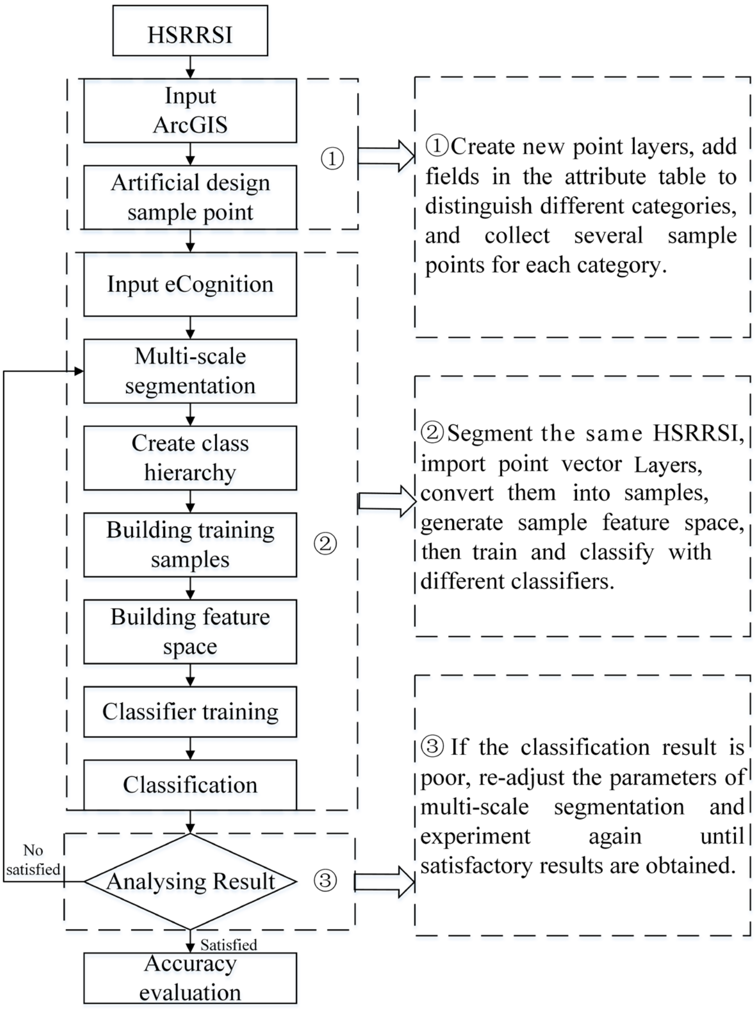

HSRRSI
Input ArcGIS
Artificial design sample point
①
①Create new point layers, add fields in the attribute table to distinguish different categories, and collect several sample points for each category.
Input eCognition
Multi-scale segmentation
Create class hierarchy
Building training samples
Building feature space
Classifier training
Classification
②
②Segment the same HSRRSI, import point vector Layers, convert them into samples, generate sample feature space, then train and classify with different classifiers.
No satisfied
Analysing Result
③
③ If the classification result is poor, re-adjust the parameters of multi-scale segmentation and experiment again until satisfactory results are obtained.
Satisfied
Accuracy evaluation

Figure 4. The traditional procedure of object-orientation recognition.

### 2.4. Application Research on Engineering Ability of the Trained Model

In order to verify the engineering application ability of the deep learning model trained in this paper, an empirical application experiment is performed on the HSRRSI of Xicheng and Daxing District of Beijing respectively, and the generalization ability of the trained model of Mask R-CNN is evaluated.

## 3. Experiments and Discussion

### 3.1. Experimental Result and Quality Assessment

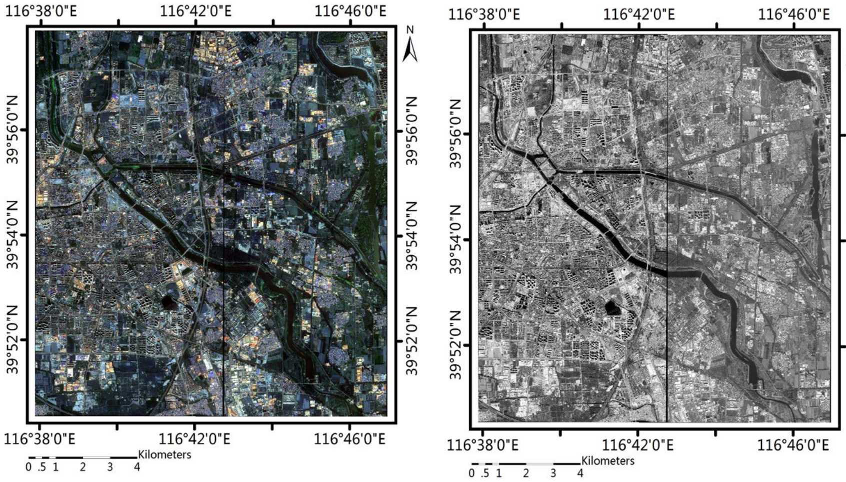
#### 3.1.1. Study Area and Data

Three images are used in the experiment. One covers Tongzhou district of Beijing taken by the WordView-3 satellite in 2014, as shown in Figure 5. The second part is the image set of Northwestern Polytechnical University NWPUVHR-10 [18]. The third part is the image set of Wuhan University team RSOD-Dataset [19]. The images for the experiment are selected from aforementioned three parts, and then collect and produce sample data sets.

Considering the actual conditions of ground features of outdoor sports venues in China, this paper chooses outdoor football field, basketball court, tennis court and baseball field as the targets to be identified. As shown in Figure 6, the characteristics of each type of feature are as follows:

**1) The characteristics of outdoor sports venues**

As shown in Figure 6(a), most of football fields have the standard geometric shape as shown in Figure 6(a1). Football field with non-standard size and shape is shown in Figure 6(a2). The football field with unconventional texture is shown in Figure 6(a3). The football field with other sports fields (composite football field) is shown in Figure 6(a4). Their outer contour feature is similar,



(a) (b)

Figure 5. Study area. The area covers 175 km$^2$ of urban area in Tongzhou, Beijing, China, and is located from 39˚50'33" (N) to –39˚57'53" (N) and from 116˚37'53" (E) to 116˚46'57" (E); (a) multispectral image (2 m); (b) panchromatic image (0.5 m).

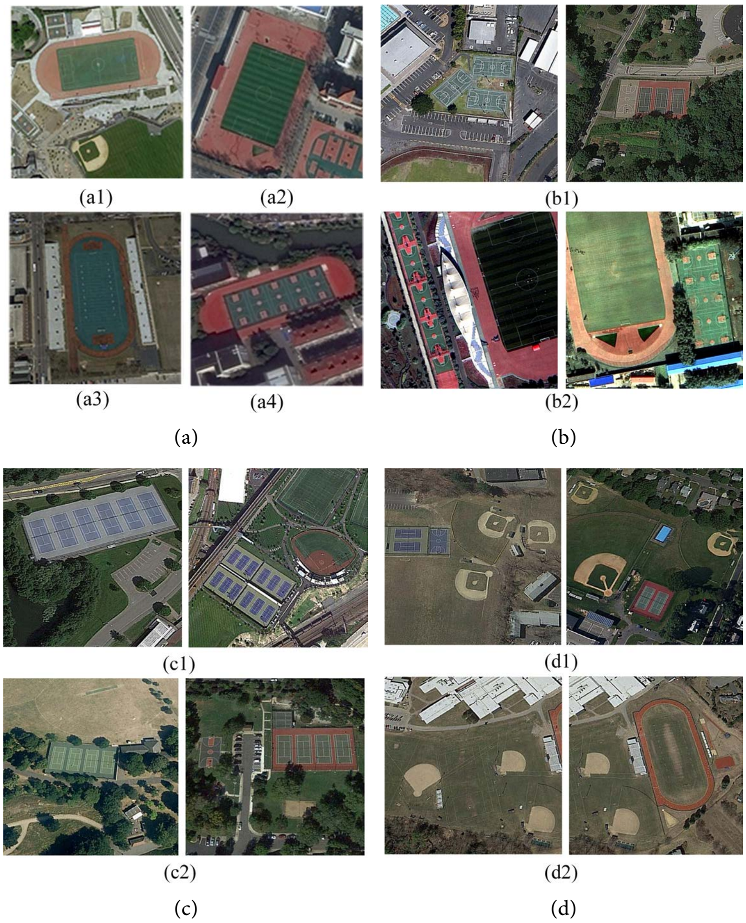


Figure 6. The example images of sample data set. (a) Football field; (b) Basketball court; (c) Tennis court; (d) Baseball field.

but the internal features are diverse, especially the composite football field, there is not much regular pattern to follow.

For the basketball court, as shown in Figure 6(b), it is mainly divided into line type and material type. The main problem in the line-type basketball court is that the line information may be missing due to lack of maintenance for a long time, and it is difficult to identify even if the image is enhanced. The material-differentiated basketball court has various spectral characteristics depending on the material.

For the tennis court, as shown in Figure 6(c), it is mainly a line type, usually with two textures of green and blue rubber. The difference between the tennis court and the basketball court texture features is obvious, but they are similar to the badminton court and the volleyball court.

For the baseball field, as shown in Figure 6(d), it is mainly divided into a solid baseball field and a non-solid baseball field. The solid baseball field is a piece of land. The non-solid baseball field has a piece of grass in the center. Most baseball courts have no outer contours and the boundaries are often unclear.

**2) The sample data set construction**

When extracting samples, the rules are:

1) All feature types must be included in the outline information when manually drawing;

2) For large-area shadows, we can choose to avoid, so as not causing false recognition of the neural network. Small area shadows (size less than 1/10) can be included to preserve the complete geometric characteristics of the object;

3) For a compound football field, there is no need to extract other type of sports ground which overlap on the football field, just follow the outline of the football field.

After the clipping of the original images, a dataset containing 613 sample images with the size of 500 * 500 is generated. 481 images are selected as training data, 102 images are used for verification, and 30 images are used as test data. Table 1 shows, the samples number of outdoor sports venues (football field, basketball court, tennis court and baseball field) used to train and test the Mask R-CNN model.

### 3.1.2. Assessment Metrics

To quantify the effectiveness of land features recognition, we use an object-based image analysis method to evaluate the verification results. Due to the difference between the recognition principle of the Mask R-CNN method and the other four methods from eCognition®, the Mask R-CNN method uses a single feature as the object unit, and the Decision Tree, Bayes, KNN and Random Trees methods use the segmentation block as the object unit. We use six common indicators to evaluate the performance of Mask R-CNN method. Precision, mPrecision, Recall, mRecall, AP, mAP, where AP and mAP are used to evaluate the overall performance of the Mask R-CNN method. The recognition results can be judged from four categories: 1) true positive (TP), positive was recognized as positive; 2) false positive (FP), negative was recognized as positive; 3) false Negative (FN), positive was recognized as negative; 4) true negative (TN), negative was recognized as negative. The evaluation indicator definition is:

$$\text{Precision} = \text{TP}/(\text{TP}+\text{FP}) \tag{1}$$

$$\text{Recall} = \text{TP}/(\text{TP}+\text{FN}) \tag{2}$$

$$\text{AP} = \int_0^1 P(R)\,\mathrm{d}R \tag{3}$$

In Formula (3), $P$ is the Precision and $R$ is the Recall.

Precision reflects the accuracy of prediction positive; Recall reflects the ability of covering positive. There is a certain constraint between these two indicators.

Table 1. Land features sample dataset.

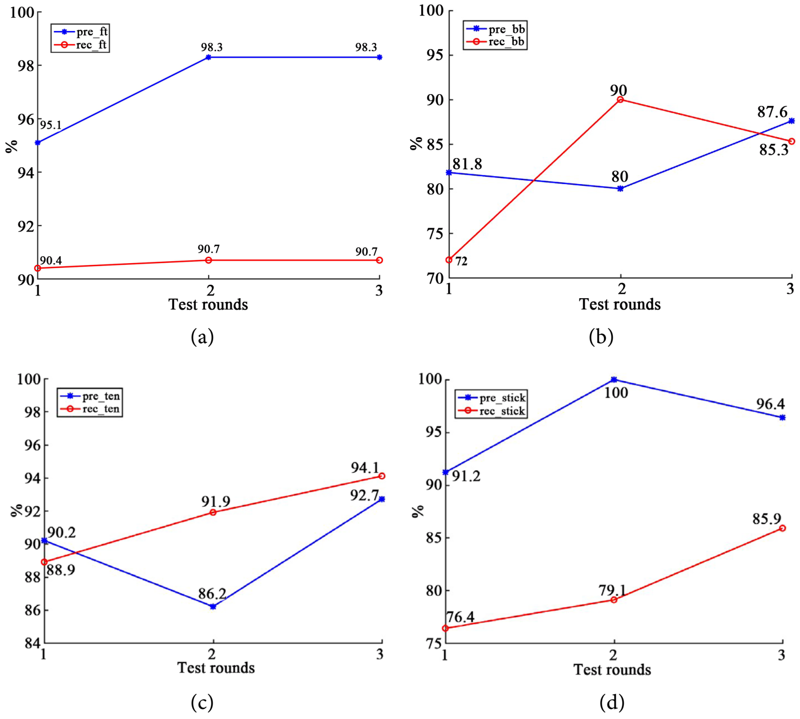
| Samples number | Football field | Basketball court | Tennis court | Baseball field |
|---|---|---|---|---|
| Training samples | 357 | 421 | 413 | 254 |
| Testing samples | 60 | 136 | 75 | 57 |
| Handover box | 417 | 557 | 488 | 311 |

When the recall is high, the number of missing recognition will decrease, and the number of wrong recognition will increase, and the accuracy will decrease. Considering the accuracy and recall of a set of data to evaluate the algorithm has limitations, so this paper cites average accuracy (AP). The AP comprehensively considers Precision and Recall to evaluate the overall performance of the Mask R-CNN method. Generally, the higher the AP value, the better the recognition effect. The mAP, mPrecision, and mRecall are the average values of all AP, Precision, and Recall when multiple classes are detected.

### 3.1.3. Experimental Result of Mask R-CNN Method

As shown in Figure 7, this study compares and analyses the accuracy of the training model for each type of feature when the number of training samples is 201, 398, and 481, respectively. As shown in Figure 7, in the first experiment, the model trained with 201 training samples shows higher recognition accuracy for football fields and tennis courts (precision: 0.951, 0.902; recall: 0.904, 0.889; respectively). The precision of the baseball field and the recall of the basketball court are not ideal (0.764 and 0.72, respectively). Therefore, the sample numbers of the four types of target objects are increased to check on the results again.

The second experiment nearly doubled the number of training data to 398. The accuracy evaluation results show that the recognition accuracy of the football field and the baseball field has been improved (the accuracy of 0.983, 1, respectively, the recall of 0.907, 0.791, respectively), but the precision of the tennis



(a)

(b)

(c)

(d)

Figure 7. Recognition results in different test rounds. (a) Recognition result of football field; (b) Recognition result of basketball court; (c) Recognition result of tennis court; (d) Recognition result of baseball field.

court and basketball court becomes lower (the precision of the basketball court dropped from 0.818 to 0.8, and the precision of the tennis court dropped from 0.902 to 0.862).

The reasons for different experiment precision are: The first training round is to test the feasibility of the experimental scheme. The images with high sharpness, rarely shaded, containing few other features of the object are selected as the training data. Generally speaking, the high quality sample data are limited, so the quality of the images used in the second experiment is slightly lower than the first, because the geometric characteristics of the football field and the baseball field are more obvious comparing with the tennis court and basketball court. So in the case of lower sample quality, increasing the number of samples (adding 46 football fields, 88 baseball fields) can still improve the recognition accuracy. The geometric characteristics of basketball courts and tennis courts are not prominent. The distinction is mainly based on internal texture features, so in the case of lower sample quality, increasing the number of samples (adding 67 tennis courts, 50 basketball courts) may lead to accuracy decrease.

In the second experiment, the first training samples are carefully selected and are enhanced by using the rotation operation to turn the sample into the training model again [20]. Using these data in the third experiment, we guarantee the quality of the training data and solve the problem of insufficient good quality samples. At the same time, the second accuracy evaluation results show that the precision of the baseball field is the lowest, so the number of baseball field samples is mainly increased in the training sample. The total training data have 481 images, by adding 46 samples of football fields, 67 samples of tennis courts, 50 samples of basketball courts and 88 samples of baseball fields to the second experiment. The third experiment Mask R-CNN method training takes 3 hours and 40 minutes. The accuracy of each feature category is shown in **Table 2**, and the partial recognition results are shown in **Figure 8**.

As shown in **Table 2**, the basketball court is inferior to other features (the precision of 0.8767, the recall of 0.8533, the mean precision of 0.8455), and the football field has the best recognition performance (0.9076 of precision, 0.9833 of recall, 0.9830 of mean precision). **Figure 8** depicts the representative recognition performance of the four types of features. We select images including only one feature and including other three types of features respectively as much as

**Table 2.** Recognition results of different objects.

| Category | Actual object | Detected object | Matching feature | Precision | Recall | AP |
|---|---|---|---|---|---|---|
| Football field | 60 | 65 | 59 | 0.9076 | 0.9833 | 0.9830 |
| Basketball court | 75 | 73 | 64 | 0.8767 | 0.8533 | 0.8455 |
| Tennis court | 136 | 138 | 128 | 0.9275 | 0.9411 | 0.9182 |
| Baseball field | 57 | 64 | 55 | 0.8593 | 0.9649 | 0.9473 |
| Overall | | | | 0.8927 | 0.9356 | 0.9235 |

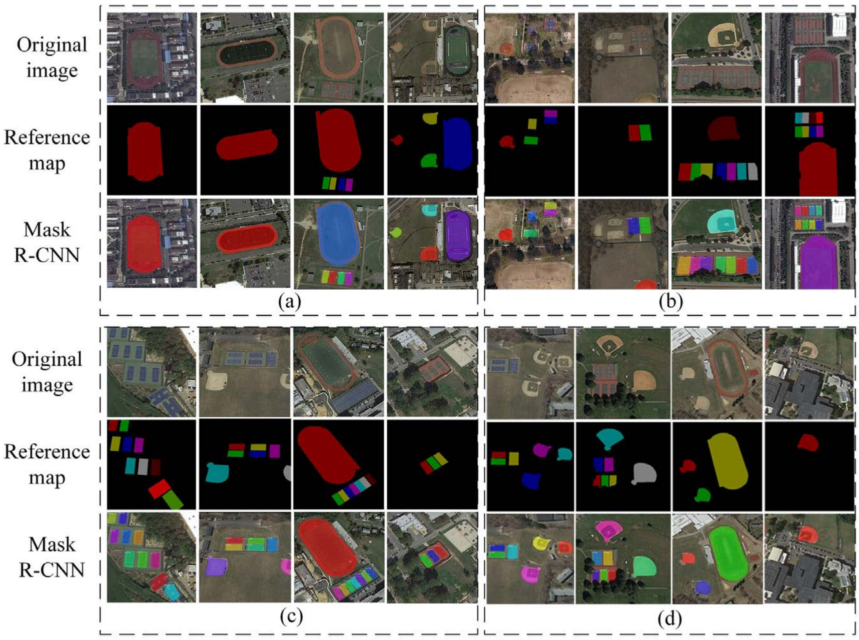


Figure 8. Recognition results based on the Mask R-CNN. Color is positive and black is negative in Reference map; the colored area is the recognized feature in Mask R-CNN. (a) Recognition results of football field; (b) Recognition results of basketball court; (c) Recognition results of tennis court; (d) Recognition results of baseball field.

possible. Each color mask area represents the identified feature area. Each recognition mask is labeled with its identified feature type and recognition confidence. It can be seen from Figure 8 that the overall recognition performance is good (the overall precision is 0.8927, 0.9356 of recall and 0.9235 of mean precision). On the other hand, there are two problems: one is that the identification of some sports venues is incomplete because of the shade or shadow; the second is that the segmentation at the edge of the feature is not accurate enough. These two problems are not effectively solved in this paper.

### 3.1.4. Experimental Result of Traditional Object-Oriented Method

In order to evaluate the performance of the model proposed in this paper, we use the eCognition® software, which plays an important role in the field of object-oriented image analysis technology, to conduct comparative experiments. We chose Decision Tree, Bayes, KNN and Random Forest as classifiers. The partial recognition results are shown in Figure 9.

Figure 9(a) shows a representative classification result of the football field. When the image only contains the football field, the recognition performance of the Decision Tree, Bayes and KNN is not very different, while the Random Forest shows more missing or wrong recognition. When the image contains football field and other features, the four classifiers still do well in recognizing the football field, clearly distinguish the football field and others.

Figure 9(b) shows a representative classification result of the basketball court. When the image mainly contains basketball court, the recognition result of Bayes can barely identify each object. We can hardly identify the location of the

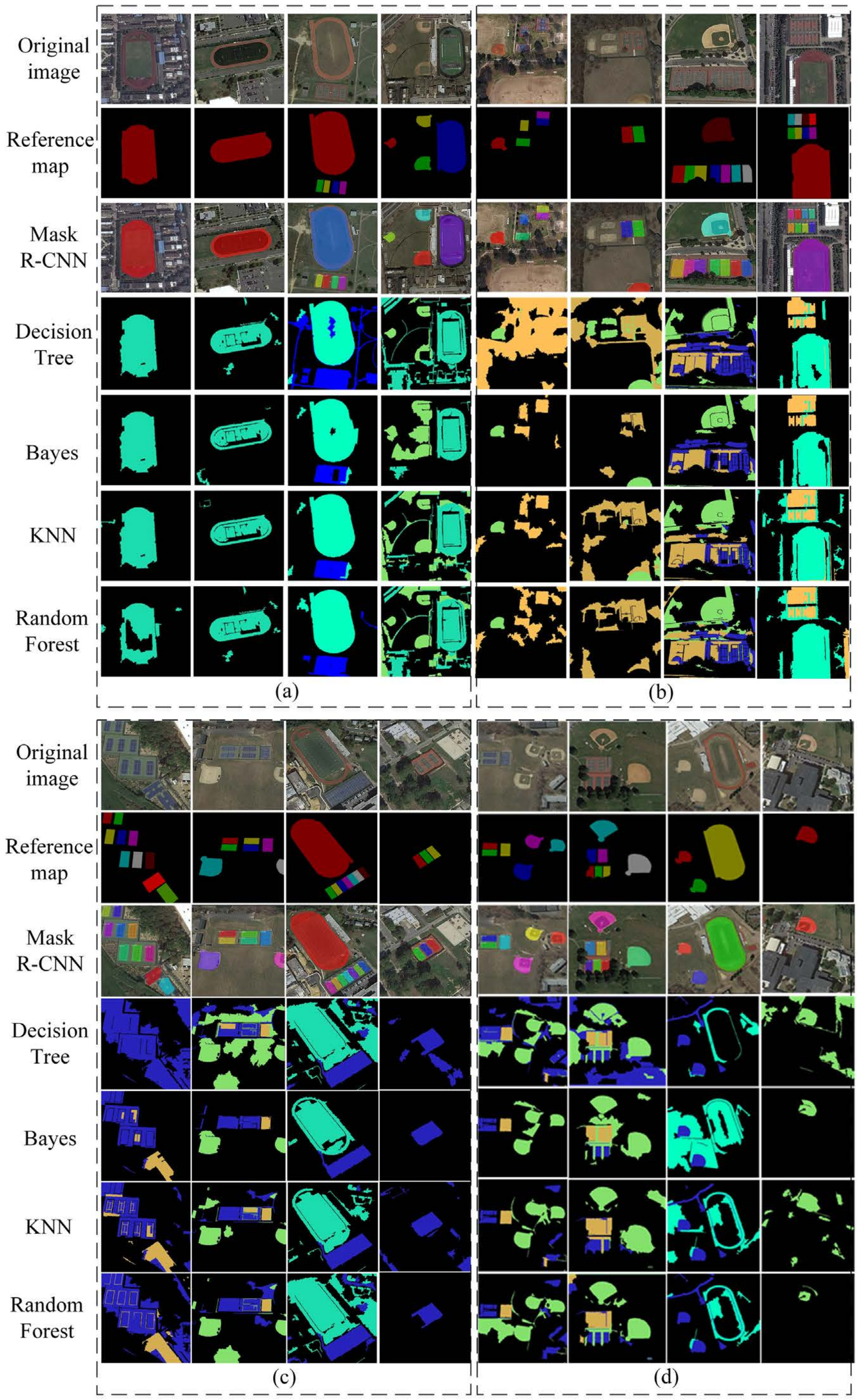


**Figure 9.** Recognition results based on the different classifiers. Color is positive and black is negative in reference map; the colored area is the recognized feature in Mask R-CNN. (a)-(d) are Recognition results of football field, basketball court, tennis court and baseball field. (a) Recognition results of football field; (b) Recognition results of basketball court; (c) Recognition results of tennis court; (d) Recognition results of baseball field.

basketball court though the results of the other three classifiers due to the large-scale missing or wrong recognition. When the image contains other features, the missing recognition with the tennis court is the most serious, and the recognition results are mixed. When several basketball courts are next to each other, the recognition results are in a single piece. It is almost impossible to identify the location and number of each basketball courts.

Figure 9(c) shows a representative classification result of the tennis court. The classification results are similar to basketball courts. The Bayes recognition results are relatively better than the other three classifiers. Basketball court and tennis field is more likely to be missed in recognition. When several tennis courts are next to each other, the recognition results are in a single piece. It is almost impossible to identify the position and number of each tennis courts.

Figure 9(d) shows a representative classification result of the baseball field. When the image only contains the baseball field, the Bayes performs better, and the other three classifiers have relatively poor classification results. When the image contains other features, the recognition results of the four classifiers are almost the same. The common problem is that for the solid baseball field, when the grass are collected as sample points, the classifier may mistake the grassland for a baseball field, causing a large area of the wrong recognition; when the grass are not collected as sample points, the classifier can hardly recognize the baseball field as a whole, causing partial missing recognition.

From a qualitative point of view, the outlines of the features recognized by the four classifiers are relatively rough; when the materials constituting the target features are different, there always occurs hollow in the recognition mask; when the target features are similar to the surrounding environment, there always are large-scale missing recognition at the same time; when several target objects are connected, the recognition results are connected into one piece. Among the four classifiers, the recognition results of Bayes are generally better, and the Decision Tree is the worst.

From a quantitative point of view, the precision, recall, and Kappa of each graph are presented, then the average values of 50 images are calculated. As can be seen from Table 3, the Kappa for all types are above 0.65, indicating that all classifiers perform well in the recognition. For the results of football and basketball courts, Bayes reaches the highest value in all indicators. For the results of tennis courts, Bayes reaches the highest value in precision and recall (0.7587 and 0.8414 respectively); and the Random Forest reaches the highest Kappa, which is 0.8152. For the results of baseball field, three indicators are distributed on different classifiers, the Decision Tree reaches the highest accuracy of 0.8361, the Bayes reaches the highest recall of 0.8798, and the Random Forest reaches the highest Kappa of 0.9254.

In order to obtain the overall evaluation results, this paper calculates the arithmetic mean of the three indicators of the four classifiers. In general, the Bayes has the best recognition performance and the Decision Tree is the worst.

## 3.2. Comparison of Different Methods

### 3.2.1. Comparison of Different Methods

The deep learning method requires a lot of manpower in the early stage to

**Table 3.** Recognition precision of object-oriented methods.

| Category | Indicator | Decision tree | Bayes | KNN | Random forest |
|---|---|---|---|---|---|
| Football field | Precision | 0.7534 | **0.8017** | 0.7831 | 0.7132 |
| | Recall | 0.8264 | **0.8652** | 0.8564 | 0.8446 |
| | Kappa | 0.6552 | **0.8341** | 0.7671 | 0.8035 |
| Basketball court | Precision | 0.7226 | **0.8313** | 0.8116 | 0.7537 |
| | Recall | 0.8451 | **0.8827** | 0.8676 | 0.8512 |
| | Kappa | 0.6769 | **0.9330** | 0.8861 | 0.8937 |
| Tennis court | Precision | 0.7313 | **0.7587** | 0.7392 | 0.7441 |
| | Recall | 0.8016 | **0.8414** | 0.7961 | 0.8083 |
| | Kappa | 0.6956 | 0.7390 | 0.7868 | **0.8152** |
| Baseball field | Precision | **0.8361** | 0.8005 | 0.8112 | 0.8278 |
| | Recall | 0.8567 | **0.8798** | 0.8465 | 0.8653 |
| | Kappa | 0.7372 | 0.8909 | 0.8207 | **0.9254** |
| Total | mPrecision | 0.7608 | **0.7980** | 0.7862 | 0.7594 |
| | mRecall | 0.8324 | **0.8672** | 0.8416 | 0.8423 |
| | Kappa | 0.6912 | 0.8492 | 0.8151 | **0.8594** |

produce a large amount of sample data that can be input into the neural network. In addition, in order to improve the generalization ability of the neural network, high-quality and multi-source data is also required, and this process requires a large amount of manual participation. When the network training is completed, the process of recognizing is completely automatic. At this time, the neural network can automatically identify the targets in the data to be detected by using the effective features learned from the samples, and no human interaction is needed. The method has certain value in the research direction of remote sensing image automatic detection and recognition of ground objects.

Traditional machine learning classification methods rely on human interaction from start to finish. From object-oriented segmentation, optimization of feature space to training samples and classification, professional experience is required to design parameters, and continuous debugging is performed to the appropriate effect. There is no accurate measurement standard and artificial error is also large. In addition, the recognition result is greatly affected by the segmentation effect and the artificially designed sample quality, so the result is unstable. The common method is to compare the results under several different conditions to identify the best. Therefore, there are many steps, relying on manpower, and the process is more complicated.

### 3.2.2. Qualitative Analysis of Recognition Results

The Mask R-CNN method expresses the recognition result by generating a translucent mask on the surface of the targets. Each target can be clearly presented on the result image, and the visualization effect is better. The mask divides the edge of the target object more accurately, and there is no fragmentation in the results. It is difficult for the traditional classifier to subdivide the object. The recognition result is very inaccurate at the outlines of the object, and always has hollow in the recognition results, resulting in incomplete recognition. When the targets are similar to the surrounding material, wrong recognition happens at large-scale.

#### 3.2.3. Quantitative Analysis of Recognition Results

As can be seen from Table 4, in the recognition of the basketball court, the Bayes achieves a recall of 0.8827, slightly higher than the 0.8533 of the Mask RCNN, but its precision is 0.8313, which is significantly lower than 0.8767 from Mask R-CNN. Therefore, the comprehensive evaluation of the Mask R-CNN is still better. In the recognition of other types, the Mask R-CNN has significantly higher indicators than the four traditional classifiers. Therefore, it can be seen that the Mask RCNN method not only in the accuracy of each type respectively, but also in the accuracy of the four classes is better than the four traditional classifiers. This shows that the traditional classification method is obviously insufficient.

### 3.3. Empirical Application and Quality Assessment

From the aspect of practical engineering application, using and validating the well-trained deep learning model for those four outdoor sports venues studied in this paper, an empirical application experiment is performed on the HSRRSI of Xicheng and Daxing District of Beijing respectively, and the generalization ability of the trained model of Mask R-CNN is evaluated.

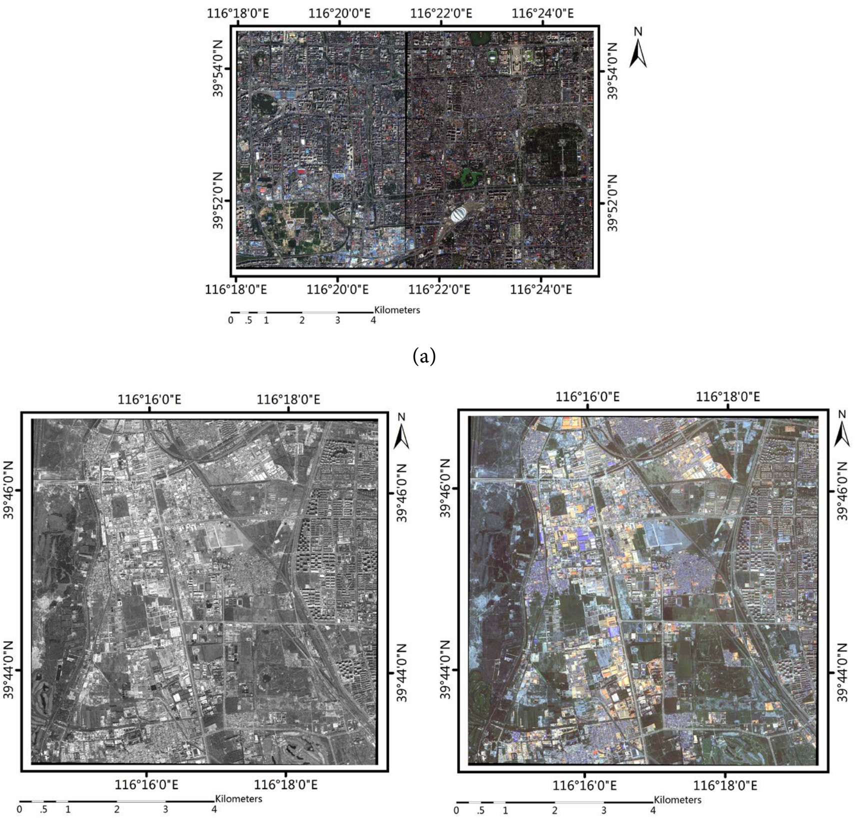
#### 3.3.1. Study Area and Data

The data uses in the empirical experiment come from two parts as follows.

One part is HSRRSI of Xicheng, Beijing, China, taken by World View satellite in 2012, as shown in Figure 10(a), covering an area of 50.70 km$^2$. The spatial resolution is 0.5 m, including three bands R, G, and B images.

Table 4. Recognition results of different classifier.

| Category | Indicator | Decision tree | Bayes | KNN | Random forest | Mask R-CNN |
|---|---|---|---|---|---|---|
| Football field | Precision | 0.7534 | 0.8017 | 0.7831 | 0.7132 | **0.9076** |
| | Recall | 0.8264 | 0.8652 | 0.8564 | 0.8446 | **0.9833** |
| Basketball court | Precision | 0.7226 | 0.8313 | 0.8116 | 0.7537 | **0.8767** |
| | Recall | 0.8451 | **0.8827** | 0.8676 | 0.8512 | 0.8533 |
| Tennis court | Precision | 0.7313 | 0.7587 | 0.7392 | 0.7441 | **0.9275** |
| | Recall | 0.8016 | 0.8414 | 0.7961 | 0.8083 | **0.9411** |
| Baseball field | Precision | 0.8361 | 0.8005 | 0.8112 | 0.8278 | **0.8593** |
| | Recall | 0.8567 | 0.8798 | 0.8465 | 0.8653 | **0.9649** |
| Total | mPrecision | 0.7608 | 0.7980 | 0.7862 | 0.7594 | **0.8927** |
| | mRecall | 0.8324 | 0.8492 | 0.8416 | 0.8423 | **0.9356** |



(a)

(b)

**Figure 10.** Empirical experiment study area. (a) Area of Xicheng, Beijing, China, three bands R, G, B images; (b) Area of Daxing, Beijing, China, panchromatic image with spatial resolution of 0.5 m (left), multispectral image of 2 m (right).

The other part is HSRRSI of Daxing, Beijing, China, taken by WorldView satellite in 2013, as shown in **Figure 10(b)**, covering an area of 1031 km$^2$, which includes panchromatic image with spatial resolution of 0.5 m and multispectral image of 2 m.

In addition to the fusion processing of the Xicheng District image, the two parts of the image data are pre-processed, which includes Gram-Schmidt Pan Sharpening fusion, image framing, 2% - 98% maximum and minimum linear stretching, Laplacian filtering. Eventually, 125 images of 500 * 500 are selected as test images. Then they are input in Labelme® to extract the target feature samples. The feature characteristics between this data and experimental data are quite different. Due to the shortage of land resources in Beijing and other reasons, the phenomenon of composite use of various sports venues is more prominent. The phenomenon of basketball courts in football stadiums is more common, and even basketball courts contain tennis courts. In addition, due to the influence of China's sports preferences, the baseball field is very limited, and there is no hollow baseball field in this data.

### 3.3.2. Assessment Metrics

All assessment metrics used are the same with those from the previous experiment.

### 3.3.3. Empirical Application Result

125 images in the empirical engineering application data set are evaluated for accuracy assessment. The results of outdoor sports venues recognition and precision are shown in Figure 11 and Table 5.

As can be seen from Table 5, the four recognition value of the football field reaches the best recognition among the all types. The main reason is that the geometric characteristics of the football field are relatively obvious, and it has a good distinction with other feature categories. Further study can include more samples of composite football fields to improve the robustness of model. The baseball field has the lowest recognition precision of 0.7778. However, due to the limited number of samples in the baseball field, it is not suitable to determine the accuracy of Mask R-CNN on the baseball field based solely on this value. A certain amount of sample field of the baseball field should be added to more accurately evaluate it. At the same time, the recall of the baseball field reaches a maximum of 1.0, indicating that there is no missing recognition in the entire test sample set. The main reason is that the geometric characteristics of the baseball field are also obvious and varied greatly from other features.

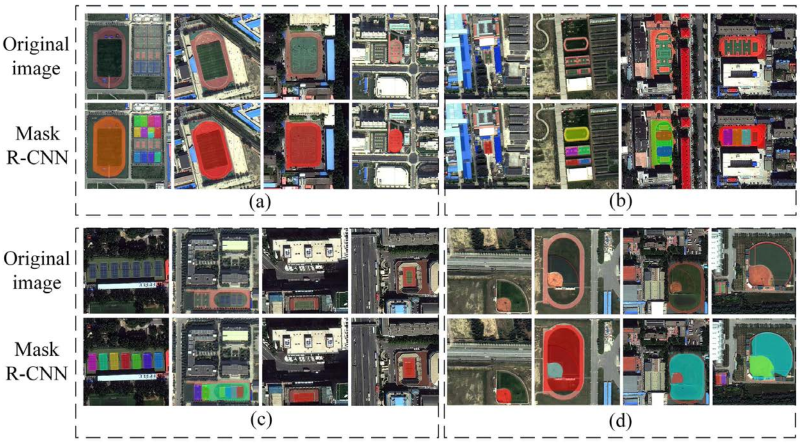


Figure 11. Recognition results based on the Mask R-CNN. The colored area is the recognized feature in Mask R-CNN. (a)-(d) are recognition results of football field, basketball court, tennis court and baseball field. (a) Recognition results of football field; (b) Recognition results of basketball court; (c) Recognition results of tennis court; (d) Recognition results of baseball field.

Table 5. Recognition results in empirical engineering application.

| Category | Actual object | Detected object | Matching feature | Precision | Recall | AP |
|---|---|---|---|---|---|---|
| Football field | 62 | 65 | 60 | 0.9230 | 0.9677 | 0.9558 |
| Basketball court | 176 | 145 | 143 | 0.8238 | 0.8125 | 0.8346 |
| Tennis court | 52 | 54 | 46 | 0.8518 | 0.8846 | 0.8725 |
| Baseball field | 7 | 9 | 7 | 0.7778 | 1.0 | 0.9657 |
| Overall |  |  |  | 0.8441 | 0.9162 | 0.9071 |

The recognition results of the above two types has been consistent with previous experiments, indicating that the Mask R-CNN algorithm is sensitive to geometric features of features.

The average precision of the basketball court is 0.8365, which is the lowest value in the four types. The reason is that the existence form of the basketball court in this experimental area is quite complicated, most of which are contained in the football field. There are also shades and shadows in the field, which has a great interference to the recognition. Improving the recognition accuracy of the composite football field is the key point in the further study.

The four recognition value of the tennis court is closest to the average of indicators, which indicates that the model has better recognition ability for the tennis court. The next step can be to increase samples to improve the recognition accuracy.

The overall precision reaches 0.8441, the recall of 0.9162, the average precision of 0.9071, indicating good recognition ability of model.

For a more intuitive display to identify the generalization ability of the Mask R-CNN model on different data sets, we compare the evaluation indicators of the experimental data and the empirical data recognition results. It can be seen from **Table 6** that the values are floating, and most of the values slightly decreased. The reason is described above, and the value of overall precision is still good, indicating that the model has certain generalization ability.

## 4. Conclusions

This paper proposed a set of object recognition methods and technical flow based on Mask R-CNN, which can be used to recognize four outdoor sports ground of football field, basketball court, tennis court and baseball field from HSRRSI. The main research achievements include:

1) The experimental results show that the trained Mask R-CNN model is effective and applicable for the recognition of four outdoor sports ground in HSRRSI, and the overall precision and recall are respectively 0.8927 and 0.9356.

**Table 6.** Comparison of two experiments.

| Category | Indicator | Experiment data | Empirical data |
|---|---|---|---|
| Football field | Precision | 0.9076 | 0.9230 |
| | Recall | 0.9833 | 0.9677 |
| Basketball court | Precision | 0.8767 | 0.8238 |
| | Recall | 0.8533 | 0.8125 |
| Tennis court | Precision | 0.9275 | 0.8518 |
| | Recall | 0.9411 | 0.8846 |
| Baseball field | Precision | 0.8593 | 0.7778 |
| | Recall | 0.9649 | 1.0 |
| Overall | mPrecision | 0.8927 | 0.8441 |
| | mRecall | 0.9356 | 0.9162 |

2) Comparing the trained Mask R-CNN model with the object-oriented Decision Tree, Bayes, KNN and Random Forest, it shows that the Mask R-CNN is better than the traditional methods not only in the technical flow but also in the recognition results.

3) Applying the well-trained Mask R-CNN model to the HSRRSI of Beijing Xicheng in 2012 and Beijing Daxing in 2013, the overall precision and recall have achieved 0.8441 and 0.9162 respectively, and the model has certain generalization ability and engineering application value.

There are still many exploration spaces in the model. In future research, we will consider modifying the internal parameters of the neural network and expanding samples to improve the model recognition performance.

## Acknowledgements

This research is funded by 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), and the Educational Science Research Project of Beijing University of Civil Engineering and Architecture(No. Y1613).

## Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

## References

[1] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V. and Rabinovich, A. (2015) Going Deeper with Convolutions. *Conference on Computer Vision and Pattern Recognition*, Boston, 7-12 June 2015, 2-8. https://doi.org/10.1109/CVPR.2015.7298594

[2] Girshick, R., Donahue, J., Darrell, T. and Malik, J. (2014) Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. *Conference on Computer Vision and Pattern Recognition*, Columbus, 23-28 June 2014, 580-587. https://doi.org/10.1109/CVPR.2014.81

[3] Girshick, R. (2015) Fast R-CNN. In: *Proceedings of the* 2015 *IEEE International Conference on Computer Vision*, IEEE Computer Society, Washington DC, 1440-1448. https://doi.org/10.1109/ICCV.2015.169

[4] He, K.M., Zhang, X.Y., Ren, S.Q. and Sun, J. (2014) Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition. *Transactions on Pattern Analysis & Machine Intelligence*, **37**, 1904-1916. https://doi.org/10.1109/TPAMI.2015.2389824

[5] Ren, S.Q., He, K.M., Girshick, R. and Sun, J. (2015) Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, **39**, 1137-1149.

https://doi.org/10.1109/TPAMI.2016.2577031

[6] Xu, Y.Z., Yu, G.Z., Wang, Y.P., Wu, X.K. and Ma, Y.L. (2017) Car Detection from Low-Altitude UAV Imagery with the Faster R-CNN. *Journal of Advanced Transportation*, **2017**, Article ID: 2823617. https://doi.org/10.1155/2017/2823617

[7] Sun, X.D., Wu, P.C. and Hoi, S.C.H. (2018) Face Detection Using Deep Learning: An Improved Faster RCNN Approach. *Neurocomputing*, **299**, 42-50.

[8] He, K.M., Gkioxari, G., Dollár, P. and Girshick, R. (2018) Mask R-CNN. *IEEE International Conference on Computer Vision*, Venice, 22-29 October 2017, 1. https://doi.org/10.1109/TPAMI.2018.2844175

[9] Zhao, K., Kang, J., Jung, J. and Sohn, G. (2018) Building Extraction from Satellite Images Using Mask R-CNN with Building Boundary Regularization. *Conference on Computer Vision and Pattern Recognition Workshops*, Salt Lake City, 18-22 June 2018, 247-251.

[10] Ji, S.P., Wei, S.Q. and Lu, M. (2018) Fully Convolutional Networks for Multisource Building Extraction from an Open Aerial and Satellite Imagery Data Set. *IEEE Transactions on Geoscience and Remote Sensing*, **57**, 574-586.

[11] Song, S.R., Liu, J.H., Pu, H., Liu, Y. and Luo, J.Y. (2019) The Comparison of Fusion Methods for HSRRSI Considering the Effectiveness of Land Cover (Features) Object Recognition Based on Deep Learning. *Remote Sensing*, **11**, 1435. https://doi.org/10.3390/rs11121435

[12] Wada, K. (2018) Image Polygonal Annotation with Python (Polygon, Rectangle, Circle, Line, Point and Image-Level Flag Annotation). https://github.com/wkentaro/labelme

[13] He, K.M., Zhang, X.Y., Ren, S.Q. and Sun, J. (2015) Deep Residual Learning for Image Recognition. *Conference on Computer Vision and Pattern Recognition*, Las Vegas, 27-30 June 2016, 770-778.

[14] Liu, J.H., Pu, H., Song, S.R. and Du, M.Y. (2018) An Adaptive Scale Estimating Method of Multiscale Image Segmentation Based on Vector Edge and Spectral Statistics Information. *International Journal of Remote Sensing*, **39**, 6826-6845.

[15] Ghatkar, J.G., Singh, R.K. and Shanmugam, P. (2019) Classification of Algal Bloom Species from Remote Sensing Data Using an Extreme Gradient Boosted Decision Tree Model. *International Journal of Remote Sensing*, **40**, 9412-9438. https://doi.org/10.1080/01431161.2019.1633696

[16] Lashkenari, M.S. and Khazaie Poul, A. (2016) Application of KNN and Semi-Empirical Models for Prediction of Polycyclic Aromatic Hydrocarbons Solubility in Supercritical Carbon Dioxide. *Polycyclic Aromatic Compounds*, **37**, 415-425.

[17] Jebur, M.N., Shafri, H.Z.M., Pradhan, B. and Tehrany, M.S. (2014) Per-Pixel and Object-Oriented Classification Methods for Mapping Urban Land Cover Extraction Using SPOT 5 Imagery. *Geocarto International*, **29**, 792-806. https://doi.org/10.1080/10106049.2013.848944

[18] Cheng, G., Zhou, P.C. and Han, J.W. (2016) Learning Rotation-Invariant Convolutional Neural Networks for Object Detection in VHR Optical Remote Sensing Images. *IEEE Transactions on Geoscience and Remote Sensing*, **54**, 7405-7415. https://doi.org/10.1109/EORSA.2016.7552845

[19] Long, Y., Gong, Y.P., Xiao, Z.F. and Liu, Q. (2017) Accurate Object Localization in Remote Sensing Images Based on Convolutional Neural Networks. *IEEE Transactions on Geoscience and Remote Sensing*, **55**, 2486-2498. https://doi.org/10.1109/TGRS.2016.2645610

[20] Nogueira, K., Penatti, O.A.B. and dos Santos, J.A. (2016) Towards Better Exploiting Convolutional Neural Networks for Remote Sensing Scene Classification. *Pattern Recognition*, **61**, 539-556. https://doi.org/10.1016/j.patcog.2016.07.001