Journal of Network Intelligence
Taiwan Ubiquitous Information
 ISSN 2414-8105 (Online)
Volume **8**, Number **4**, November **2023**

Remote Sensing Image Object Detection Based on Improved Sparse R-CNN

Li-Quan Zhao

Key Laboratory of Modern Power System Simulation and Control & Renewable Energy Technology, Ministry of Education
Northeast Electric Power University
Jilin 132012, China
zhao_liquan@163.com

Chun-Lu Chen

School of Electric Power Engineering
Northeast Electric Power University
Jilin 132012, China
neepuccl@163.com

Tie Zhong*

Key Laboratory of Modern Power System Simulation and Control & Renewable Energy Technology, Ministry of Education
Northeast Electric Power University
Jilin 132012, China
516104653@qq.com

Ying Cui

Zhuhai Power Supply Bureau
Guangdong Electric Power Corporation
Zhuhai 519000, China
cuiying794758706@126.com

Yan-Fei Jia

School of Electrical and Information Engineering
Beihua University
Jilin 132013, China
jia_yanfei@163.com

*Corresponding author: Tie Zhong

Received November 17, 2022, revised January 3, 2023, accepted March 17, 2023.

Abstract. *A more precise sparse region convolutional neural network algorithm is developed to increase the accuracy of remote sensing object recognition. Firstly, a new intersection over union method is proposed to solve the problem that the intersection over union used in sparse region convolutional neural network cannot well measure the offset between the ground-truth box and the proposal box. It introduces multiple geometric factors that are overlap area, center point distance, vertex distance between the ground-truth box and the proposal box, and diagonal distance of minimum enclosing of the ground-truth box and the proposal box and into the new intersection over union. The proposed intersection over union can measure the offset between the ground-truth box and the proposal box to optimal the proposal box. Secondly, to improve the detection precision of sparse region convolutional neural network, a dual branches dynamic instance interaction head is designed. The proposed dynamic interaction head consists of a fully connected branch and a convolution branch. The fully connected head is used for both bounding box regression and object classification. The convolution branch is only used for bounding box regression. Compared with original sparse region convolutional neural network, the average mAP and recall are 3.55% and 5.75% higher, and the average detection speed is only 0.85 FPS slower. The proposed method realizes balance between detection precision and detection speed.*

Keywords: Sparse region convolutional neural network, Deep convolution neural networks, Remote sensing object detection, Detection precision

1. **Introduction.** Remote sensing images are image data collected by imaging sensors on board satellites or spacecraft, which contain detailed information about the features of the observed scene. The aim of remote sensing image object detection is to identify the object class from the information contained in the remote sensing image. Remote sensing image object detection can provide data support for practical applications such as natural catastrophe assessment, urban planning, agriculture, forestry, transportation, and the military [1, 2]. The large height span of remote sensing image acquisition results in the same class of objects having different scale sizes on different remote sensing images. In addition, the remote sensing image contains a large number of objects and they have smaller sizes. Therefore, it is a larger challenge for remote sensing image detection [3].

Traditional object detection methods require an artificially designed feature extractor to extract features [4]. The robustness of the designed feature extractor and the quality of extracted features tend to be below. Advances in deep learning technology have led to the development of convolution neural networks that can be used to extract features without relying on an artificially designed feature extractor [5]. The deep convolution neural networks has been widely used in remote sensing image object detection [6], renewable energy forecast [7], intelligent transportation [8, 9] and human parameters recognition [10]. Many object detection methods based on deep convolution neural networks also have been used in remote sensing image object detection [11, 12]. They have better feature expression ability and higher detection accuracy than traditional detection methods. However, compared with other images, remote sensing images have a large number of small-sized objects due to the long distance of shooting. In addition, due to the influence of lighting and meteorological conditions, the image often has the problem of uneven brightness or even the existence of a bright surface and shadow of the object. The aforementioned issues make it challenging to identify objects in remote sensing images. Therefore, it is required to improve the performance of the remote sensing object detection method.

Sparse R-CNN (Sparse region convolutional neural network) is a special object detection method based on region convolutional neural network [13]. Compared with other methods based on R-CNN (region convolutional neural network) that require tens of thousands of proposal boxes, Sparse R-CNN requires only a small number of proposal boxes

(hundreds of proposal boxes) to achieve object detection and obtain better detection performance. Especially these sparse features of proposal boxes do not require interaction with all features in the whole image. Both features and proposal boxes of proposal boxes are learnable, so they can be optimized in the network together with other parameters. Besides, the Sparse R-CNN also further improves its performance by employing a cascade structure. It has a faster detection speed with acceptable detection accuracy than other region convolutional neural network serials methods. The performance of object detection method based on deep convolution neural networks directly affects the remote sensing images detection performance. To improve the remote sensing object detection accuracy based on Sparse R-CNN method, we proposed an improved sparse region convolutional neural network with higher accuracy for remote sensing image. Compared with original Sparse R-CNN, our proposed method has higher detection accuracy and acceptable detection speed.

The main highlight of this paper are as follows:

1. It propose a new IoU (Intersection over Union) method to better compute the differences in position and scale between ground-truth box and the proposal box in remote sensing image detection. We introduced the overlap area, center point distance, vertex distance between the ground-truth box and the proposal box, and diagonal distance of minimum enclosing of the proposal box and the ground-truth box into IoU. The proposed IoU can better optimize the proposal box and make it closer to the ground-truth box, which is useful to improve remote sensing image detection precision.

2. It propose a dual branches dynamic instance interaction head to improve detection boxes' precision of Sparse R-CNN. The proposed interaction head consists of a fully connected branch and a convolutional branch. The fully connected branch is mainly responsible for both bounding box regression and object classification. The convolution branch is only responsible for bounding box regression. The convolution layer is better at performing bounding box regression due to its structural properties, so the dual branches dynamic instance interaction head that has been presented has the potential to improve the remote sensing image's object detection precision.

2. **Related work.** The object detection techniques based on deep convolution neural network are divided into two categories: two-stage object detection technique and one-stage object detection technique. For two-stage object detection technique, the detecting process is split into two stages. In the first stage, the two-stage object detection technique generates region proposals. In the second stage, the two-stage object detection technique regresses the bounding box and candidate regions and classifies the objects. For one-stage object detection technique, it directly generates detection boxes and classifies the objects. Compared with the two-stage object detection technique, the one-stage object detection technique does not require to generate region proposals. Therefore, the one-stage object detection technique has a faster detection speed and lower detection precision.

The one-stage object detection technique is better suited for simple or fast detection tasks. Representative methods for the one-stage technique are YOLO serial method and SSD methods. Redmon et al. firstly proposed the YOLOv1 method in 2016 [14]. In 2017 and 2018, they also proposed the YOLOv2 and YOLOv3, respectively [15, 16]. Although Redmon et al. withdrew the research on the computer version, many scholars still try to improve the YOLO serial methods. Wu et al. [17] proposed the improved YOLOv5 algorithm based on YOLOv5. To improve the detection precision of smaller targets and the detection adaptability for images with different sizes, it introduce the multi-scale anchor mechanism that used in faster region convolutional neural network into the YOLOv5. The proposed method has complex network architecture, which is useful to

improve the detection precision of remote sensing objects. Wulamuet al. [18] designed a novel network with U shaped architecture and introduced atrous spatial pyramid pooling into network for detecting remote sensing objects. It has higher detection precision in road detection from remote sensing image. Qu et al. [19] improved the YOLOv3 model with an auxiliary network and applied it to remote sensing image object detection by using an image processing module to obtain a fixed size image and then adding a convolution block attention module between the backbone network and auxiliary network. The convolution block attention module can reduce the loss of key feature information. In addition, they introduced the DIoU method into loss function to reduce the convergence time of YOLOv3 and improve the detection precision. Finally, the adaptive feature fusion technique was also utilized to reduce inference overhead to improve object detection speed. The method has a higher mAP than the original YOLOv3 model on remotely sensed images. Liu et al. firstly proposed the SSD (single shot multiBox detector algorithm) method [20] and some improved methods are also used in the remote sensing image object. Lv et al. [21] proposed an improved single shot multi-Box detector algorithm for detecting smaller remote sensing objects. It combined the top-down structure of feature pyramid networks with the single shot multiBox detector algorithm and was able to extract more useful feature information. The method has higher remote sensing image detection precision than the original single shot multiBox detector algorithm without increasing the computational effort.

The size is smaller and the characteristics are not obvious for remote sensing object. It is a challenge to successfully detect objects from remote sensing image. The two-stage object detection technique has better performance in detection precision than the one-stage object detection technique. Therefore, we use the two-stage object detection technique to detect the remote sensing objects. The representative methods for the two-stage object detection technique are the R-CNN serials methods. Girshick et al. proposed the first R-CNN method. It firstly generated candidate regions and then extracts features from generated candidate regions. Secondly, it used the support vector machine method to classify features. In the end, each proposal box was corrected by bounding-box regression. In the region convolutional neural network method, all proposal boxes were retained, which caused a large computational effort and slow detection. In order to speed up the detection speed, Girshick et al. proposed fast R-CNN base on region convolutional neural network [22]. They first used the selective search method to filter the redundant proposal boxes and obtain the feature matrix by mapping the filtered boxes to the feature map. Secondly, the region of interest pooling operation was used to resize the feature matrix to the same scale. In the end, the feature matrix was spread into one dimension and imputed to the detection head to achieve object detection. The training time of fast region convolutional neural network was one over nine lower than region convolutional neural network and the detection speed of fast region convolutional neural network was 213 times faster than region convolutional neural network. Besides, the fast region convolutional neural network also had higher precision than region convolutional neural network. Ren et al. proposed a region proposal network (RPN) to design Faster region convolutional neural network [23]. The faster region convolutional neural network can be seen as a combination of fast region convolutional neural network and region proposal network. The region proposal network generated many anchors that were the regions of interest (RoI), and then the classification branch determined whether the regions of interest belonged to the foreground or background by softmax function. The regression branch corrected the anchor box to form proposal boxes. In the end, the optimal proposal boxes were obtained by confidence score and non-maximum suppression. When the coordinates of the anchor box generated in RPN are floating points, RoI pooling will quantize and round them. The approximate rounding operation will result in some deviation between the quantized

proposal box and the original proposal box, and the extracted features will not be exactly precise. For large-size objects, these quantitative losses occur on the edges of objects, so the impact is not significant. However, for small objects, the detection performance is greatly affected. He et al. proposed the Mask R-CNN [24] by designing a new pooling operation that is RoI Align. RoI Align avoids two rounding operations through a bilinear interpolation algorithm. Therefore, it greatly improves detection precision. Based on the Mask R-CNN, Wu et al. [25] proposed an enhanced Mask R-CNN. They replaced the feature pyramid network with a region proposal network to extract more effective feature information. It has better performance in fine-grained remote sensing object detection. On the same dataset, the enhanced Mask R-CNN is better than other methods. When the IoU threshold is small, the quality of the proposal boxes is usually poor, and the final detection boxes are prone to false detection. When the IoU threshold is too large, the number of positive samples is significantly reduced. It will aggravate the foreground-background imbalance problem [26] and the model is more prone to be overfitted. To overcome the problem, the Cascade R-CNN is designed by Cai et al. [27]. They used the cascading multiple same models with different IoU values. It enables the detector to focus on proposal boxes with IoU values in a certain range to improve object detection precision. Besides, an improved Faster R-CNN network is also designed by Chen et al. They combined a feature pyramid network (FPN) and deformable convolutional network (DCN) [28]. The FPN combined the structural information in shallow layers and semantic information in deep layers to obtain richer multi-level features. A deformable convolution network can enhance the effect of feature extraction. In addition, the shared convolutional layers enable the improved network to realize end-to-end training.

To improve the ship detection precision from remote sensing images, Wang et al. [29] proposed a high-performance ship detection method with low computation and efficient storage. The method preprocessed complex and diverse remote sensing image by using an accurate segmentation algorithm. It extracted ship object candidate regions using multivariate Gaussian distributions to improve the recall. The method has a few parameters and strong detection robustness against interference such as reefs and noise. Li et al. [30] designed a new detection network model for the detection of airport aircraft from remote sensing images based on depth transferable. It adopts hard example mining and skip-layer feature fusion to the training efficiency and improve the expression ability of the object of the detection network. Besides, it also introduces a cascaded region proposal network with soft-judgment non-maximum suppression into the network. It solves the over-fitting problem. In complex backgrounds, It also can quickly and accurately detect different airport objects. Tian et al. [31] designed a detection framework based on fast region convolutional neural network. To locate the boundary of large objects and avoid losing small objects, DetNet was used as the backbone network to fix the spatial resolution of the deep layer. It used convolutional projection to expand the bottleneck to increase the difference between input and output feature maps, and then fused the extracted scene features and regional features. To improve the regression performance, the framework also adopted a Cascade structure. The Cascade structure had multiple stages, and each stage had an independent classifier and repressor. The results obtained from the previous stage will be used as the input for the next stage to improve the detection accuracy stage by stage. To improve the performance of small target detection in remote sensing images, Courtrai et al. [32] proposed an improved SANET-SR method based on the SA-NET network. It firstly designed a super-resolution module to enhance the remote sensing image and then detected object from the enhanced image. Cui et al. [33] proposed a new anchor-free rotating ship detection framework which was called SKNet. It used the rotation coordinates of the ship as the key points of detection instead of using the traditional rectangular box

coordinates. The SKNet paid more attention to the central key points and shape of the ship object (including width, height, and rotation angle). It designed two customized modules: orthogonal pooling soft rotate non-maximum suppression and soft rotate non-maximum. Orthogonal pooling was used to improve the prediction accuracy of central key points and morphological dimensions, and soft rotate non-maximum suppression was used to effectively remove redundant rotating object detection frames. Hua et al. [34] proposed a cascaded convolution neural network framework for realizing real-time remote sensing object detection. It consisted of two fully convolution networks. The first network is the object accurate detection fully convolution network and the second network is the self-attention pre-screening fully convolution network. It has faster remote sensing object detection speed and an acceptable remote sensing object detection precision.

3. Improved Sparse R-CNN.

3.1. Proposed improved IoU.

For the object detection methods based on deep convolution neural network, the loss function generally consists of two parts: regression loss and classification loss. The IoU (Intersection over Union) loss is used as regression loss. It can quantitatively measure the degree of overlap between the ground-truth box and the proposal box to classify positive and negative samples. When the value of IoU between ground-truth box and the proposal box is greater than a threshold, we consider that the objects in both boxes belong to the same class. The proposal box will be classified as a positive sample. On the contrary, it will be judged as a negative sample.

The original IoU is the area ratio of the intersection region and concurrent region between the proposed box and the ground-truth box. When the proposed box and the ground-truth do not intersect, the value of IoU is zero, and the loss function cannot return the gradient, so the gradient descent cannot be carried out to optimize the proposal box. To solve the problem, Rezatofighi et al. [35] proposed GIoU(Generalized Intersection Over Union) by introducing minimum enclosing box between the ground-truth box and the proposal box. However, the GIoU is equivalent to IoU and they cannot well measure the difference between the two boxes when the proposal box is completely contained by the ground-truth box. To solve the problem, DIoU was proposed by introducing the distance of center points between two boxes [36]. However, the IoU, GIoU, and DIoU don't consider the size difference between the two boxes, so they cannot well measure the two boxes. In Figure 1, the blue boxes are the ground-truth box and the red boxes are the proposal box. They have the same center points, ground-truth boxes and different proposal boxes with different ratios of width and high in Figure 1. (a) and Figure 1. (b). It can be seen that the proposal box in Figure 1.(b) is more close to the ground-truth box than in Figure 1. (a). However, the IoU, GIoU, and DIoU are the same in Figure 1. (a) and Figure 1. (b). This means that the IoU, GIoU, and DIoU cannot reflect well the relationship between ground-truth box and the proposal box.

To overcome the problem, we propose a new IoU method which called multifactorial intersection over union (MIOU) and get new loss. The MIoU introduces the overlap area, center point distance, vertex distance between the ground-truth box and the proposal box, and diagonal distance of minimum enclosing of the ground-truth box and the proposal box. The MIoU loss is expressed as follows:

$$L_{MIoU} = 1 - \frac{A \cap B}{A \cup B} + \frac{|C - (A \cup B)|}{|C|} + \frac{d_0^2}{c^2} + \frac{D^2}{(1 - \frac{A \cap B}{A \cup B}) + D} \tag{1}$$

where the red box A, green box B, and box C are the proposal box, ground-truth box, and the minimum enclosing box of the proposal box and ground-truth box in Figure

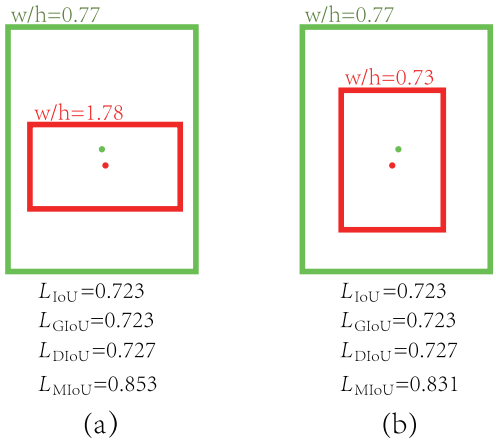

FIGURE 1. The losses of IoU, GIoU, DIoU and MIOU for the same ground-truth box and different proposal boxes.

2, respectively. d_0 is the center point distance between the ground-truth box and the proposal box. d_1 and d_2 are distances between non-adjacent vertices of the proposal box and ground-truth box, respectively. The c is the diagonal length of the C box. The D is expressed as follows:

$$D = \frac{\sqrt{3\sum_{i=0}^{2}(d_i - \overline{d})^2}}{3\overline{d}} \tag{2}$$

where $\overline{d}$ is the mean value of d_0, d_1 and d_2. The $\frac{A\cap B}{A\cup B}$ is used to measure the overlapping area between the A box and the B box. The closer the A box and B box are, the larger the overlapping area is, and the smaller the $1-\frac{A\cap B}{A\cup B}$ is, when A and B intersect. On the contrary, the larger the $1-\frac{A\cap B}{A\cup B}$ is. The $\frac{|C-(A\cup B)|}{|C|}$ is used to measure the proportion of areas except for $A\cup B$ in box C. The closer the A box and B box are, the larger the overlapping area is, and the smaller the $\frac{|C-(A\cup B)|}{|C|}$ is.When A and B not intersect, the $\frac{|C-(A\cup B)|}{|C|}$ still has a non-zero value and is useful to optimize the proposal box. The $\frac{d_0^2}{c^2}$ is used to measure the relative center distance between the proposal box and the ground-truth box. The closer the position and scale of A box are to B box, the smaller $\frac{d_0^2}{c^2}$ is.

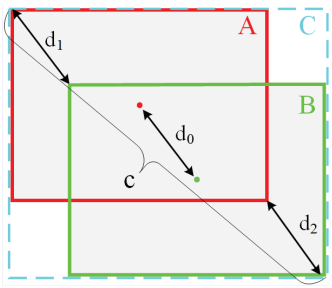

FIGURE 2. The geometric distance between bounding boxes.

The D in (2) is used to measure the size differences between box A and box B. The D equals zero when the two boxes have the same height and width. When the A box and B

box completely coincide, the MIoU loss equals zero. The proposed MIoU can measure the differences in size and location between the ground-truth box and the proposal box. In Figure 1, the values of MIoU loss are 0.853 and 0.831 for Figure 1. (a) and Figure 1. (b), respectively. The MIoU loss for Figure 1. (b) is smaller than Figure 1.(a). This means that the proposal box in Figure 1.(b) is closer to the ground-truth box in the aspect of size and location than the proposal box in Figure 1.(a). It looks like this is right from Figure 1. However, the IoU, GIoU, and DIoU have the same values in Figure 1.(a) and Figure 1.(b). They cannot well measure the difference between the proposal box and the ground-truth box. This will affect detection precision.

3.2. **Dual branches dynamic instance interaction head.** Sparse R-CNN uses a single fully connected head as the detection head in the dynamic instance interaction head. This fully connected layer is not only responsible for the regression of bounding boxes, but also for the classification of the object. Each output of the fully connected layer can be viewed as the result of multiplying each neuron in the previous layer by the corresponding weight coefficients and adding a bias. The fully connected layer can integrate the features of the same class with different positions in the image, so it is more suitable for classification. The fully connected layer ignores the spatial structure of the detected objects, so it is not sensitive to the location information of the object. Therefore, the location information of an object is easily lost in the bounding box regression. Compared with the convolution layer, it is more suitable for classification.

To improve the detection precision of Sparse R-CNN, we propose a dual branches dynamic instance interaction head. The convolution layer implements two-dimensional mutual correlation convolution through convolution kernels. The convolution kernel moves from the top leftmost part of image to the bottom rightmost of image according by moving form the left to right. The perceptual fields in the window are multiplied by elements with the convolution kernel array and summed, and finally, the elements at the corresponding positions in the output array are obtained. So convolution layer is more sensitive to the position information of the object. Therefore, we use the convolution layer to construct the regression detection head. Our proposed deal branches dynamic instance interaction head (DDIIH) is shown in Figure 3. The proposed DDIIH consists of a feature interaction module and detection head module.

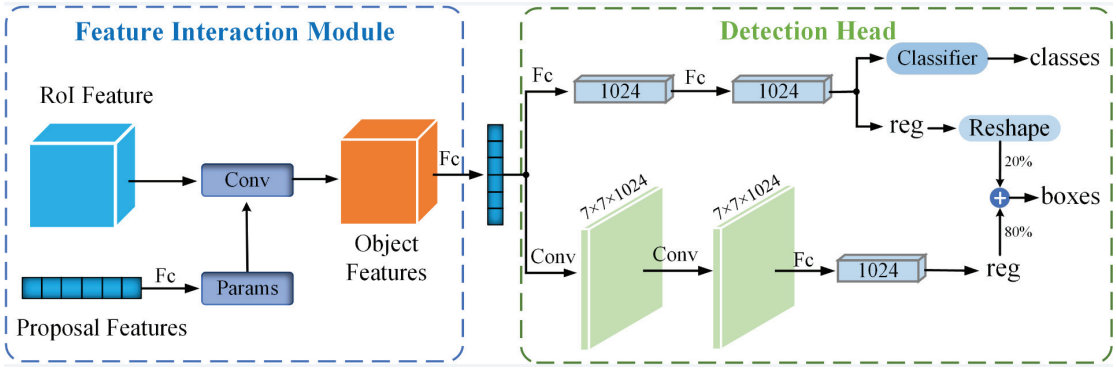

FIGURE 3. The structure of DDIIH.

The RoI feature is the region of interest feature that is extracted from all proposal boxes. The proposal feature is the feature set of each proposal box in different locations. The dynamic convolution parameters are generated by proposal features according to the location and size of different proposal boxes. In dynamic convolution, the convolution

kernels are variable, and their size and step of movement are determined by the dynamic convolution parameters. The RoI feature and proposal features are interacted by such a dynamic convolution process. The purpose of feature interaction is to reason about the feature relationships between local regions and regions of interest to obtain the object features of each class. The object features are used as the input of the detection head. The detection head composed fully connected branch and a convolution branch. The fully connected branch consists of two fully connected layers, a classifier, and a regressor. The number of neurons in each fully connected layer is 1024. The fully connected layer is more suitable for the classification task, so we mainly use this branch for classification, and randomly select 20% of the regression information to be used for generating the detection boxes with 80% of regression information obtained from the other branch. Dimension transformation is used to adjust the dimension of the regression information obtained by the fully connected branch to make it have the same dimension as regression information obtained by the convolution branch. In the end, the regression information obtained by two branches is fused to generate detection boxes.

The convolution branch consists of two convolution layers with 1024 channels, a fully connected layer, and a regressor. The convolution layers are mainly used for extracting the position information of bounding boxes, and the fully connected layer is mainly used for the fusion and mapping of regression information. After fusing the regression information obtained by the convolution branch and the fully connected branch, the information of the detection boxes is obtained which is used to generate the final detection boxes. Since the convolution layer is more suitable for extracting location information and the fully connected layer is more suitable for extracting class information. The proposed dual branches dynamic instance interaction head can extract more effective and richer location information, improving the precision of detection.

We design the improved Sparse R-CNN through multiple DDIIH cascades. The structure of improved Sparse R-CNN is shown in Figure 4.We take the Resnet and feature pyramid network as the backbone of the feature extraction network, which can obtain a multi-level feature map. Then 300 proposal boxes are generated on the feature map through Gaussian initialization. The RoI Align module extracts RoI feature and proposal features by pooling operation. RoI feature and proposal features are used as the inputs of feature interaction in DDIIH. Feature interaction is realized by dynamic convolution in the feature interaction module of DDIIH to obtain the object features. Finally, the detection module of DDIIH implements the classification and regression tasks at the current stage according to the object features. The improved Sparse R-CNN contains five cascaded RoI Align modules and DDIIH modules. The detection boxes obtained in the current DDIIH module will also become the proposal boxes in the next cascaded RoI Align module. Such a cascade structure can improve the quality of the proposal boxes stage by stage.

3.3. **Loss function of improved Sparse R-CNN.** The total loss function of the improved Sparse R-CNN is composed of classification loss and regression loss. Expressed as the followings:

$$Loss = \beta_1 * L_{cls} + \beta_2 * L_{reg} \tag{3}$$

where L_{cls} is the classification loss, L_{reg} is the regression loss, β_1 and β_2 are the weights of classification loss and regression loss in the total loss. The expression of the classification loss function is as follows:

$$L_{cls} = -\alpha_t (1 - p_t)^{\gamma} \log(p_t) \tag{4}$$

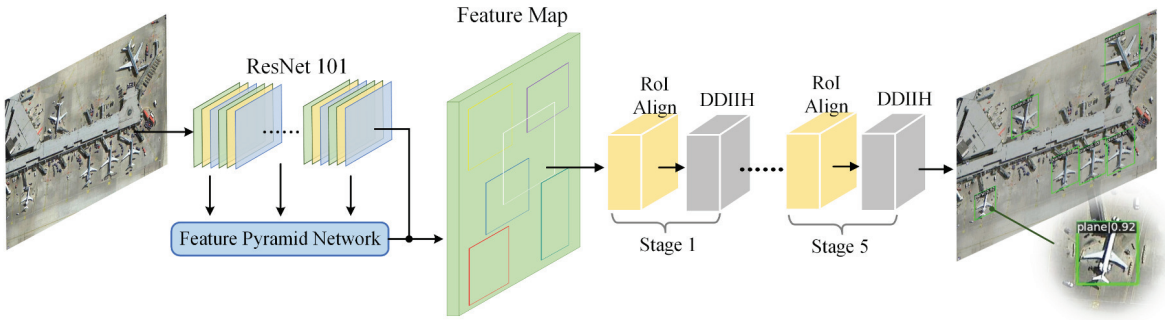

FIGURE 4. The complete structure of improved Sparse R-CNN.

where α_t is the weight of cross-entropy, which is used to adjust the proportion of positive and negative samples to solve the imbalance between them. The $(1-p_t)^\gamma$ is the modulation coefficient, where p_t is the probability of a class, γ is the focusing parameter which can make the loss focus more on difficult and misclassified samples. The regression loss function can be expressed as:

$$L_{reg} = 0.2 * L_{fc}^{reg} + 0.8 * L_{conv}^{reg} \tag{5}$$

Where $Ł_{fc}^{reg}$ and $Ł_{conv}^{reg}$ are the regression losses obtained from the fully connected branch and the convolution branch, respectively. The L_{fc}^{reg} is expressed as followings:

$$L_{fc}^{reg} = 1 - \frac{A \cap B}{A \cup B} + \frac{|C - (A \cup B)|}{|C|} + \frac{d_0^2}{c^2} + \frac{D^2}{(1 - \frac{A \cap B}{A \cup B}) + D} \tag{6}$$

where A is the ground-truth box and B is the detection box obtained from the fully connected branch. The L_{conv}^{reg} has the same expression L_{fc}^{reg}. The only difference is that the B is detection box obtained from the convolution layer branch. The (6) is also our proposed MIoU loss that is given in (1).

4. Simulation and Discussion.

4.1. Datasets and evaluation metrics.

To expand the scale of the dataset to better train the model, we merge two remote sensing image datasets that are NWPU_VHR_10 and HRSC_2016, into a dataset in PASCAL VOC format. The merged remote sensing image dataset contains 12 classes with a total of 3000 images, and the ratio of the training set, validation set, and test set is 6:2:2. This paper mainly uses mAP (mean average precision) and recall to measure the precision of the algorithm, and FPS (frame per second) to measure the speed of the algorithm. Precision and recall are represented as follows:

$$precision = \frac{TP}{TP + FP} \tag{7}$$

$$recall = \frac{TP}{TP + FN} \tag{8}$$

where TP, FP and FN are the the number of true positive samples, false positive samples and false negative examples,respectively. In practice, the values of TP, FP, and FN correspond to the number of correct, false, and missing detection, respectively.

All experiments are implemented on the MMDetection platform. All methods use resnet50 + FPN and resnet101 + FPN as backbones to extract features, respectively. During the training, we use random flipping to expand the dataset. We use the Adamw

optimizer and apply the linear preheating strategy at the same time. The hyper parameters are consistent with those in the original Sparse R-CNN, that is, the weights of each loss item in the loss function are 2 and 5,respectively.

4.2. **Module analysis.** In the improved Sparse R-CNN, the IoU algorithm and the structure of the dynamic instance interaction head have been improved. The relevant calculation details and model structure have changed. Therefore, it requires conducting parameter experiments to determine a set of appropriate parameters again to ensure the performance of the model. Since the number of proposal features and proposal boxes are equal, we are only required to determine the number of proposal boxes and cascading modules.

The number of suggestion boxes generated by initialization has a great impact on the detection performance of Sparse R-CNN. When the number of proposal boxes is insufficient, the detection precision is often not ideal, and the performance improvement of the detector is easy to reach the bottleneck. It is suggested that the detection precision can be guaranteed when the number of proposal boxes is too large, but the training time of the model will increase significantly, and the detection speed is also not ideal. We set the number of cascaded modules to 6. The mAP, FPS, and training time for different numbers of proposal boxes are shown in Table 1. As the number of proposal boxes increases, the mAP and training time gradually increase, and the FPS gradually decreases. When the number of proposal boxes is equal to 300, the mAP increases slowly. Therefore, the number of proposal boxes is set to 300.

Table 1. Impact of the number of proposal boxes on performance.

Proposal boxes	mAP(%)	FPS	Training time(h)
100	67.9	19.7	35.8
200	74.6	16.2	42
300	78.7	13.2	53
400	79.1	9.8	48.7
500	79.7	8.2	60.5

We test our proposed method for different numbers of cascaded modules with the same number of proposal boxes which is 300. The results are shown in Table 2. We can see that the mAP increases with the increase of the number of cascaded modules. The larger the number of the cascaded module is, the smaller the FPS is and the longer the training time is. When the number of cascaded modules is equal to 5, the mAP increases slowly. Therefore, we set the number of cascaded modules to 5 to balance the precision and speed. According to Table 1 and Table 2, the number of proposal boxes and cascaded modules are set to 300 and 5, respectively.

Table 2. Impact of the number of Cascaded modules on performance.

Cascaded module	mAP(%)	FPS	Training time(h)
2	44.4	24.7	12.6
3	61.0	22.2	20.4
4	72.7	18.6	28.3
5	77.9	14.9	38.2
6	78.7	13.2	53.0
10	79.2	7.0	96.4

4.3. **Ablation study.** To test the performance of the different IoU loss, we take IoU loss, GIoU loss, and MIoU loss as the regression loss of Sparse R-CNN, respectively. The relationship between the mAP and epoch for Sparse R-CNN based on IoU, GIoU and MIoU are shown in Figure 5. Although the mAP fluctuates with the increase of the number of the epoch, the overall trend of mAP gradually increases for Sparse R-CNN methods based on different IoU loss. The mAP of Sparse R-CNN based on our proposed MIoU is still larger than the Sparse R-CNN based on IoU and Sparse R-CNN based on GIoU for the same epoch when the number of the epoch is greater than 10. All curves tend to be smooth when the number of the epoch is greater than 30. When the number of the epoch is 36, the mAP of Sparse R-CNN based on IoU loss, GIoU loss, and MIoU loss are 72.4%, 74.1%, and 75.6%, respectively. The Sparse R-CNN based on MIoU loss has the highest mAP than others. This shows the Sparse R-CNN based on proposed MIoU has higher precision than Sparse R-CNN based on IoU loss and GIoU loss.

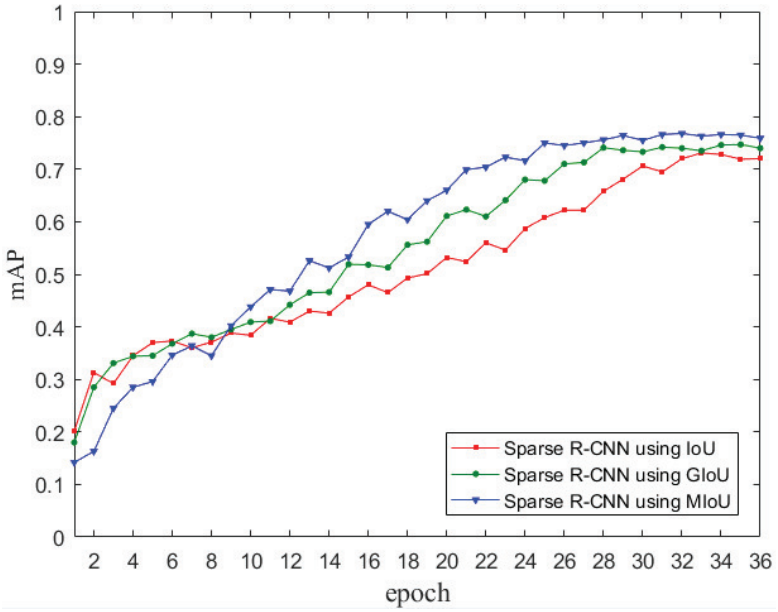

FIGURE 5. Relationship of mAP and epoch for Sparse R-CNN based different IoU losses.

We use ResNet101+FPN as the backbone, and the max number of the epoch is 36 for Sparse R-CNN and our proposed Sparse R-CNN. We select three different images from the test dataset to test the original Sparse R-CNN that uses GIoU loss as regression loss, Sparse R-CNN+MIoU, and Sparse R-CNN + MIoU+ DDIIH. The detection results of different methods are shown in Figure 6. For the first picture, the original Sparse R-CNN missed the detection of three planes that are marked in the red circle. Sparse R-CNN+MIoU missed one plane that is also marked in a red circle. Our complete proposed network that is Sparse R-CNN + MIoU+DDIIH successfully detects all objects. For the second picture, the original Sparse R-CNN mistakenly detected irrelevant objects as a baseball court. The Sparse R-CNN+MIoU and our complete proposed network successfully detect all objects. For the third picture, the Sparse R-CNN, Sparse R-CNN+MIoU, and our complete proposed network mistakenly detected three objects, two objects and one object, respectively. All error detected objects also are marked in red circles. Based on the detection results, our complete network has a better performance in precision than the original Sparse R-CNN.

When we use ResNet101+FPN as backbone, the values of mAPfor Fast R-CNN, SSD, RetinaNet, YOLOv4, Faster R-CNN, Sparse R-CNN, Cascade R-CNN and Our Method are 58.8%, 63.4%, 69.6%, 72.5%, 73.4%, 74.1%, 80.5% and 77.9%, respectively. Recall values for Fast R-CNN, SSD, RetinaNet, YOLOv4, Faster R-CNN, Sparse R-CNN, Cascade R-CNN and our method are 66.8%, 72.4%, 73.2%, 80.4%, 81.7%, 81.5%, 90.4% and 86.3%, respectively. The Cascade R-CNN still has the highest mAP and recall, followed by our method. The FPS for Fast R-CNN, SSD, RetinaNet, YOLOv4, Faster R-CNN, Sparse R-CNN, Cascade R-CNN, and our method are 12.6, 15.5, 14.6, 19.3, 14.7, 15.6, 5.1, 14.9, respectively. The YOLOv4 has the fastest detection speed, followed by RetinaNet and SSD.Based on Table 4, the Cascade R-CNN still has the highest mAP and recall, followed by our method. Compared with Cascade R-CNN, the mAP and recall of our method are 1.4% and 2.0% lower for ResNet50+FPN backbone and 2.6% and 4.1% lower for ResNet101+FPN backbone, respectively. However, compared with Cascade R-CNN, the detection speed of our method is about two times faster. Compared with Sparse R-CNN, the proposed method still has higher mAP and recall for both backbones.

(a) (b) (c)

FIGURE 6. Object detection results for original Sparse R-CNN and improved methods. (a) Original Sparse R-CNN;(b) Sparse R-CNN+MIoU;(c) Sparse R-CNN + MIoU+ DDIIH.

We also use all images in the test dataset to test our proposed method and original method. The detection precision and speed indxes are shown in Table 3. The mAP of the Sparse R-CNN, Sparse R-CNN+MIoU, and Sparse R-CNN+MIoU+ DDIIH are 74.1%, 75.6%, and 77.9%, respectively. Compared with the original Sparse R-CNN which is the Sparse R-CNN+GIoU, the mAP of our complete network is 3.8% higher. The recall of the Sparse R-CNN, Sparse R-CNN+MIoU, and Sparse R-CNN+MIoU+ DDIIH are 81.5%, 82.7%, and 86.3%, respectively. Compared with the original Sparse R-CNN, the recall of our final network is 4.8% higher. The FPS of the Sparse R-CNN, Sparse R-CNN+MIoU, and Sparse R-CNN+MIoU+ DDIIH are 15.6, 15.2, and 14.9, respectively. Compared with the Sparse R-CNN, the FPS of our complete network is only reduced by 0.7. Based on the above analysis, the proposed MIoU and DDIIH are effective, and the complete network has the best performances in mAP and recall. The structures of the proposed MIoU and DDIIH are more complex than the original Sparse R-CNN, so it has lower FPS than the original Sparse R-CNN.

Table 3. Performance of Sparse R-CNN, Sparse R-CNN+MIoU, and Sparse R-CNN+MIoU+DDIIH.

MIoU	DDIIH	mAP(%)	Recall(%)	FPS
-	-	74.1	81.5	15.6
√	-	75.6	82.7	15.2
√	√	77.9	86.3	14.9

4.4. **Performance comparison.** We also compare our complete Sparse R-CNN with SSD, RetinaNet, YOLOv4, Fast R-CNN, Faster R-CNN, Cascade R-CNN and Sparse R-CNN. We select an image that contains six planes from the test dataset to test the performance of different methods. The ResNet101+FPN is used as the backbone of different detection methods. The detection results of different methods are shown in Figure 7. Although our proposed method, Cascade R-CNN and Sparse R-CNN successfully detected all planes, the detection boxes do not completely enclose the largest plane and the confidence score of the largest plane is lower for Sparse R-CNN. The RetinaNet and SSD only detect five planes. They miss the largest plane. The Faster-RCNN and YOLOv4 detect seven objects. The Faster R-CNN detects lane tail as a new plane. The YOLOv4 detects other object as an aircraft carrier. The Fast R-CNN misses and mistakenly detects three planes and has redundant detection boxes. In Figure 7, our proposed method and Cascade R-CNN have higher detection precision than other methods.

We also use ResNet50+FPN and ResNet101+FPN as the backbone of different detection methods. The performance indexes of different methods with different backbone networks are shown in Table 4. When we use ResNet50+FPN as a backbone, the mAPs of Fast R-CNN, SSD, RetinaNet, YOLOv4, Faster R-CNN, Sparse R-CNN, Cascade R-CNN, and our method are 56.0%, 59.3%, 68.1%, 71.3%, 71.9%, 72.5%, 77.2% and 75.8%, respectively. The values of recalls for Fast R-CNN, SSD, RetinaNet, YOLOv4, Faster R-CNN, Sparse R-CNN, Cascade R-CNN, and our method are 65.3%, 70.6%, 69.8%, 76.5%, 79.6%, 78.0%, 86.7% and 84.7%, respectively. The Cascade R-CNN has the highest mAP and recall, followed by our method. The FPS for Fast R-CNN, SSD, RetinaNet, YOLOv4, Faster R-CNN, Sparse R-CNN, Cascade R-CNN, and our method are 17.4, 18.2, 18.3, 24.2, 17.8, 18, 8.3, 17, respectively. The YOLOv4 has the fastest detection speed, followed by RetinaNet and SSD.

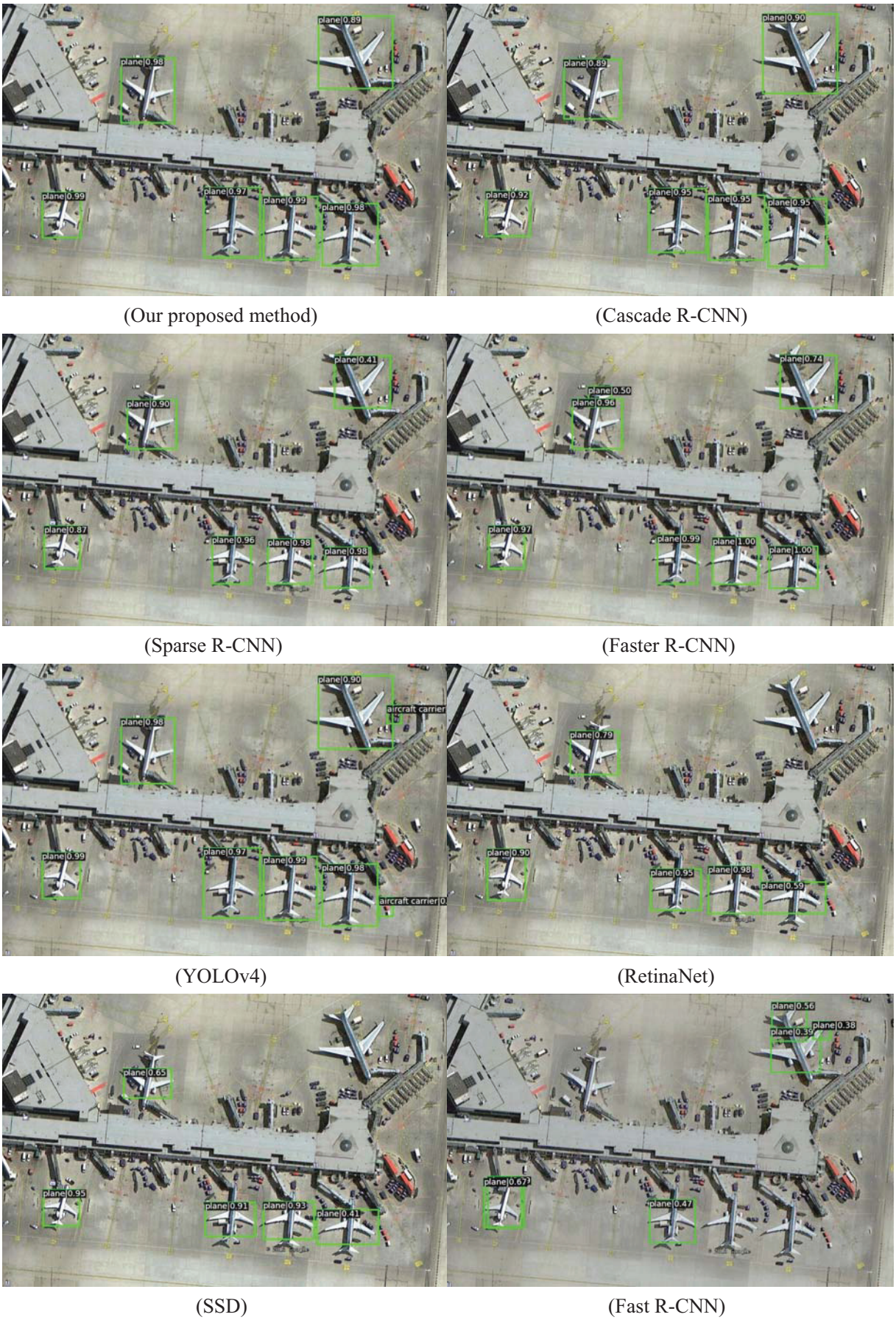

FIGURE 7. Detection results of different methods.

TABLE 4. Performance comparison between improved Sparse R-CNN and other methods.

Method	Backbone	mAP(%)	Recall(%)	FPS
Fast R-CNN	ResNet50+FPN	56.0	65.3	17.4
Fast R-CNN	ResNet101+FPN	58.8	66.8	12.6
SSD	ResNet50+FPN	59.3	70.6	18.2
SSD	ResNet101+FPN	63.4	72.4	15.5
RetinaNet	ResNet50+FPN	68.1	69.8	18.3
RetinaNet	ResNet101+FPN	69.6	73.2	14.6
YOLOv4	ResNet50+FPN	71.3	76.5	24.2
YOLOv4	ResNet101+FPN	72.5	80.4	19.3
Faster R-CNN	ResNet50+FPN	71.9	79.6	17.8
Faster R-CNN	ResNet101+FPN	73.4	81.7	14.7
Sparse R-CNN	ResNet50+FPN	72.5	78.0	18.0
Sparse R-CNN	ResNet101+FPN	74.1	81.5	15.6
Cascade R-CNN	ResNet50+FPN	77.2	86.7	8.3
Cascade R-CNN	ResNet101+FPN	80.5	90.4	5.1
Our Method	ResNet50+FPN	75.8	84.7	17.0
Our Method	ResNet101+FPN	77.9	86.3	14.9

5. **Conclusions.** This paper proposes improved Sparse R-CNN for remote sensing object detection. It proposes an improved IoU by introducing the multiple geometric factors that are overlap area, center point distance, vertex distance between the ground-truth box and the proposal box, and diagonal distance of minimum enclosing of the ground-truth box and the proposal box into IoU. Compared with the IoU and GIoU, the proposed IoU can better measure the offset between the proposal box and the ground-truth box to improve the quality of the detection box. Besides, it also designs a dual branches dynamic instance interaction head. It consists of a fully connected branch and a convolution branch. The fully connected branch is mainly used for classification and the convolution branch is completely used for bounding box regression. The fully connected layers are suitable for classification and the convolution layers are suitable for the location. Therefore, the proposed dual branches dynamic instance interaction head can improve the precision of remote sensing object detection.

When all methods use the ResNet50+FPN as a backbone, compared with Cascade R-CNN that has the highest detection mAP and recall, the mAP and recall of our method are 1.4% and 2.0% lower, the average detection speed of our method is about 2.1 times faster. Compared with the original Sparse R-CNN, the mAP and recall of our method are 3.3% and 6.7% higher, and the average detection speed is only 1.0 FPS slower. When all methods use the ResNet101+FPN as a backbone, compared with the Cascade R-CNN that has the highest detection mAP and recall, the mAP and recall of our method are 2.6 % and 4.1% lower, the average detection speed of our method is about 3 times faster. Compared with original Sparse R-CNN, the mAP and recall of our method are 3.8% and 4.8% higher, and the average detection speed is only 0.7 FPS slower. On the whole, compared with the Cascade R-CNN, our proposed method has lower detection accuracy and faster detection speed. Compared with Sparse R-CNN, our proposed method has higher detection accuracy and slower detection speed. The proposed method realizes the balance between detection accuracy and detection speed.

In future work, to improve the detection accuracy of remote sensing image in foggy weather, we will consider using remote sensing image dehazing method based on generative adversarial network to enhance the remote sensing image.

Acknowledgment. This work is partially supported by the Jilin Provincial Department of Education Project(JJKH20230125KJ)

REFERENCES

[1] L. Zhen, Z. Bin, and Y. -X. Zhu, "High resolution representation-based Siamese network for remote sensing image change detection," *IET Image Processing*, vol. 16, no. 9, pp. 2506-2517, 2022.

[2] Z. -Q. Zhang, Z. -L. Xiong, B. Zhang, Y. -X. Yang, and E. -K. Fu, "Detection for Small Target Ship in Remote Sensing Image Based on Super Resolution Reconstruction Technology," *Journal of Northeast Electric Power University*, vol. 42, no. 2, pp. 33-40, 2022.

[3] A. -P. Shaik, M. -K. Manoharan, A. -K. Pani, R. -R. Avala, and C. -M. Chen, "Gaussian Mutation-Spider Monkey Optimization (GM-SMO) Model for Remote Sensing Scene Classification," *Remote Sensing*, vol. 14, no. 24, 6279, 2022.

[4] F. -Q. Zhang, T. -Y. Wu, J. -S. Pan, G. -Y. Ding, and Z. -Y. Li, " Human Motion Recognition Based on SVM in VR Art Media Interaction Environment," *Human-centric Computing and Information Sciences*, vol. 9, 40, 2019.

[5] K. Wang, C. -M. Chen, M. -S. Hossain, G. Muhammad, S. Kumar and S. Kumari, " Transfer reinforcement learning-based road object detection in next generation IoT domain," *Computer Networks*, vol. 193, 108078, 2021

[6] K. Zhang, C. Hu, and H. Yu, "Remote Sensing Image Land Classification Based on Deep Learning," *Scientific Programming*, vol. 2021, no. 14, pp. 6203444.1-6203444.12, 2021.

[7] L. -Z. Long, B. Xiao, and L. -G. Sun, "Conditional Depth Convolution Generation of Confrontation Network Method for Scenery Output Scenario Generation," *Journal of Northeast Electric Power University*, vol. 41, no. 6, pp. 90-99, 2021.

[8] F. -Q. Zhang, T. -Y. Wu, Y. Wang, R. Xiong, G. -Y. Ding, P. Mei, and L. -Y. Liu, "Application of quantum genetic optimization of LVQ neural network in smart city traffic network prediction," *IEEE Access*, vol. 8, pp. 104555-104564, 2020.

[9] S. -M. Zhang, X. Su, X. -H. Jiang, M. -L. Chen, and T. -Y. Wu, " A traffic prediction method of bicycle-sharing based on long and short term memory network," *Journal of Network Intelligence*, vol. 4, no. 2, pp. 17-29, 2019

[10] M. -G. Tang, G. Li , P. -L. Liu, R. -W. Zhao, Z. -C. Liu, X. -Z. WU, and J. Gao, "Human Parameters Recognition with Infrared Images Based on Deep Learning Method," *Journal of Northeast Electric Power University*, vol. 42, no. 4, pp. 18-27, 2022

[11] D. Yu, and S. Ji, "A New Spatial-Oriented Object Detection Framework for Remote Sensing Images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, 4407416, 2021.

[12] H. Chen, Z. Qi and Z. Shi, "Remote Sensing Image Change Detection With Transformers," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, 5607514, 2022.

[13] P. -Z. Sun, R. -F Zhang, Y. Jiang, T. Kong, C. -F Xu, and W. Zhan, "Sparse R-CNN: End-to-End object detection with learnable proposals," *in IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, 2021, pp. 14454-14463.

[14] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," *in IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2016, pp. 779-788.

[15] J. Redmon, and A. Farhadi, "YOLO9000: Better, faster, stronger," *in IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2017, pp. 7263-7271.

[16] J. Redmon, and A. Farhadi, "YOLOv3: An incremental improvement," *arXiv preprint*, 2018. [Online]. Available: https://arxiv.org/abs/1804.02767.

[17] W. -T. Wu, H. Liu, L. -L. Li, Y. -L. Long, and X. -D. Wang, "Application of local fully Convolutional Neural Network combined with YOLO v5 algorithm in small target detection of remote sensing image," *Plos One*, vol. 16, no. 10, e0259283, 2021.

[18] A. Wulamuet, Z. -X. Shi, D. -Z. Zhang, and Z. -Y. He, "Multiscale Road Extraction in Remote Sensing Images," *Computational Intelligence and Neuroscience*, vol. 2019, 2373798, 2019.

[19] Z. Qu, F. Zhu, and C. Qi, "Remote Sensing Image Target Detection: Improvement of the YOLOv3 Model with Auxiliary Networks," *Remote Sensing*, vol. 13, no. 16, 3908, 2021.

[20] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, and S. Reed, "SSD: Single shot multibox detector," *in International European Conference on Computer Vision*. IEEE, 2016, pp. 21-37.
[21] S. -D. Lv, L. Zhu and W. Wang, "Improving SSD for detecting small target in Remote Sensing Image," *in Chinese Automation Congress*. CAC, 2020, pp. 567-571.
[22] R. Girshick, "Fast R-CNN," *in International Conference on Computer Vision (ICCV)*. IEEE, 2015, pp. 1440-1448.
[23] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137-1149, 2017.
[24] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," *in International Conference on Computer Vision*. IEEE, 2017, pp. 2961-2969.
[25] Q. Wu, D. Feng, C. Cao, X. Zeng, Z. Feng, and J. Wu, "Improved mask R-CNN for aircraft detection in remote sensing images," *Sensors*, vol. 21, no. 8, 2618, 2021.
[26] K. Oksuz, B. -C. Cam, S. Kalkan, and E. Akbas, "Imbalance problems in object detection: A review," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 10, pp. 3388-3415, 2020.
[27] Z. Cai, and N. Vasconcelos, "Cascade R-CNN: High Quality Object Detection and Instance Segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 5, pp. 1483-1498, 2019.
[28] X. Chen, Q. Zhang, J. Han, X. Han, Y. Liu, and Y. Fang, "Object detection of optical remote sensing image based on improved faster RCNN," *in Fifth International Conference on Computer and Communications (ICCC)*. IEEE, 2019, pp. 1787-1791.
[29] N. Wang, B. Li, X. Wei, Y. Wang, and H. Yan, "Ship detection in spaceborne infrared image based on lightweight CNN and multisource feature cascade decision," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 5, pp. 4324-4339, 2020.
[30] S. Li, Y. Xu, M. Zhu, S. Ma, and H. Tang, "Remote sensing airport detection based on end-to-end deep transferable convolutional neural networks," *IEEE Geoscience and Remote Sensing Letters*, vol. 16, no. 10, pp. 1640-1644, 2019.
[31] Z. Tian, W. Wang, R. Zhan, Z. He, J. Zhang, and Z. Zhuang, "Cascaded detection framework based on a novel backbone network and feature fusion," *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 12, no. 9, pp. 3480-3491, 2019.
[32] L. Courtrai, M. -T. Pham, and S. Lefèvre, "Small object detection in remote sensing images based on super-resolution with auxiliary generative adversarial networks," *Remote Sensing*, vol. 12, no. 19, 3152, 2020.
[33] Z. Cui, J. Leng, Y. Liu, T. Zhang, P. Quan, and M. Zhao, "SKNet: Detecting rotated ships as keypoints in optical remote sensing images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 59, no. 10, pp. 8826-8840, 2021.
[34] X. Hua, X. Wang, T. Rui, H. Zhang, and D. Wang, "A fast self-attention cascaded network for object detection in large scene remote sensing images," *Applied Soft Computing*, vol. 94, no. 1, 106495,2020.
[35] H. Rezatofighi, N. Tsoi, J. Gwak, A. Sadeghian, I. Reid, and S. Savarese, "Generalized intersection over union: a metric and a loss for bounding box regression," *in International Conference on Computer Vision and Pattern Recognition*. IEEE, 2019, pp. 658-666.
[36] Z. Zheng, P. Wang, W. Liu, J. Li, R. Ye, and D. Ren, "Distance-iou loss: faster and better learning for bounding box regression," *in Association for the Advancement of Artificial Intelligence Conference on Artificial Intelligence*. AAAI, 2020, pp. 12993-13000.