ISSN: 2688-8645

Invoice Detection and Classification based on Improved YOLOv5s

Weihua Niu[1, 2], Qiaoyue Liu[1]

[1]Department of Computer, North China Electric Power University, Baoding Hebei 071000, China

[2]Engineering Research Center of Intelligent Computing for Complex Energy Systems, Ministry of Education, Baoding Hebei 071000, China

Abstract

For the complex problems of invoice occlusion, invoice deformation, dark environment, excessive noise and so on in invoice detection, this paper proposes an improved YOLOv5s invoice detection and classification method. In order to improve the generalization ability of the model, the attention mechanism is introduced to improve the feature extraction ability of the network. By adding cavity convolution to the YOLOv5S backbone network and the neck network, and adding context transformation network to the backbone network, the robustness of the model is improved. For model output, flexible non-maximum suppression is used to replace non-maximum suppression to improve the detection effect. Comparative experiments show that the accuracy, recall and average accuracy of the proposed method are greatly improved.

Keywords

Invoices Classification; YOLOv5s; CotNet; HDC; Soft NMS.

1. Introduction

At present, popular object detection algorithms are often used to deal with computer vision tasks, which mainly include automatic driving, face recognition, video surveillance, defect detection and other fields. The traditional invoice detection and classification technology is optical character recognition (OCR) technology [1], but due to the image pixel or invoice itself and other problems, the detection results are not ideal. Deep learning network is widely used. The application of target detection algorithm in invoice detection and classification can greatly reduce the cost of manual invoice classification and improve the speed of invoice detection and classification. In the field of target detection, it can be divided into two categories: target detection algorithm based on single-stage regression and target detection algorithm based on two-stage candidate region. Object detection algorithms based on single-stage regression, such as YOLO[2], SSD[3], etc.; Two-stage target detection algorithms based on candidate regions, such as Fast-RCNN[4], Faster-RCNN[5], etc. Among them, the single-stage target detection algorithm based on regression can directly output the detection results, while the two-stage target detection algorithm will not directly output the detection results, but first generate the candidate region, and then output the detection results. Generally speaking, the two-stage algorithm has higher accuracy but longer running time, so it is not suitable for real-time detection tasks. The single - stage algorithm is faster, but the accuracy of the intersection - two - stage algorithm is lower.At present, there are still many difficulties in invoice classification. Due to uneven lighting conditions, excessive deformation of invoices, serious occlusion of invoices, damage of invoices, excessive noise, small data size and other factors, it is more difficult to detect and classify invoice targets. At the same time, the existing target detection algorithms are not suitable for development on local common hardware devices due to the large number of parameters, large memory consumption and long training time. YOLOv5s model has the advantages of fast speed, high accuracy and few model parameters, which has

ISSN: 2688-8645

good application value in practice and development application. However, the recognition effect of images taken in dark environment and fuzzy image needs to be improved. Based on the traditional YOLOv5s model, a feature extraction network model with more detailed classification is designed in this paper.The main work of this paper is as follows: (1) Lacking the standard invoice data set, 8 classified invoice data sets were built by ourselves, including 2235 data sets for training and 400 data sets for testing. (2) The YOLOv5s model is improved, and the proposed method is verified by comparative experiments. (3) The cloud server is used to train the network, and the network model and data set are deployed on the cloud server, so that it can modify parameters locally and control operation.

2. Overview of Invoice Classification

In this task, invoices are mainly divided into 'invoice', 'invoice_taxi', 'invoice_train', 'invoice_highway', 'attachment_delivery', 'attachment', 'attachment_list' and 'attachment_receipt' have eight categories. There are many low-quality invoice images in actual computer vision tasks. YOLOv5s does not have a good learning degree of low-quality images, so it is necessary to modify the model to make it suitable for this task.

3. YOLOv5s Principle and Improvement

3.1. Principle of YOLOv5s Algorithm

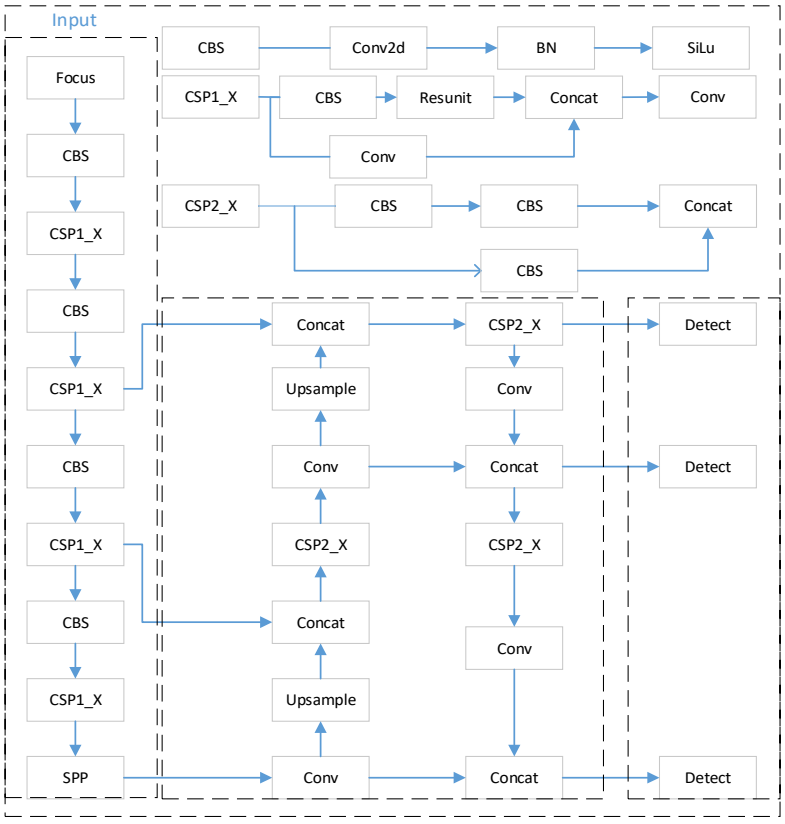

Figure 1. YOLOv5s structure

YOLOv5s is an improvement based on YOLOv4[6]. YOLOv5 series algorithms mainly include YOLOv5l, YOLOv5x, YOLOv5s and YOLOv5m. This paper adopts YOLOv5s as the basic model of the experiment. The YOLOv5s model is usually composed of Backbone network, Neck network and Head network. In addition, data enhancement (Mosaic) module is added. The backbone network mainly extracts the features of the invoice image. Combined with FPN [7], the neck

ISSN: 2688-8645

network mainly fuses the features extracted from invoices, which are mainly from the backbone network. The header network can screen the detected boxes through non-maximum suppression operation [8]. Data enhancement can enrich data and improve the learning degree of the model. YOLOv5s adopts Mosaic data enhancement to enrich training data set, randomly arranging, stitching and cropping four images into a new image. The backbone network mainly includes the Focus layer, CSPNet network, and Leaky ReLU activation function. The Focus layer replaces the three convolution layers of the original network and performs slicing and down sampling operations on the input image. The Focus layer can save the number of parameters and the running time of the model.

CSPNet network extracted the feature information in the image, in which Resunit residual block adopted Resnet50[9] as the network model. Leaky ReLU activation function updates parameters in the network and improves the expression capability of the network. The neck network adopts image pyramid structure (FPN) and proposes a multi-scale algorithm to accurately locate the target by absorbing the underlying feature information. The header network can reflect the coincidence degree between the predicted box and the real box [10]. The structure flow chart of YOLOv5s is shown in Figure 1, where CBL module is a combination module of Conv, normalized BN and Leaky ReLU activation function.

3.2. YOLOv5s Improvement

3.2.1. Soft NMS Soft Non-maximum Suppression

In order to ensure the recall rate, the output of the network is usually more than one candidate box. If these candidate boxes are not restricted, the same target object may have multiple detection boxes. In order to optimize the prediction result, the NMS non-maximum suppression method is used to suppress the borders with high scores and high overlap. However, the non-maximum suppression uses confidence score to forcibly suppress other candidate boxes, which has some defects. When the objects are relatively dense, a variety of candidate frames will be generated, and forcing the suppression of non-maximum candidate frames will reduce the recall rate of the model. When the number of detected prediction frames is large, the time of model detection is longer. It is difficult to determine the threshold of the non-maximum suppression method of the NMS. A high threshold may cause detection errors, and a low threshold may reduce the recall rate of the model, resulting in poor detection results. In this paper, based on the NMS non-maximum suppression, Soft NMS suppression score is used to reduce the score of the candidate box instead of making its confidence level to 0. The advantage of suppression score lies in that there is no need to retrain the model and only need to add a suppression score mechanism to the object detection algorithm without increasing the calculation amount. Soft NMS non-maximum suppression algorithm is more general. The NMS calculation formula is Formula (1), and the Soft NMS calculation formula is Formula (2).

$$s_i = \begin{cases} s_i, iou(M, b_i) > N_t \\ 0, iou(M, b_i) \geq N_t \end{cases} \tag{1}$$

$$s_i = s_i e^{-\frac{iou(M,b_i)^2}{\sigma}}, \forall b_i \notin D \tag{2}$$

3.2.2. Context Transform Network (CoTNet)

Based on YOLOv5s add context network (Contextual Transformer Networks) [11], the existing network for Transformer style rich context between adjacent secret key tend to be ignored. Context converter (CoT) is added on the basis of YOLOv5s network, so that dynamic attention learning is guided by rich context information. CoT blocks can easily replace any 3×3 convolution in the ResNet50 system to produce a context transform network (CoTNet), as shown in Table 1. Especially for low illumination image, the image information of object is usually covered by dark area or a merger with a dark background, which can only get the

ISSN: 2688-8645

invoice image outline, cannot capture the details outline, invoice for detection and classification of test results is challenging [12]. Therefore, a bottom-up context transformation network is introduced in the backbone network for feature compensation in the process of convolution from lower level to higher level. The input query is combined with the encoded key, and the combined result is input into two 1×1 convolution to get the attention matrix. Dynamic context representation is to multiply the obtained attention matrix with the input value, and finally output the static and dynamic context representation into the next layer of the network.

Table 1. CoTNet

Layer_name	Output_size	Resnet50		**Resnet50+ Cotnet**	
Conv1	112×112	7×7,64,stride 2		7×7,64,stride 2	
Conv2	56×56	3×3 max pool, stride 2		3×3 max pool, stride 2	
Conv3	28×28	1×1,64 3×3,64 1×1,256	×3	1×1,64 ***CoT***,64 1×1,256	×3
Conv4	14×14	1×1,128 3×3,128 1×1,512	×4	1×1,128 ***CoT***,128 1×1,512	×4
Conv5	7×7	1×1,256 3×3,256 1×1,1024	×6	1×1,256 ***CoT***,256 1×1,1024	×6
Conv6	112×112	1×1,512 3×3,512 1×1,2048	×3	1×1,512 ***CoT***,512 1×1,2048	×3

3.2.3. Improved Boundary Frame Loss Function

The more accurate the model is for target positioning, the smaller the value of the loss function. The loss function used in the model includes three aspects, namely Lobject confidence loss, Lbox loss and Lclass loss calculation function [13]. The loss function of YOLO series models is shown in Equation (3). Among them, the weights are a1=0.7, a2=0.05 and a3=0.3. An impact factor Lasp is added to the improved loss function on the basis of the original loss function. The impact factor Lasp takes into account the loss of intersection ratio between the prediction box and the real box, the distance between the prediction box center and the real box center and the loss of transverse and longitudinal ratio. The mathematical expression of the new boundary frame loss function is shown in Equation (4). The variables in Formula (4) are shown in equations (5) - (6):

$$L=a_1 \times L_{obiect}+a_2 \times L_{box}+a_3 \times L_{class} \tag{3}$$

$$L_{box}=L_{IoU}+L_{dis}+L_{asp} \tag{4}$$

$$L_{Iou}=1-\left|\frac{B \cap B_t}{B \cup B_t}\right| \tag{5}$$

$$L_{dis}=\frac{d^2(b,\ b_t)}{c^2} \tag{6}$$

Where, B is the prediction box, Bt is the real box, d is the distance between the prediction box and the center point of the real box, b is the center point of the prediction box, bt is the center point of the real box, c is the diagonal distance of the outer boundary rectangular box, h and w are the height and width of the prediction box, ht and wt are the height and width of the real box, cw and ch are the width and height of the outer rectangular box. By replacing the boundary frame loss function in YOLOv5s with the new loss function, the model converges faster in training, which is helpful to improve the accuracy of the model in the training process.

ISSN: 2688-8645

3.2.4. Fusing the HDC Void Convolution Module

In this paper, a void convolution module [14] conforming to HDC design criteria is integrated. In order to reduce information loss, a void convolution method is introduced. The main idea is to add gaps in the convolution region to enlarge the receptive field. The convolution kernel size of the empty convolution and the ordinary convolution is actually the same, but there are some gaps between the sliding Windows of the empty convolution. These gaps are the expansion factors in the empty convolution, and the contents of these gaps are all expressed by 0. The empty convolution has one more hyperparameter than the standard convolution, and this hyperparameter is the expansion factor. Empty convolution is an ordinary convolution. Hollow convolution enlarges the receptive field while keeping the size of the original feature image unchanged, so as to better grasp the overall information of the image, and has a good effect on the recognition of blocked image and overlapping image. Generally, the expansion factor is assumed to be a, and the relationship between the original convolution kernel and the hollow convolution kernel is shown in formula (7). The size of the actual hollow convolution kernel is set to M, the size of the original convolution kernel is set to m, and the value of the expansion factor a should meet the design criteria of HDC. The formula for calculating the hollow convolution receptive field is shown in Formula (8), where rn is the size of the receptive field of this layer, the size of the convolution kernel of this layer is set as kn, and the step size of the ith layer is set as si, that is, the hollow convolution will increase the receptive field.

$$M = M + (m-1)(a-1) \tag{7}$$

$$r_n = r_{n-1} + (k_n - 1)\prod_{n=1}^{n-1} s_i \tag{8}$$

3.3. Improved YOLOv5s Network Structure

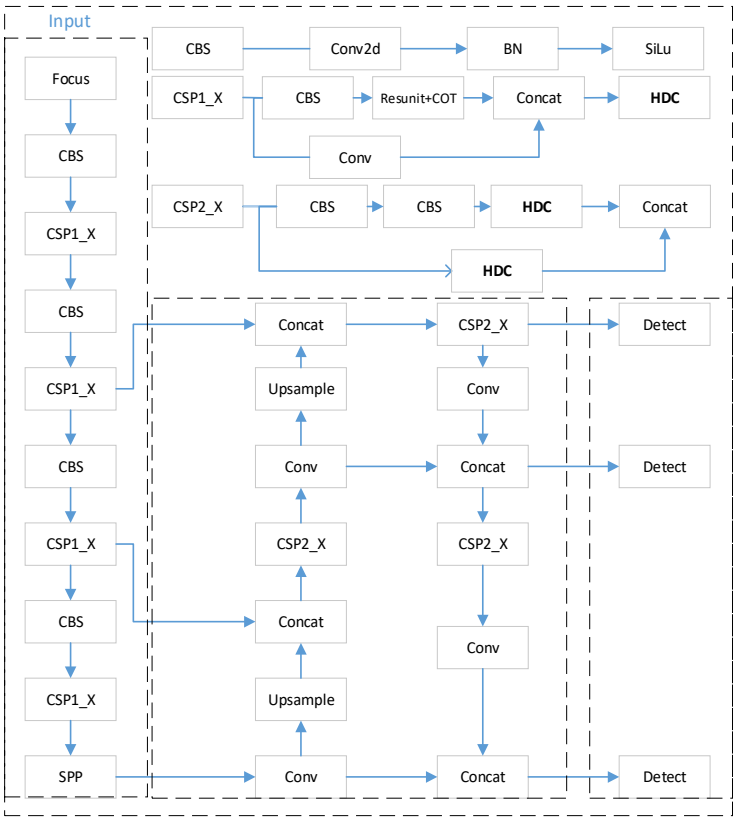

Figure 2. Improved YOLOv5S structure

ISSN: 2688-8645

In this paper, the YOLOv5s backbone network is improved by adding the following transform network, improving the boundary frame loss function, and replacing common convolution with void product. The feature extraction capability of the improved network is enhanced, and the detection effect of dark image and occlusion image is significantly improved. The improved YOLOv5s structure is shown in Figure 2.

4. DataSet

In this paper, self-built data set was used, labelImg annotation tool was used to annotate the data set, and the following categories of data set were classified: 'invoice', 'invoice-train', 'invoice-taxi', 'invoice-highway', 'attachment', 'attachment-list', 'Attachment-receipt', 'attachme' nt-delivery 'eight categories. There are 2235 data sets used for training, among which invoice accounts for about 34% of the total, "invoice-train" accounts for 13%, "invoice-taxi" accounts for 12.9% and "invoice-highway" accounts for 17%. 'attachment' accounts for about 33%, 'Attachment-list' for 10%, 'Attachment-receipt' for 9% and 'Attachment-delivery' for 4%. The detection process and the shooting process will have different effects of light, in the case of insufficient light, image quality will decline, color distortion, contrast decline, and then there are various defects in the invoice itself. Because current target detection tends to focus on images under normal lighting, while low-light images are often ignored and lack of data under dark conditions, in order to expand the data set shot under low light environment and improve the generalization ability of the model, low light images are artificially generated by Gaussian noise and gamma transform [15] to expand the data set and improve the generalization ability of the model. 400 data sets were used for testing. The size of the unified input image is 640×640, and the number of iterations is 600. Except for the model modification mentioned in this paper, other parts remain unchanged, and default parameters are used.

The data set is expanded by rotation, clipping, grayscale change, scaling, random occlusion, adding noise, dark processing (Gaussian noise and gamma transform) and other operations. Enrich the data number of low-quality images by means of data enhancement such as dark processing, so as to increase the proportion of low-quality images in the data set and improve the accuracy of low-quality images in the test.

5. Experimental Results and Analysis

5.1. Experimental Process and Experimental Environment

This experiment is carried out in cloud server based on cloud computing. The cloud server is a group of computer clusters with powerful hardware configuration, high fault tolerance and low failure rate, which can reduce the extra cost caused by local miscontact or power failure and save computing time. The experimental environment and parameter configuration are shown in Table 2.

Table 2. Experimental environment and parameters

parameter	configuration1	Training parameter name	Parameter value
CPU	6x Xeon E5-2680 v4	Initial learning rate	0.0100
GPU	NVIDIA RTX A2000	Learning rate decline parameter	0.0001
system environment	Ubuntu 18.04	Weight attenuation coefficient	0.0005
frame	Pytorch 1.8.1	momentum	0.9370
language	Python 3.7	batch size	16

ISSN: 2688-8645

5.2. Evaluation Index

In order to evaluate the performance of the model for invoice detection, some comparative indicators such as Precision rate P (Precision), mAP (mean Average Precision) of detection, R (Recall), Average precision AP (Average Precision) are used. Detection speed speed. The average precision mean (mAP) and the time taken can be used for preliminary evaluation of the model to verify the availability of the improved method [16].

Accuracy P represents the ratio of the predicted positive sample (TP) to the sum of the predicted positive sample (TP) and the predicted positive negative sample (FP). The higher P is, the lower the error detection rate is. The calculation formula of accuracy P is formula (9) .

$$P = \frac{TP}{TP + FP} \tag{9}$$

$$R = \frac{TP}{TP + FN} \tag{10}$$

Recall rate R is the ratio of positive samples (TP) predicted by the model to positive samples (TP) predicted by the model and positive samples (FN) predicted by the model to negative. The higher R is, the more samples are correctly detected and processed, and the lower the missed detection rate [17]. The calculation formula of recall rate R is Formula (10) .

5.3. Comparative Analysis of Models

YOLOv5s-Ⅰ is a network model with modified boundary frame loss function, YOLOv5s-Ⅱ is a network model with context transformation network added on the basis of YOLOv5s-Ⅰ, and YOLOv5s-Ⅲ is a network model with cavity convolution added on the basis of YOLOv5s-Ⅱ. To verify the effect difference between the improved YOLOv5s model in this paper and other commonly used model, verify the difference between the improved new method and YOLOv4 and Faster-RCNN. The experiment was verified on the same data set. In the task of invoice classification, the number of training sets is 2235, and the number of test sets is 400. The training iteration epoch is 600, the batch-size is 16, and the single machine and single card are used for training with one GPU. In this paper, YOLOv5s is adopted as the training and detection model.

As can be seen from Table 3, the accuracy of the improved YOLOv5s model has been greatly improved. Although the detection speed is slower to some extent, it still has great advantages compared with other popular network models.

This paper was compared with the traditional invoice classification method: Optical Character Recognition (OCR) technology [18] in the task of invoice classification. The principle of the experiment is to identify specific words through ocr, and then judge the category of invoice through these words. In this paper, the original training data set was identified by ocr. Due to the different accuracy of ocr technology for image text recognition, the data set in this experiment was classified into two categories: high quality data set and low-quality data set (fuzzy, uneven illumination, incomplete image).

Failure to recognize key information will result in classification errors. Traditional optical character recognition is not effective for image recognition in dark environment, which is easy to lose important information and lead to classification errors. ocr technology is not suitable for the identification and classification of low-quality invoices.

5.4. Analysis of Experimental Results

Through the improvement of the trunk network and the loss function, the longitudinal experiment is compared, and it can be seen that the feature extraction of the model shows a better effect. Experimental results show that the accuracy rate P, recall rate R, average detection accuracy mAP@0.5, mAP@0.5:0.95, and average detection speed of an image are 0.330s on

ISSN: 2688-8645

invoice data set after the improved method. Compared with the original method, the accuracy of the new improved method is improved by 5.6% in P, 6.6% in R, 3.1% in mAP@0.5, and 3% in mAP@0.5:0.95. The horizontal comparison experiment proves that deep learning is more effective than optical character recognition technology in invoice classification. Under the condition of similar detection speed, the improved model has better detection accuracy. The improved YOLOv5s works well in the invoice sorting task and can be put into use in the invoice sorting application. The experimental results are shown in Figure3. The Precision tends to decline when the number of iterations is about 400, indicating that the model is overfitting and the number of iterations should be reduced, and regularization processing should be adopted in the algorithm.

Table 3. Indicators show

Model	P /%	R /%	mAP@ 0.5 /%	mAP@ 0.5: 0.95 /%	speed /(s/sheet)
YOLOv5s	0.858	0.893	0.938	0.880	0.300
YOLOv5s- Ⅰ	0.866	0.899	0.935	0.887	0.323
YOLOv5s- Ⅱ	0.890	0.922	0.946	0.891	0.320
YOLOv5s-Ⅲ	0.914	0.959	0.969	0.910	0.330
YOLOv4	0.813	0.841	0.879	0.849	0.540
Faster-RCNN	0.742	0.767	0.815	0.772	0.830

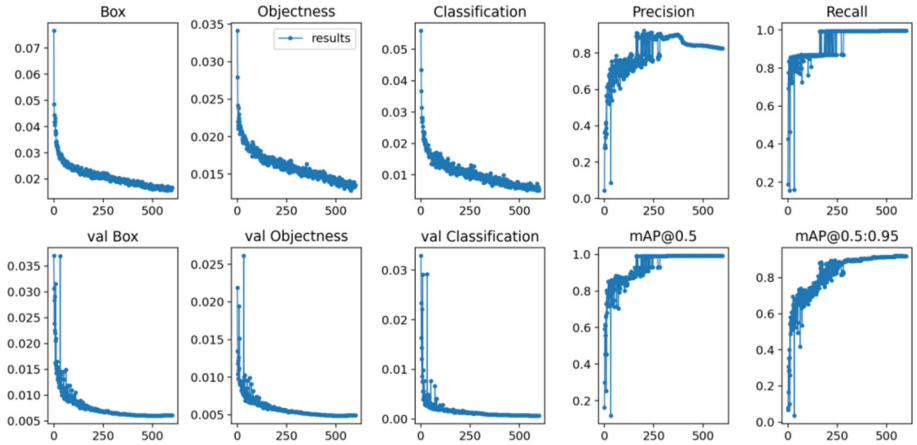

Figure 3. The experimental results

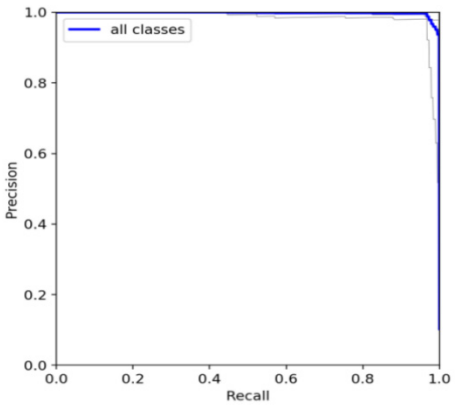

Figure 4. PR curve of the improved model

In Figure 4, Box represents the mean value of loss function; Objectness represents the mean value of detection target loss; Classification represents the mean value of classification loss;

ISSN: 2688-8645

Precision represents the accuracy rate; Recall represents the recall rate; mAP@0.5 represents the average AP value of all categories when IoU is 0.5. mAP@0.5:0.95 represents the average AP value for all categories of different IoU between 0.5 and 0.95. The smaller the first three values, the better. mAP@0.5 and mAP@0.5:0.95, the larger the value, the better.

It can be seen from Figure4 that the PR curve of the improved YOLOv5s model and the coordinate axis have a large besieged area and better performance, which is suitable for invoice classification.

6. Conclusion

Aiming at the problem of low accuracy of invoice classification, this paper firstly collects data. Considering the small amount of data in low-light images, low-light images are synthesized artificially by gamma transform and Gaussian noise method to make data sets. YOLOv5s with high practicality was selected as the basic model of the experiment, and the model was further improved. The new loss function is used to calculate the loss of the boundary frame, and the convergence rate is faster for the model when training the invoice, which is helpful to improve the average detection accuracy of the model. Context transform network and void convolution are used in backbone network to extract image features better. The improved model performs better than the original model in accuracy, recall rate and average detection accuracy. The experiment shows that the image recognition effect is better for occlusion, dark environment and damage.

References

[1] Wang Xing, Zheng Yong-feng. Research on bill recognition algorithm based on OCR[J]. INTELLIGENT COMPUTER AND APPLICATIONS, 2021, 011 (011) : 101-106.

[2] Redmon J, Divvala S, Girshick R, et al. You Only Look Once: Unified, Real-Time Object Detection [C]// Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 779-788.

[3] LIU W, ANGUELOV D, ERHAN D, et al. SSD: Single shot multibox detector [C]//European conference on computer vision. Springer, Cham, 2016: 21-37.

[4] GIRSHICK R. Fast R-CNN [C]//Proceedings of 2015 TEEE International Conference on Computer Vision. Santiago: IEEE, 2015: 1440-1448.

[5] REN S Q, HE K M, GIRSHICK R, et al. Faster R-CNN; towards real-time object detection with region proposal network[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(6) : 1137-1149.

[6] Bochkovskiy A, Wang C Y, Liao H Y M. YOLOv4: Optimal speed and accuracy of object.

[7] Lin T Y, Dollar P , Girshick R , et al. Feature Pyramid Networks for Object Detection[C]// 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE Computer Society, 2017.

[8] Xu Yin-yun, Jiang Ming, Li Yun-fei, et al. Fruit target detection based on improved YOLO and NMS[J]. Journal of Electronic Measurement and Instrument, 2022, 036(004) : 114-123.

[9] Alsabhan Waleed, Alotaiby Turky. Automatic Building Extraction on Satellite Images Using Unet and ResNet50 [J]. Computational Intelligence and Neuroscience, 2022, 2022.

[10] Zhang Cheng-jun, Hu Xiao-bing, Niu Hong-chao. Vehicle object detection based on improved YOLOv5 method [J/OL]. Journal of Sichuan University(Natural Science Edition), 2022, 59(5) : 053 001.

[11] Li Y, Yao T, Pan Y, et al. Contextual Transformer Networks for Visual Recognition[J]. 2021.

[12] Shu Zi-ting, Zhang Ze-bin, Song Yao-zhe, et al. Low-light Image Object Detection Based on an Improved YOLOv5[J/OL]. Laser & Optoelectronics Progress, 2022-07-17.

[13] Zhang Rui-ping, Ning Qian, Lei Yin-jie, et al. Garbage detection based on Mask R-CNN[J/OL]. Computer Engineering & Science, 2022-09-22.

ISSN: 2688-8645

[14] Huang Z, Chen P , Wang P . System and method for semantic segmentation using hybrid dilated convolution (HDC), US11010616B2[P]. 2021.

[15] Xiao Y X, Jiang A W, Ye J H, et al. Making of Night Vision: Object Detection Under Low Illumination[J]. IEEE Access, 2020, (8): 123075-123086.

[16] Liu Hong-yu, Yuan Guo-yu. Detection of Cigarette Appearance Defects Based on Improved YOLOv5s. COMPUTER TECHNOLOGY AND DEVELOPMENT, 2022, 32(08) : 161-167.

[17] Zhang Tong, Meng Ling. Recognition of diabetic retinopathy based on attention neural network. Computer Engineering & Science,2022, 44(03) : 479-485.

[18] Sun Rui-bin, Qian Kui, Xu Wei-min, et al. Adaptive recognition of complex invoices based on Tesseract-OCR. Journal of Nanjing University of Information Science & Technology (Natural Science Edition), 2021, 13(03) : 349-354.