IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received July 6, 2019, accepted July 26, 2019, date of publication August 2, 2019, date of current version August 16, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2932731

An Improved Faster R-CNN for Small Object Detection

CHANGQING CAO, BO WANG, WENRUI ZHANG, XIAODONG ZENG, XU YAN, ZHEJUN FENG, YUTAO LIU, AND ZENGYAN WU
School of Physics and Optoelectronic Engineering, Xidian University, Xi'an 710071, China

Corresponding author: Bo Wang (wangbo599026@sina.com)

This work was supported in part by the Central Universities under Grant K5051399208, in part by the Major Instruments of the Ministry of Science and Technology under Grant 2012YQ12004702, and in part by the 111 Project under Grant B17035.

ABSTRACT With the increase of training data and the improvement of machine performance, the object detection method based on convolutional neural network (CNN) has become the mainstream algorithm in field of the current object detection. However, due to the complex background, occlusion and low resolution, there are still problems of small object detection. In this paper, we propose an improved algorithm based on faster region-based CNN (Faster R-CNN) for small object detection. Using the two-stage detection idea, in the positioning stage, we propose an improved loss function based on intersection over Union (IoU) for bounding box regression, and use bilinear interpolation to improve the regions of interest (RoI) pooling operation to solve the problem of positioning deviation, in the recognition stage, we use the multi-scale convolution feature fusion to make the feature map contain more information, and use the improved non-maximum suppression (NMS) algorithm to avoid loss of overlapping objects. The results show that the proposed algorithm has good performance on traffic signs whose resolution is in the range of (0, 32], the algorithm's recall rate reaches 90%, and the accuracy rate reaches 87%. Detection performance is significantly better than Faster R- CNN. Therefore, our algorithm is an effective way to detect small objects.

INDEX TERMS CNN, faster R-CNN, small object detection.

I. INTRODUCTION

Visual information plays more than 90% of human cognition [1], and various types of optoelectronic imaging devices are also widely used in areas closely related to human production and life. With the continuous development of machine learning methods, computer vision has successfully implemented the processing of images and other information in many industries [2]–[4]. Object detection, as the core research problem of computer vision, has attracted more and more attention by researchers [5], [6]. The object detection usually includes two steps, searching the object in the image, and then using the bounding box to locate the object. In recent years, the convolutional neural network achieved excellent performance with object detection [7], [8].

In 2012, Alex and Hinton used the convolutional neural network-based AlexNet [2] to achieve great success in the ImageNet dataset [9], which has set off a wave of convolutional neural network applications in the field of computer vision. In 2014, Ross Girshick *et al.* applied convolutional neural networks to object detection and proposed the region-based CNN (R-CNN) [10], the algorithm first uses selective search algorithm to generate a series of region proposals for each input picture, and then uses a convolutional neural network to extract the features of these regions and train the support vector machine (SVM) for classification. R-CNN is computationally intensive and has limitations on the size of the input image, so Kaiming He *et al.* proposed SPP-net [11], which solves the problem that size of input images are limitation by using spatial pyramid pooling, making the speed of SPP-net several times higher than the R-CNN. In 2015, Ross Girshick proposed fast region-based CNN (Fast R-CNN) [12], which uses two parallel different fully connected layers to complete the classification and positioning tasks respectively, and solves the need to train SVM separately in R-CNN and SPP-net algorithms, and the drawback of occupying a large amount of storage space. Fast R-CNN still uses the selective search algorithm to extract region proposals with fixed position and size, the whole algorithm is not an end-to-end network, the back-propagation

The associate editor coordinating the review of this manuscript and approving it for publication was Yan-Jun Liu.

algorithm cannot improve the extraction process of region proposals. S. Ren *et al*. proposed the Faster R-CNN [5] algorithm, which uses neural networks to extract region proposal networks (RPN), shares the parameters of the convolutional layer, thus greatly improving the speed of object detection. But Faster R-CNN is not applicable to perform object detection for other image datasets directly. Furthermore, it is difficult for Faster R-CNN to identify objects from low resolution of images due to its weak capacity to identify local texture. In recent years, researchers have achieved good results in small object detection in optical remote sensing images through the improved Faster R-CNN [13]. A strategy for combining the Faster-RCNN model with two different convolutional neural networks (VGG-16 and ResNet-50) [14], which has good robustness in the specified vehicle datasets, it also shows that integrating the Faster-RCNN model with VGG-16 is better than ResNet-50. According to the characteristics of convolutional neural networks, some scholars proposed a new solution that modified the structure of Faster R-CNN, the network can integrate low-level and high-level features [15] for small object Detection. Liang Zhenwen *et al.* also proposed to use the deep feature pyramid networks [16] to solve the detection problems of small objects and achieved good performance.

By analyzing the structure of Faster R-CNN, this paper proposes an improved method based on Faster R-CNN for small object detection. In the positioning stage, we propose an improved loss function for bounding box regression, and use bilinear interpolation to improve the regions of interest (RoI) pooling operation to solve the problem of positioning deviation, in the recognition stage, we use the algorithm with multi-scale convolution feature fusion based on VGG-16 to make the feature map contains more information, and use the improved non-maximum suppression algorithm (NMS) to avoid missing overlapping objects. This algorithm has achieved good performance on the small traffic signs, and its performance is greatly improved compared to Faster R-CNN.

II. POSITIONING STAGE

A. IMPROVED INTERSECTION OVER UNION

Intersection over Union (IoU) is an important indicator in the system of object detection. In the regression task, the most direct indicator for judging the distance between the predicted bounding box and Ground Truth box is IoU, the formula for IoU is as follows:

$$IoU = \frac{S_{\text{DetectionResult}} \cap S_{\text{GroundTruth}}}{S_{\text{DetectionResult}} \cup S_{\text{GroundTruth}}} \tag{1}$$

Object detection relies on regression of bounding boxes to achieve accurate positioning. However, the IoU-based L_1 norm or L_2 norm loss function used in the process of regression is not good.

In order to solve the above problem, we have improved the IoU, denoted as IIoU (Improved IoU):

$$\text{IIoU} = \text{IoU} - \frac{C - (A \cup B)}{C} \tag{2}$$

Designated area S, where the minimum area C (C⊆S) containing A, B. The range of IoU is [0, 1], and the range of IIoU is [−1, 1]. The maximum value is 1 when the two regions coincide, and the minimum value is −1 when there is no intersection between the two regions and infinity, so IIoU is a good distance metric, and because IIoU introduces the minimum area C containing A and B, it not only pays attention to overlapping areas, but also other non-overlapping areas, even if A and B do not coincide, they can still be optimized.

B. LOSS FUNCTION

The definition and calculation of IIoU is simple. The calculation of coincident area is the same as IoU. When calculating the minimum closure area C, only the maximum and minimum coordinate values of the two bounding boxes are needed. The rectangle enclosed by these two coordinate values is C.

The coordinates of the top left and bottom right corner are used to represent each bounding box. The predicted bounding box is recorded as: $B = (x_1, y_1, x_2, y_2)$ and the bounding box of Ground Truth is recorded as: $B^* = (x_1^*, y_1^*, x_2^*, y_2^*)$.

Calculate the areas S^* and S of B^* and B:

$$\begin{aligned} S^* &= \left|\left(x_2^* - x_1^*\right) * \left(y_2^* - y_1^*\right)\right| \\ S &= \left|(x_2 - x_1) * (y_2 - y_1)\right| \end{aligned} \tag{3}$$

Calculate the overlap area S^I of B^* and B:

$$S^I = \begin{cases} \left|\left(x_2^I - x_1^I\right) * \left(y_2^I - y_1^I\right)\right| & x_2^I > x_1^I, \;\; y_2^I > y_1^I \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

x_1^I, x_2^I, y_1^I, and y_2^I are defined as:

$$\begin{aligned} x_1^I &= max(x_1, x_1^*), \quad x_2^I = min(x_2, x_2^*) \\ y_1^I &= max(y_1, y_1^*), \quad y_2^I = min(y_2, y_2^*) \end{aligned} \tag{5}$$

Find the smallest rectangle $C = \left(x_1^C, y_1^C, x_2^C, y_2^C\right)$ contains B^* and B:

$$\begin{aligned} x_1^C &= min(x_1, x_1^*), \quad x_2^C = max(x_2, x_2^*) \\ y_1^C &= min(y_1, y_1^*), \quad y_2^C = max(y_2, y_2^*) \end{aligned} \tag{6}$$

Calculate the area S^C of C:

$$S^C = \left|\left(x_2^C - x_1^C\right) * \left(y_2^C - y_1^C\right)\right| \tag{7}$$

Get IoU:

$$IoU = \frac{s^I}{S + S^* - S^I} \tag{8}$$

Calculate IIoU with the equation (2):

$$IIoU = IoU - \frac{S^C - \left(S + S^* - S^I\right)}{S^C} \tag{9}$$

The regression loss function of the bounding box:

$$L_{IIoU} = 1 - IIoU \tag{10}$$

where L_{IIoU} is non-negative and $L_{IIoU} \subseteq [0, 2]$. It is obviously that based on the L_n norm as a loss function, its local optimum is not necessarily the optimal value of IoU. Moreover, compared with IoU, the L_n norm is sensitive to the scale of object. In the process of optimization, although the central distance between the predicted bounding box and ground truth box is same, the overlap of the two boxes is different because the dimensions of the predicted bounding boxes are different. But IoU is a concept of ratio and therefore not sensitive to scale. It shows there is still a gap between optimizing the L_n norm-based loss function and the real situation of IoU. In this case, IIOU clearly shows the overlap between the predicted bounding box and ground truth box in the range of values, and also pay attention to overlapping and non-overlapping regions at the same time, the loss function L_{IIoU} based on it can also better optimize the position of the regression box and improve the problem of inconsistent optimization of the loss function and the actual situation of IIOU, and the disadvantage of directly using IoU as a loss function is also avoided.

TABLE 1. **Comparison of the detection performance of L_{IIoU} and smooth L_1 loss function.**

	AP	AP75
L_1	0.361	0.346
L_{IIoU}	0.389	0.395

Table 1. shows that the use of L_{IIoU} as the loss function for the bounding box regression has a certain improvement over the effect of detection by using the Smooth L_1 loss function. Average precision (AP) is the area that enclosed by the precision-recall (PR) curve of precision and recall. The 75 of AP75 indicates that the IoU value is greater than 0.75, and Fig.1 shows that the Faster R-CNN used L_{IIoU} has a lower missed detection rate and better effect than the Faster R-CNN which used Smooth L_1 Loss function.

C. POSITIONING DEVIATION

The region proposals are extracted by the RPN are sent to the subsequent fully connected network for further classification and fine-tuning of the bounding box. However, due to the limitation of the fully connected layer, the output region proposals of the RPN are different in size, so it is necessary to introduce an regions of interest (RoI) pooling layer. Since the scales of extracted region proposal are all based on the original image, they need to be mapped to the feature map, and then the feature map corresponding to each region proposal is divided into $k \times k$ bins, and the maximum pooling is performed for each bin, no matter how large the input is, the output through the RoI pooling layer is always $k \times k$.

However, in the process of mapping region proposals from the original image to the feature map, the coordinates of the mapped region proposals on the feature map are generally decimal, the rounding operation is shown in Fig.2. In addition, in the process of dividing the region proposals after the rounding operation into $k \times k$ bins, the rounding operation is also performed for each bin, and the rounding operation is shown in Fig.3. they cause the deviation of feature image to be mapped to the original image to be larger, so the positioning of the bounding box loses accuracy.

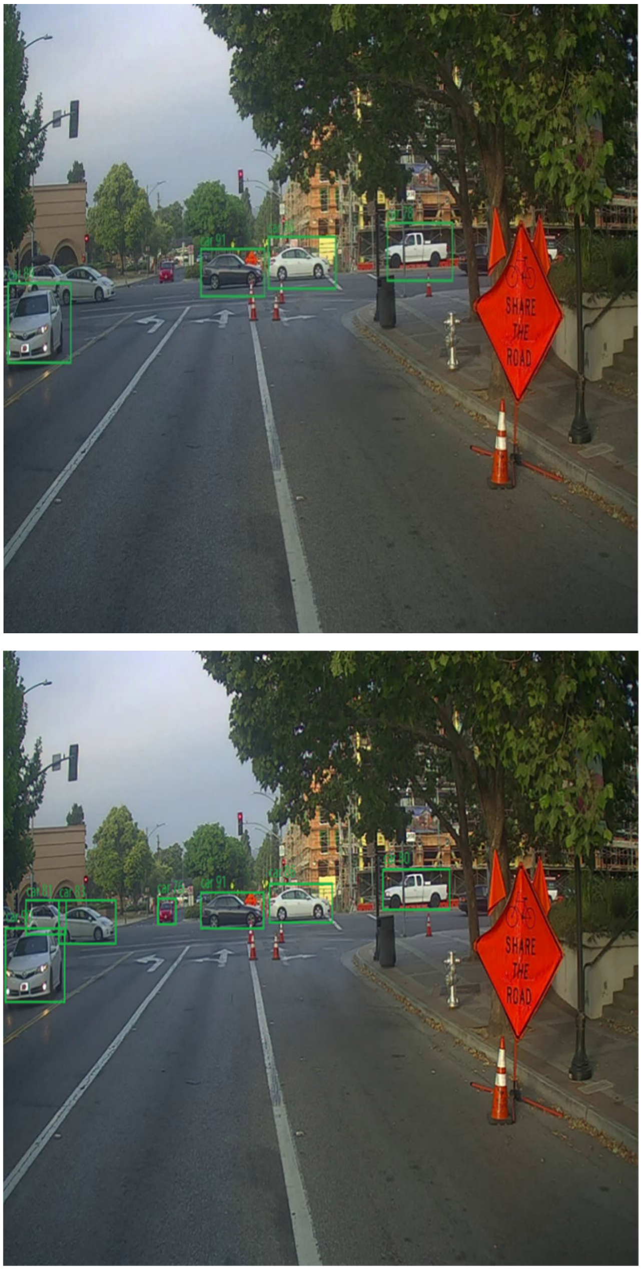


FIGURE 1. **Comparison of the performance detected by the two loss functions, the first picture used smooth L_1 loss function and the second picture used L_{IIoU}.**

D. BILINEAR INTERPOLATION

Bilinear interpolation is a good way to solve this problem, Fig.4 shows that Bilinear interpolation essentially involves linear interpolation in two directions. The values of the points Q_{11}, Q_{12}, Q_{21} and Q_{22} are known, and now we want to know the value at the point P.

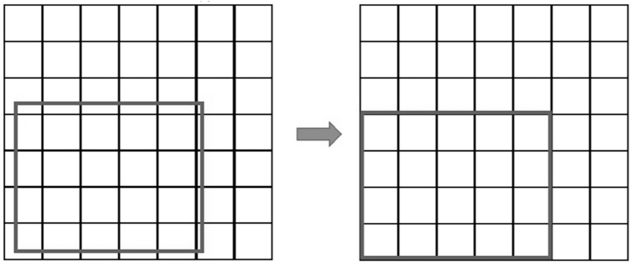


FIGURE 2. The mapped region proposal is rounded on the feature map.

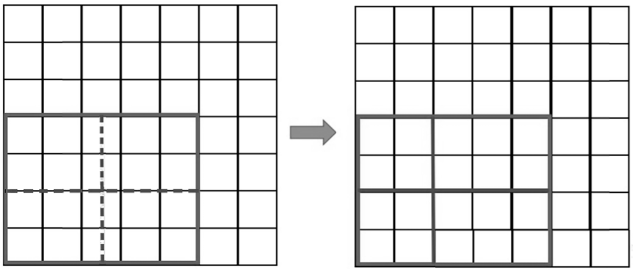


FIGURE 3. Rounding operation for each bin.

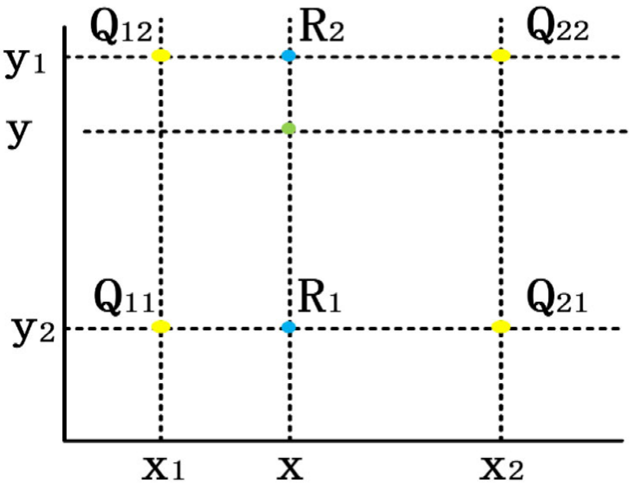


FIGURE 4. Bilinear interpolation schematic.

In the x direction, linear interpolation of Q_{11} and Q_{21} is worth the value of $R_1(x, y_1)$. Similarly, linear interpolation of Q_{12} and Q_{22} is performed to obtain the value of $R_2(x, y_2)$:

$$f(R_1) \approx \frac{x_2 - x}{x_2 - x_1} f(Q_{11}) + \frac{x - x_1}{x_2 - x_1} f(Q_{21}) \tag{11}$$

$$f(R_2) \approx \frac{x_2 - x}{x_2 - x_1} f(Q_{12}) + \frac{x - x_1}{x_2 - x_1} f(Q_{22}) \tag{12}$$

where linearly interpolating with R_1 and R_2 in the y direction to obtain:

$$f(P) \approx \frac{y_2 - y}{y_2 - y_1} f(R_1) + \frac{y - y_1}{y_2 - y_1} f(R_2) \tag{13}$$

Therefore, in order to solve the positioning deviation, the rounding operation of the RoI pooling layer is cancelled, and the decimal value is retained, the pixel value of the place where the coordinates are decimal that are obtained by the bilinear interpolation method described above, the operation is shown in Fig.5.

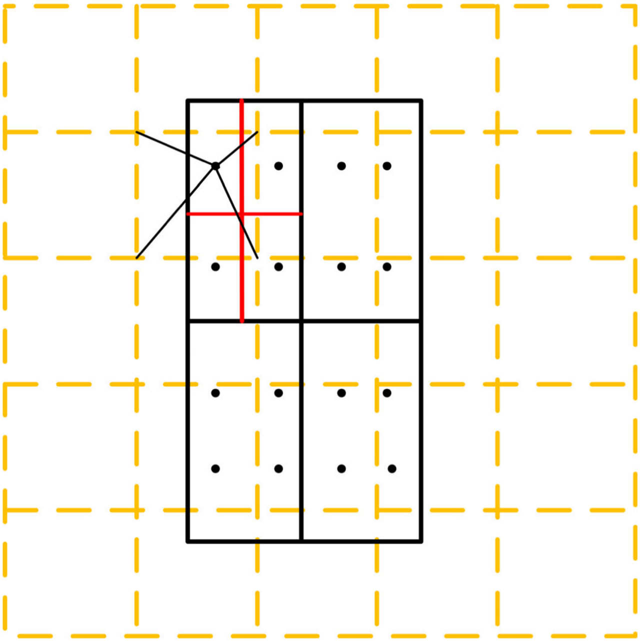


FIGURE 5. Bilinear interpolation for RoI pooling layer.

The dotted line in Fig.5 indicates the feature image, the intersection point of dotted line is the pixel point, and the black solid line indicates the region proposal. The region proposal is divided into 2 × 2 bins, the sampling point of each bin is set to 4, and the bin is evenly divided into 4 small areas, as shown by the red line, the center point of each small area is the sampling point, but the coordinates of the sampling point are usually decimal, so it is needed Bilinear interpolation of the pixel values of the sampling points, as shown by the four arrows, after obtaining the pixel values of the point, and then perform the maximum pooling operation on the four sampling points of each bin.

After the RoI pooling operation is improved, the backpropagation formula is adjusted at the same time. The object positioning performance of the two methods is shown in Fig.6. It can be observed that the normal pooling operation bounding box has a significant offset from the real object, and the rounding operation causes the predicted bounding box of region proposal to not match the Ground Truth, resulting in a traffic sign was missed, but with the improved RoI pooling operation, all three small traffic signs are correctly located.

III. RECOGNITION STAGE

A. CONVOLUTION FEATURE FUSION

Single-layer convolutional feature maps often lack some information of image. By the convolution and pooling operations of convolutional neural networks, the deeper the network, the smaller the feature maps are extracted, making it difficult to fully express the features of small objects.

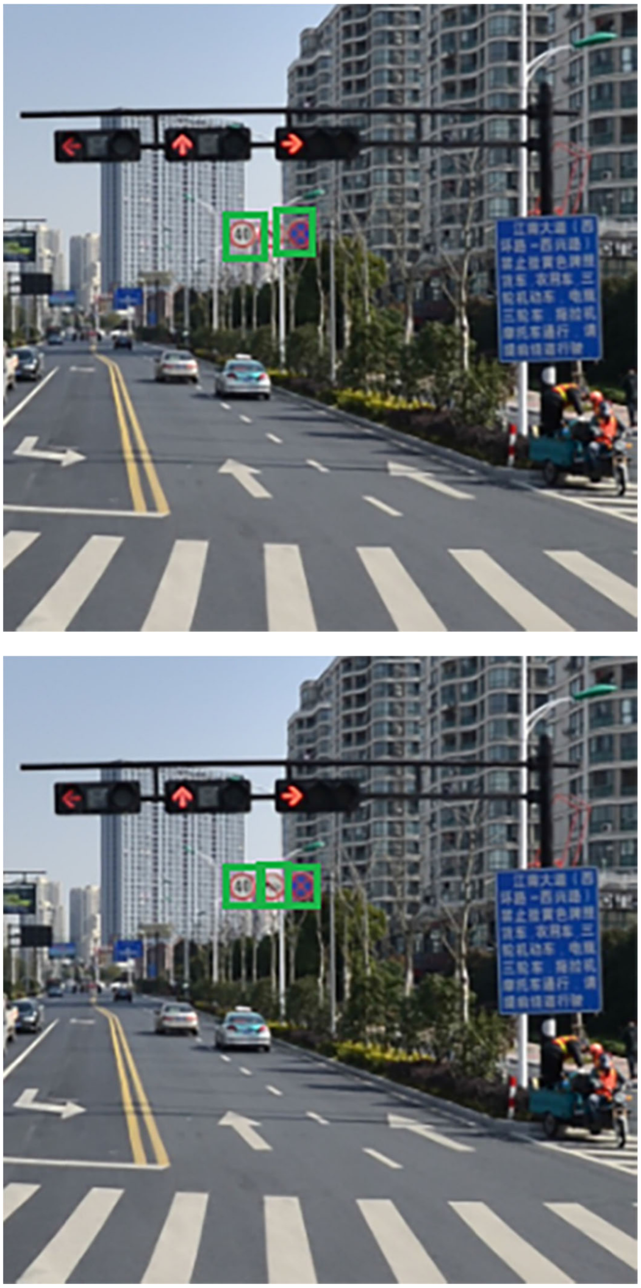


FIGURE 6. Comparison of two pooling methods, the first picture used normal RoI pooling operation and the second used improved RoI pooling operation.

The structure of VGG-16 network is shown in Fig.7. Faster R-CNN only uses the output of Conv5_3 layer as the feature map to the subsequent network. Therefore, we combine the features extracted by Conv3_3, Conv4_3 and Conv5_3 convolutional layers according to the addition of elements.

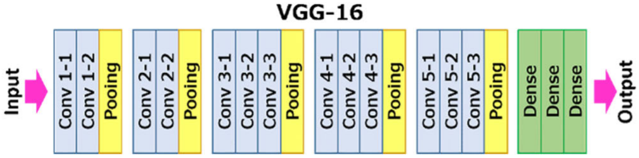


FIGURE 7. The structure of VGG-16 network.

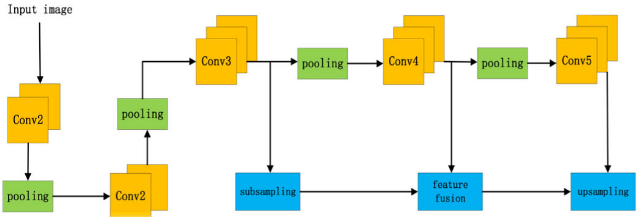


FIGURE 8. The extraction process of feature based on convolution feature fusion.

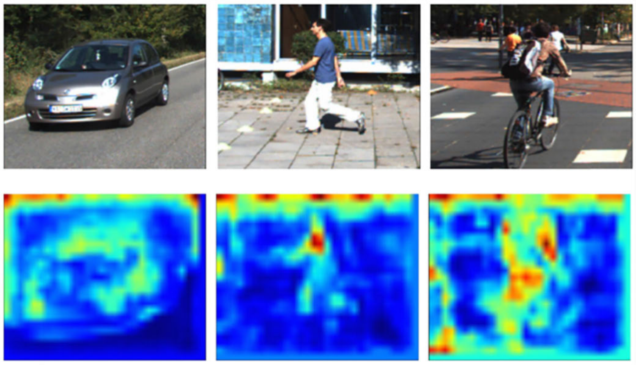


FIGURE 9. The performance of convolution feature fusion.

The framework of feature extraction network based on multi-scale convolution feature fusion is shown in Fig.8. The size of the feature maps is generated by each convolution layer are different, keep the size of feature map of Conv4_3 unchanged, and change the size of feature map of Conv3_3 and Conv5_3 to the size of the Conv4_3 feature map. The maximum pooling by subsampling is adopted for the feature map of Conv3_3, and the upsampling is used to improve the resolution of Conv5_3 feature map to make them consistent with the Conv4_3 feature map.

Use of bilinear interpolation does not require training features to improve resolution. Finally, the merged feature map can be obtained by adding the feature map included by subsampling the output of Conv3_3, upsampling by the output of Conv5_3, and the feature map of Conv4_3. Before the fusion of the three-layer convolutional feature maps, we first use local response normalization (LRN) to process each feature map so that the activated values of the feature map are the same.

The performance of merged feature map is shown in Fig.9, the object contour is faintly visible, and the detailed information are rich and contain abstract semantic information.

B. IMPROVED NMS

The object detection algorithm generates a large number of region proposals, and each region proposal has a corresponding score, and adjacent region proposals have relevant scores, which may cause false detection results and may result in some overlapping objects are missed. To solve this problem, non-maximum suppression algorithm (NMS) is proposed.

Non-maximum suppression algorithm sets an IoU threshold for a specific category object, the bounding box M have the largest score and it is selected from the generated series of bounding boxes B, removed from B and placed in the final detection result R, at the same time, the bounding boxes with the IoU of M greater than the threshold are removed from B. The non-maximum suppression algorithm repeats the above process until B is empty, and finally outputs the set D.

The non-maximum suppression algorithm is defined as:

$$s_i = \begin{cases} s_i & IoU(M, B_i) < p \\ 0 & IoU(M, B_i) \geq p \end{cases} \tag{14}$$

where p is the threshold of IoU. It can be seen from the above equation, the non-maximum suppression algorithm directly changes the bounding box score of the adjacent category to 0, causing some overlapping objects to be missed.

Then we used the Soft-NMS algorithm which improved the NMS algorithm and rescored the bounding box. If a bounding box overlaps with M most, the bounding box will get a low score. If the degree of overlap is low, the score is unchanged. The selected Soft-NMS mathematical is defined as:

$$s_i = \begin{cases} s_i & IoU\ (M, B_i) < p \\ s_i \times (1 - IoU\ (M, B_i)) & IoU\ (M, B_i) \geq p \end{cases} \tag{15}$$

where p is the threshold of IoU, and then the lower scores of boxes are removed. At the same time, as the bounding box with the highest score is M, if there are bounding boxes with high scores or the IoU with M are greater than 0.9, make them recombined. The positions of the recombined bounding boxes are weighted and averaged by the corresponding score weights to the original coordinates of bounding box, and the combined scores of bounding boxes are set to the average value.

IV. THE ALGORITHM AND ITS RESULTS

A. ALGORITHM OF THIS PAPER

In this paper, we adopted in two-stage detection algorithm, the process of positioning + recognition. The small object detection algorithm based on deep learning is designed. Fig.10 shows the overall framework of algorithm of this paper.

- Using the VGG-16 network to extract features, the feature maps with high-level semantic information are extracted by Conv5_3 layer are upsampled, and the feature maps with more detailed information are extracted by Conv3_3 layer are subsampled, merge with the features extracted by the Conv4_3 layer as input to the subsequent models. This part is shown in Box I.
- In order to detect small objects, redesign the anchor size of the RPN network, the feature map is merged into the RPN network to generate region proposals, and the softmax function is used as a loss function to determine whether the region proposal contains the object. The L_{IIoU} is used as the loss function for bounding box regression. This loss function is applied to the portion by the red dashed boxes II and V.
- Through classification and regression of the improved RPN network, a series of region proposals that may contain objects are output, and use the Soft-NMS to eliminate some overlapping region proposals in the Proposal layer. The algorithm is applied to the red dotted line in Box III.
- Make the region proposal generated by the Proposal layer and the previously merged feature map into the improved RoI pooling layer, and use the bilinear interpolation method to perform the pooling operation to obtain the fixed region proposals. This part of the algorithm is shown in the red dashed box IV. Finally, through the full connection layer into the subsequent bounding box regression, the border trimming and specific category classification by softmax are performed.

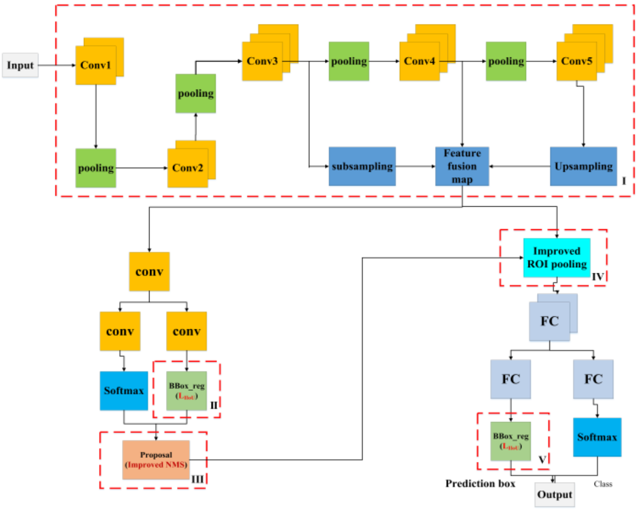


FIGURE 10. The overall algorithm framework of this paper.

B. DATASET

In order to test the accuracy of the algorithm that proposed in this paper on small objects, we select the traffic sign dataset TT100K (Tsinghua-Tencent 100K) jointly issued by Tsinghua University and Tencent. The dataset is characterized by the object of interest (traffic sign) whose resolution is very low, mostly between 12 and 60 pixels. Since the number of samples in each category of the data set are not balanced, IoU>0.67 is regarded as a positive sample in the training phase, and the region proposal with IoU<0.3 is used as a negative sample. The initial learning rate α during training is set to 0.001, and the momentum-based small batch gradient descent momentum weight ε is set to 0.9. To prevent overfitting, the dropout method is adopted, and the probability of random culling is set to 0.5.

TABLE 2. Classification result matrix.

The true situation	Forecast result	
	Positive	Negative
Positive	TP	FN
Negative	FP	TN

TABLE 3. Comparison with ours and other three algorithms based on the traffic signs whose resolution in range of (0,32).

	Faster R-CNN	Zhu *et al.*	Yan *et al.*	Ours
Recall	46%	87%	89%	90%
Accuracy	74%	82%	84%	87%

C. EVALUATION INDICATORS

Combine the categories and real conditions detected by the model into true positive (TP), false positive (FP), false negative (FN) and true negative (TN) as shown in Table 2:

Recall is defined as:

$$Recall = \frac{TP}{TP + FN} \tag{16}$$

Accuracy is defined as:

$$Accuracy = \frac{TP + TN}{TP + FN + TN + FP} \tag{17}$$

The accuracy rate indicates the proportion of the correctly detected samples to the total number of samples detected. The recall rate indicates the proportion of the correctly detected samples in all samples that should be detected.

D. ANALYSIS

The algorithm is trained and tested on the TT100K dataset and compared with algorithms such as Faster R-CNN, Zhu *et al.* [17] and Li *et al.* [18]. The results are shown in Table 3.

Table 3. shows that the recall rate and Accuracy rate of the algorithm whose resolution is in range of (0,32] is better than the other three algorithms, indicating that the algorithm has better detection effect on small objects.

E. EFFECT PICTURE

The overall effect diagram is detected by the algorithm in this paper is shown in Fig.11. The partial effect diagram is provided for easy observation, it can be seen that the positioning and recognition of small targets are better and the detection accuracy is higher.

Fig.12 shows the effect diagram of Zhu *et al*, Yan *et al* and the algorithm of this paper. Figs (a) are the effect diagrams of Zhu *et al.* and ours. Figs (b) are the effect diagrams of Yan *et al.* and ours. Compared with the algorithm which are proposed in this paper, the missed detection rate is significantly reduced.

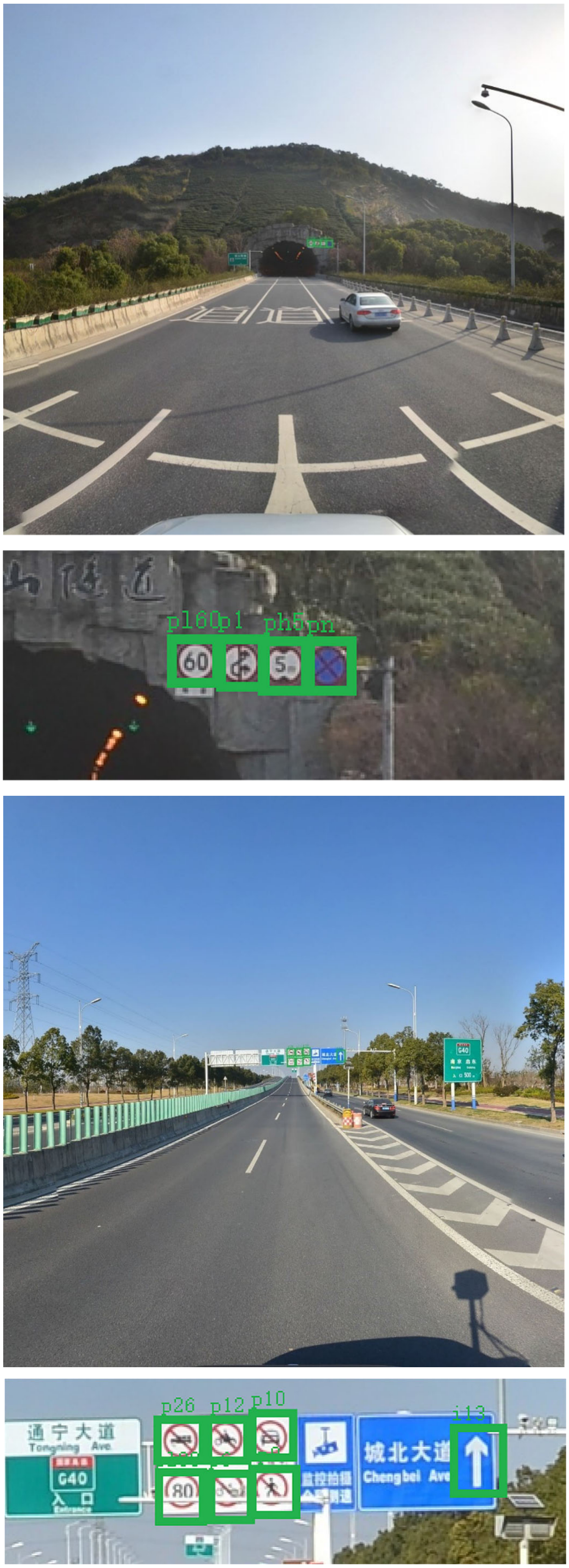
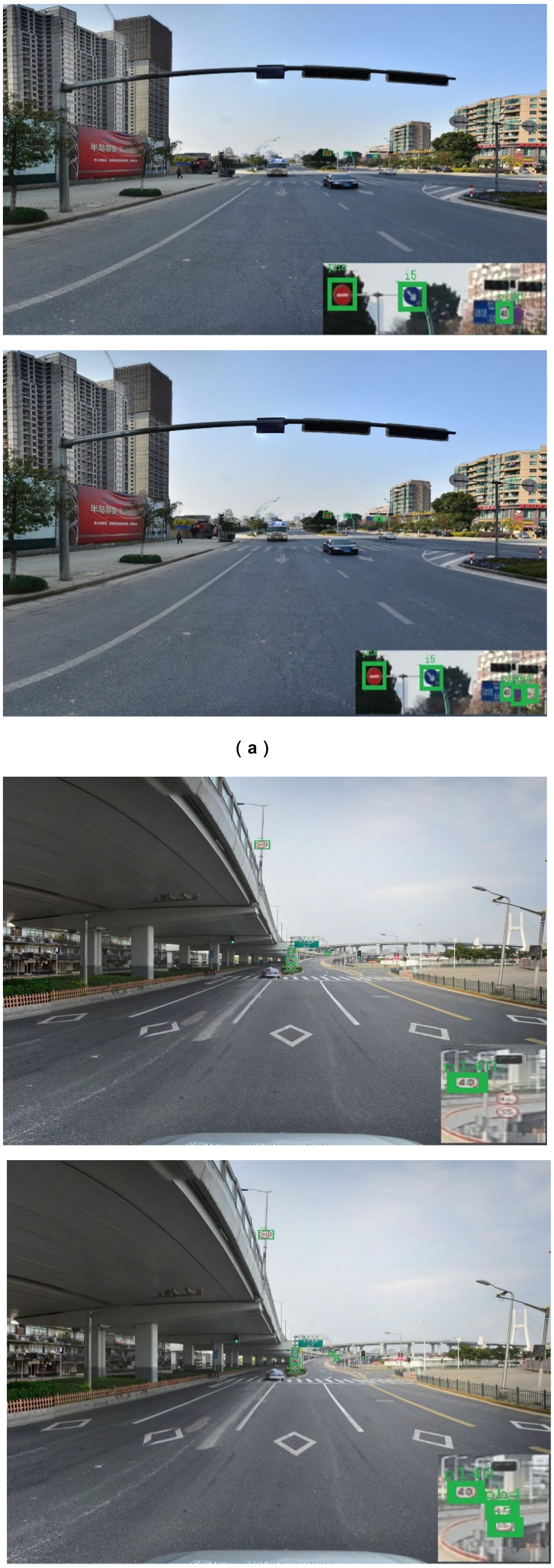


FIGURE 11. Overall and partial effect pictures.



(a)

(b)

FIGURE 12. Comparison of our algorithm with Zhu *et al.* and Yan *et al.* Figs (a) are the effect diagrams of Zhu *et al.* and ours. Figs (b) are the effect diagrams of Yan *et al.* and ours.

V. CONCLUSION

Based on the Faster R-CNN, this paper proposes the improved L_{IIOU} loss function in the positioning, which solves the problem that the L_1 norm and L_2 norm loss function cannot accurately reflect the overlap between the prediction region proposal and the ground truth, and the bilinear interpolation is used in the RoI pooling operation to solve the positioning deviation caused by traditional method. The use of convolutional feature fusion and soft-NMS in recognition also greatly improve the accuracy of target recognition. The performance of the algorithm on the specified data set whose resolution is in range of (0,32] is also better than that of Faster R-CNN, Zhu *et al.* [17]. and Li *et al* [18]. The good performance of the proposed algorithm on small traffic signs has a high reference value in the field of intelligent driving, and it has broad application prospects in civil or military applications.

REFERENCES

[1] T. Bretschneider and K. Odej, "Content-based image retrieval," in *Encyclopedia of Data Ware Housing Mining*. Hershey, PA, USA: Idea Group Publishing. 2015, pp. 212–216.

[2] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 25, 2012, pp. 1097–1105.

[3] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3431–3440.

[4] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, May 2015.

[5] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards realtime object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.

[6] D. Hossain, G. Capi, and M. Jindai, "Object recognition and robot grasping: A deep learning based approach," in *Proc. 34th Annu. Conf. Robot. Soc. Jpn. (RSJ)*, Yamagata, Japan, Sep. 2016, pp. 1–5.

[7] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, "CNN features off-the-shelf: An astounding baseline for recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR) Workshops*, Jun. 2014, pp. 806–813.

[8] A. B. Yandex and V. Lempitsky, "Aggregating local deep features for image retrieval," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 1269–1277.

[9] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

[10] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Region-based convolutional networks for accurate object detection and segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 1, pp. 142–158, Jan. 2015.

[11] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 9, pp. 1904–1916, Sep. 2015.

[12] R. Girshick, "Fast R-CNN," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015, pp. 1440–1448.

[13] Y. Ren, C. Zhu, and S. Xiao, "Small object detection in optical remote sensing images via modified faster R-CNN," *Appl. Sci.-Basel.*, vol. 8, no. 5, p. 813, 2018.

[14] Z. Huang, M. Fu, K. Ni, H. Sun, and S. Sun, "Recognition of vehicle-logo based on faster-RCNN," in *Proc. ICSINC*, 2018, pp. 75–83.

[15] H. Jipeng, S. Yinghuan, and G. Yang, "Multi-scale faster-RCNN algorithm for small object detection," *Comput. Res. Develop.*, vol. 56, no 2, pp. 319–327, 2019.

[16] L. Zhenwen, J. Shao, D. Zhang, and L. Gao, "Small object detection using deep feature pyramid networks," in *Proc. 19th Pacific-Rim Conf. Multimedia (PCM)*, 2018, pp. 554–564.

[17] Z. Zhu, D. Liang, S. Zhang, X. Huang, B. Li, and S. Hu, "Traffic-Sign Detection and Classification in the Wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 2110–2118.

[18] J. Li, X. Liang, Y. Wei, T. Xu, J. Feng, and S. Yan, "Perceptual generative adversarial networks for small object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 1951–1959.

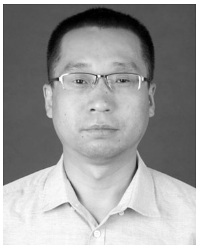
CHANGQING CAO received the Dr.Eng. degree. He completed all of his studies with Xidian University, where he was an Associate Professor, in 2011. His research interests include laser technology and its applications.

BO WANG received the bachelor's degree from Xi'an Technological University. He is currently pursuing the master's degree with Xidian University. His main research interests include deep learning and object detection.

WENRUI ZHANG received the master's degree from Xidian University, in 2010, where he is currently pursuing the Ph.D. degree. His major research interests include free space optical communications.

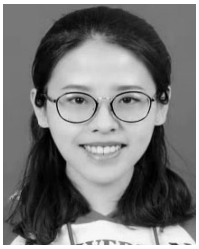
XIAODONG ZENG graduated from Xidian University, in 1996. He received the Dr.Eng. degree. He was with Xidian University, where he is currently a Professor. His research interests include optoelectronic technology and its applications.

XU YAN graduated from Xidian University, where he is currently pursuing the master's degree. His main research interests include photoelectric detection and image processing.



ZHEJUN FENG graduated from Xidian University, in 2008, where he is currently pursuing the Ph.D. degree. He is an Associate Professor. His research interests include photoelectric detection and signal processing.

YUTAO LIU received the master's degree from Yanshan University, in 2015, where he is currently pursuing the Ph.D. degree. His research interest includes heterodyne detection.



ZENGYAN WU received the bachelor's degree from the North University of China. She is currently pursuing the master's degree with Xidian University. Her main research interest includes coherent optical communications.