Hindawi
Mobile Information Systems
Volume 2021, Article ID 9374465, 13 pages
https://doi.org/10.1155/2021/9374465

Research Article

Surface Defect Detection of Nonburr Cylinder Liner Based on Improved YOLOv4

Yongbin Chen, Qinshen Fu, and Guitang Wang

Guangdong University of Technology, Guangzhou, China

Correspondence should be addressed to Guitang Wang; wanggt@gdut.edu.cn

Received 8 May 2021; Revised 1 June 2021; Accepted 18 June 2021; Published 29 June 2021

Academic Editor: Sang-Bing Tsai

Cylinder liner plays an important role in the internal combustion engine. The surface defects of cylinder liner will directly affect the safety and service life of the internal combustion engine. At present, the surface defect detection of cylinder liner mainly relies on manual visual inspection, which is easily affected by subjective factors of inspectors. Aiming at the bottleneck of traditional visual inspection technology in appearance inspection, this paper proposes a surface defect detection algorithm based on deep learning to realize defect location and classification. Based on the characteristics of the research object in this paper, the surface defect detection algorithm based on the improved YOLOv4 model is proposed, the model framework is constructed, and the data enhancement method and verification method are proposed. Experiments show that the proposed method can improve the detection accuracy and speed and can meet the requirements of the nonburr cylinder surface defect detection. At the same time, the method can be extended to other surface defect detection applications.

1. Introduction

Surface defects will directly affect the quality of the product, further affecting the chemical and physical properties of the product surface. As an important component of the internal combustion engine, the appearance of cylinder liner surface defects such as cracks and air holes will mean that there are major internal quality problems in the cylinder liner, which may lead to abnormal operation of the internal combustion engine and then lead to safety problems. Therefore, manufacturers and users put forward higher and higher requirements for the appearance quality of cylinder liner. At present, the detection of cylinder liner surface quality mainly depends on manual detection. The manual detection method not only cannot meet the production needs in terms of work efficiency but also is affected by the subjective experience of detection personnel. Some defects of products are small in size, so it is difficult for human eyes to observe these defects, and it is easy to miss inspection. At the same time, eyes working for a long time is harmful to the health of testing personnel. Therefore, manual detection cannot meet the requirements of the current mass industrial production. The rapid development of the image detection algorithm promotes the development of surface defect detection technology. Compared with manual detection, the detection technology based on machine vision not only improves the efficiency and accuracy of detection but also has the advantages of safety and reliability because of its noncontact. However, the traditional machine vision detection algorithm has poor flexibility in feature extraction, so it needs to build a feature extraction algorithm according to the type of surface defects of products. Because the shape and size of surface defects of industrial products are different, using an image algorithm for feature extraction needs a lot of resources for algorithm design, which shows that its universality for the target object is poor.

Compared with the traditional visual detection algorithm, the surface defect detection based on the deep learning algorithm not only shows high adaptability and stability but also has high detection accuracy in the face of changing scenes and targets. In this paper, a method of product surface defect detection based on deep learning is proposed, which is improved on the basis of YOLOv4 to make it more suitable for industrial product surface defect

detection. In this paper, two main surface defects "slag" and "sunken" in nonburr cylinder liner are defined as follows:

Sunken: the defects with a diameter more than 3 mm are judged as defects. There are three defects of 1 to 3 mm in the same field of vision, which can be judged as sunken.

Slag: no matter how small the size of slag inclusion is, it is not allowed to exist. Even if there is only one such defect on the surface of the cylinder liner, the product will be regarded as unqualified.

2. Related Work

The feature extraction ability of CNN is better than that of an artificial designed feature extraction operator, so using CNN for target detection has become a research hotspot in the field of contemporary target detection. At present, target detection algorithms based on the convolutional neural network can be divided into three categories: two-stage target detection algorithm, one-stage target detection algorithm, and anchor free target detection algorithm.

2.1. Two-Stage Target Detection Algorithm. At the 2014 CVPR (IEEE Conference on computer vision and pattern recognition) conference, Girshick et al. first proposed a two-stage target detection algorithm R-CNN [1] model. Later, a large number of scholars conducted research in this field and proposed SPPNet [2], Fast R-CNN [3], Faster R-CNN [4], and other two-stage target detection models. A two-stage serial target detection algorithm is formed. The flow chart of two-stage algorithms is shown in Figure 1. The two-stage target detection algorithm is mainly divided into two steps: using a specific algorithm to generate candidate regions, CNN is used to extract the features of the candidate regions to realize the classification of the candidate regions, and CNN is used to fine-tune the frame of the candidate regions to get the final detection results.

The network structure of R-CNN is shown in Figure 2. The model uses SS (Selective Search) [5] algorithm to generate candidate regions and then uses an image processing algorithm to scale the candidate regions to a fixed size. The processed regions are input into the designed CNN network for feature extraction, and the region classification is completed under the effect of the SVM classifier. At the same time, finish the fine-tuning of the border and finally get the target information. Although the accuracy of the algorithm is high, it takes a lot of computing time to generate candidate regions. Meanwhile, when the image processing algorithm is used to fix the size of the region, there is also the problem of image distortion, which leads to the confusion of information. In addition, a large number of candidate regions show the problem of computational redundancy in CNN.

In order to solve the problem of information loss caused by solidifying the size of candidate regions, Fast R-CNN improves R-CNN, and its network structure is shown in Figure 3. Different from R-CNN, Fast R-CNN inputs the whole image into CNN for calculation, and under the effect of ROI pooling, it fixes the output of CNN to a certain size of eigenvector. In this model, classification and regression are implemented in different networks, so although the detection accuracy is high, the detection speed is low.

Fast R-CNN does not solve the problem that it needs a lot of computation time to generate candidate regions. On the contrary, it also increases the computation of the model to a certain extent. In order to solve this problem, Ren S et al. proposed the target detection model of Faster R-CNN.

Different from the previous two models, under the effect of ROI pooling and the corresponding hardware conditions, the model can accept any size of the input image. The model designs the backbone network to extract the features of the input image to get the corresponding feature map, which is shared by the RPN and the full connection layer of the surface, reducing the amount of calculation to a certain extent. In order to solve the time problem of generating candidate regions, an RPN network is designed to generate candidate regions. Its structure is shown in Figure 4, and the subsequent processing is the same as fast R-CNN.

RPN network is a full convolution model. Through the application of the RPN network, the extraction efficiency of candidate boxes is greatly improved. In the process of using the RPN network, Anchor mechanism and NMS algorithm are used. Under the effect of these algorithms, the accuracy of the model is improved.

In addition, many scholars have been studying in this direction, and based on Faster R-CNN, they have proposed better detection algorithms, such as Mask R-CNN [6] and Cascade R-CNN [7].

2.2. One-Stage Target Detection Algorithm. Although the two-stage target detection model has been improved a lot and the detection accuracy has been greatly improved, due to its complex model, the parameters of the model are too many and the training time is too long. Moreover, this kind of algorithm divides the classification and regression into two parts, which leads to low time efficiency in calculation. At the same time, the region recommendation algorithm still brings an extra burden to the calculation. For this reason, Redmon et al. proposed the YOLOv1 algorithm [8]. The algorithm successfully integrates regression and classification tasks into the overall CNN structure and obtains the target category information and location information directly through a convolution neural network. Under the action of the Anchor mechanism, the region recommendation algorithm is canceled. These two improvements greatly improve the detection efficiency of the network. In terms of real-time performance and accuracy, the YOLOv1 algorithm can better meet the needs of industrial detection.

Since the introduction of YOLOv1, many scholars have done a lot of research on the target detection algorithms in this field, and there have been some classic one-stage target detection algorithms such as YOLOv2 [9], SSD [10], and YOLOv3 [11], among which YOLO series is the main representative.

In order to improve the accuracy of the model, based on the YOLOv1, the YOLOv2 algorithm is improved by introducing Batch Normalization, anchor box mechanism,

Figure 1: Two-stage series algorithm flow chart.

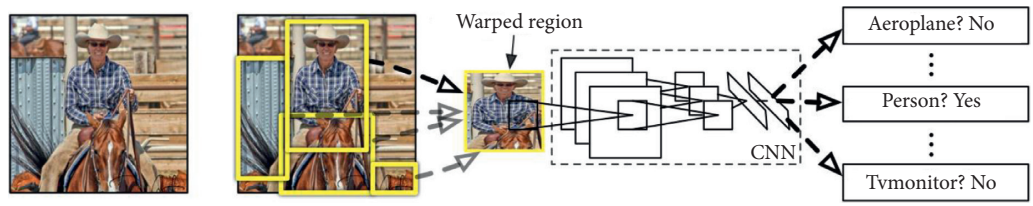


Figure 2: R-CNN network structure.

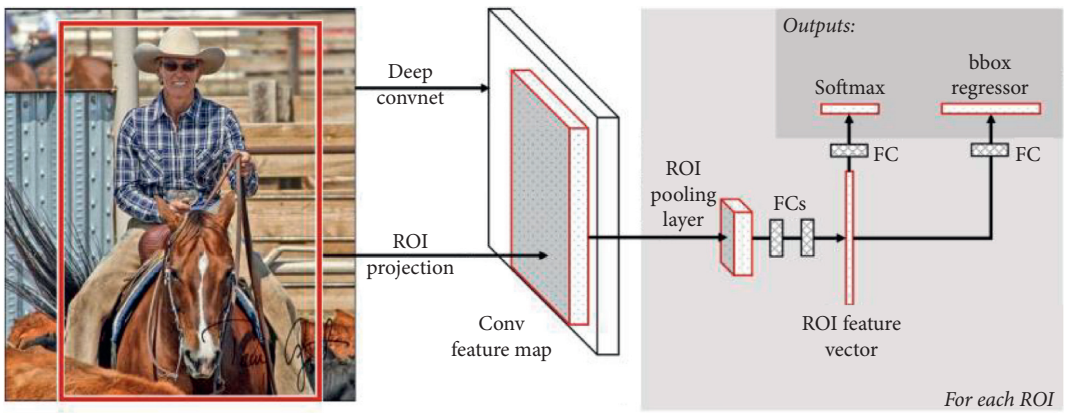


Figure 3: Fast R-CNN network structure.

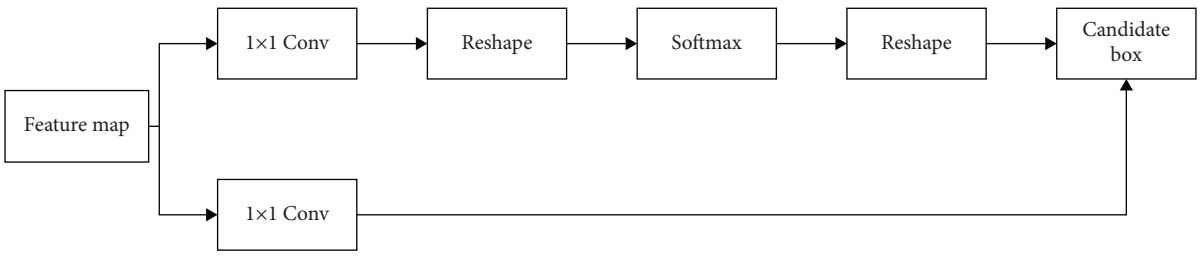


Figure 4: RPN structure.

multiscale training method, absolute position prediction, and so on. At the same time, the dropout mechanism is canceled. These improvements solve the problem of overfitting to a great extent. Meanwhile, the anchor mechanism is used to set up nine candidate windows in each center of the grid. These candidate windows are called anchor, and the width and height of an anchor are not supposed to be set, but the results are obtained by K-means clustering of the training set.

In order to improve the detection ability of small targets, a more refined detection algorithm of YOLOv3 is proposed based on YOLOv2. The new backbone network DarkNet-53 is adopted in the model. The backbone network improves the feature extraction ability of the model. At the same time,

under the ideas of DenseNet and FPN, the model can detect small targets more accurately, the feature pyramid is formed, and the fusion between features is realized, which expands the semantic information of low feature level. Moreover, the YOLOv3 algorithm improves the loss function. In the part of category loss, logical regression is used to replace the softmax function.

2.3. Target Detection Algorithm Based on Anchor Free. At present, the target detection algorithms mainly rely on anchor mechanism, but the regression box sizes of different targets are different. Fixed anchor is not particularly suitable for target detection of different sizes. In order to adapt to various sizes of objects, a large number of scholars have done a lot of research on anchor free direction. This kind of algorithm transforms the idea of box regression into the idea of key point regression.

Among them, the more typical model structures are the CornerNet model [12] and CenterNet model [13]. The structure of the CornerNet model is shown in Figure 5. In CornerNet, the input image is input into a convolution neural network to obtain the feature map of certain semantic information. At the same time, the feature map is input into two different branches to predict the coordinate information of the upper left corner and the lower right corner of the target box, respectively. These two branches will pass through the network of corner pooling, and finally get three different output results, which are heat map, offset, and embedding information. The coordinates of the upper left corner and the lower right corner of the object frame are obtained from the thermal graph. Meanwhile, the coordinates of the corner are fine-tuned according to the offset. Then, the coordinates of the upper left corner and the lower right corner of the same object are matched according to the embedding information to get the target frame.

This section mainly introduces the basic theory and typical architecture of CNN, further introduces the typical target detection algorithm based on CNN and the corresponding development trend, and analyzes various target detection algorithms, which provides a theoretical basis for the research of this paper.

3. System Architecture

3.1. Improvement of Backbone Network. As shown in Figure 6, it is the system framework of YOLOv4, in which CSPDarkNet is the backbone network. In order to improve the feature extraction ability of the backbone network, this study modified the backbone network of YOLOv4. The backbone network of YOLOv4 is mainly composed of five CSPNets, and each CSPNet is composed of several residual blocks. In order to further improve the feature extraction ability of the backbone network to the target information, the main network of YOLOv4 is composed of five CSPNets. At the same time, the backbone network should pay more attention to the global characteristics.

In January 2021, Google proposed a new self-attention structure, Bottleneck Transformer (BoT) [14]. This structure studied CNN and transformer and got better results after combining them. However, Google only replaced the convolution block of 3×3 with this structure in the last three residual blocks of Resnet50, and it not only greatly improves the feature extraction ability of the network but also reduces the network parameters. The parameters are 0.833 times the original Resnet50 model, and the calculation time of the model is 2.33 times faster than that of EfficientNet.

Therefore, in this study, the first four CSPNet structures of the YOLOv4 backbone network are retained, and only the last CSPNet structure is modified. In order not to increase the difficulty of network training, the first three residual structures are retained in the last CSPNet, only the last residual structure is adjusted, and the last residual structure is changed to a Bottleneck Transformer structure. The core component of the structure is Multi-Head Self-Attention (MHSA) structure, and its structure is shown in Figure 7.

The structure of MHSA [15] is relatively complex, which mainly uses the mathematical operation between matrices and the convolution operation of 1×1. When the structure receives the input x with the input dimension of $H \times W \times d$, it initializes two parameters $R_h$and R_w. The dimensions are $H \times 1 \times d$and $1 \times W \times d$, where $H$and $W$are the length and width of input x, and $D$is the number of channels of input x. These two are added by the broadcast mechanism to obtain spatial location information. The input $x$passes through three 1×1 convolution layers, and three outputs q, k, and $v$are obtained, where $q$represents the query value, $k$represents the key value, and $v$represents the value of attention function. The matrix dot product between $q$and $k$is calculated, and the correlation between $q$and $k$is obtained, which is content-content. Simultaneously, q, R_h, and $R_w$are calculated to get the location information of the query value in the space, that is, content-position. After adding content-content and content-position, the probability distribution of attention is obtained by using the softmax activation function. The matrix dot product of the probability distribution and $v$is calculated to get the attention degree of input information under the attention distribution.

Based on the MHSA structure, a new residual block-Bottleneck is constructed to replace the original residual block. Its structure is shown in Figure 8. It can be seen that it is not much different from the original residual block. The bigger difference is that the convolution structure of 3×3 is replaced by the MHSA structure.

In the Bottleneck structure, the convolution layer of 1×1 is always reserved. The main function of this structure is to transform the channel dimension. We can see that the convolution layer of 3×3 is changed into an MHSA module in this structure. At the same time, after the result of the short connection part of the residual structure is obtained, the Mish activation function is used for processing. The result of the backbone network obtained by replacing the last residual block in the last SCPNet in the backbone network with the Bottleneck structure is shown in Figure 9. The last residual block is changed into the Bottleneck structure. The purpose of this modification is to reduce the training difficulty, and changing all the residual blocks into a Bottleneck structure will greatly increase the training difficulty of the model.

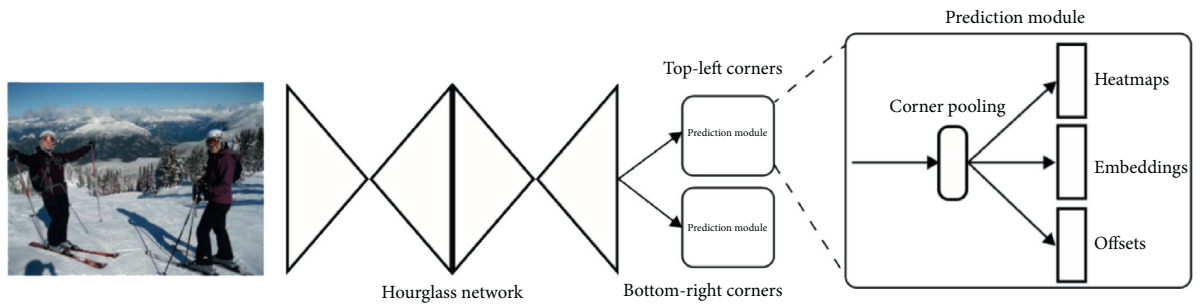


Figure 5: CornerNet model structure.

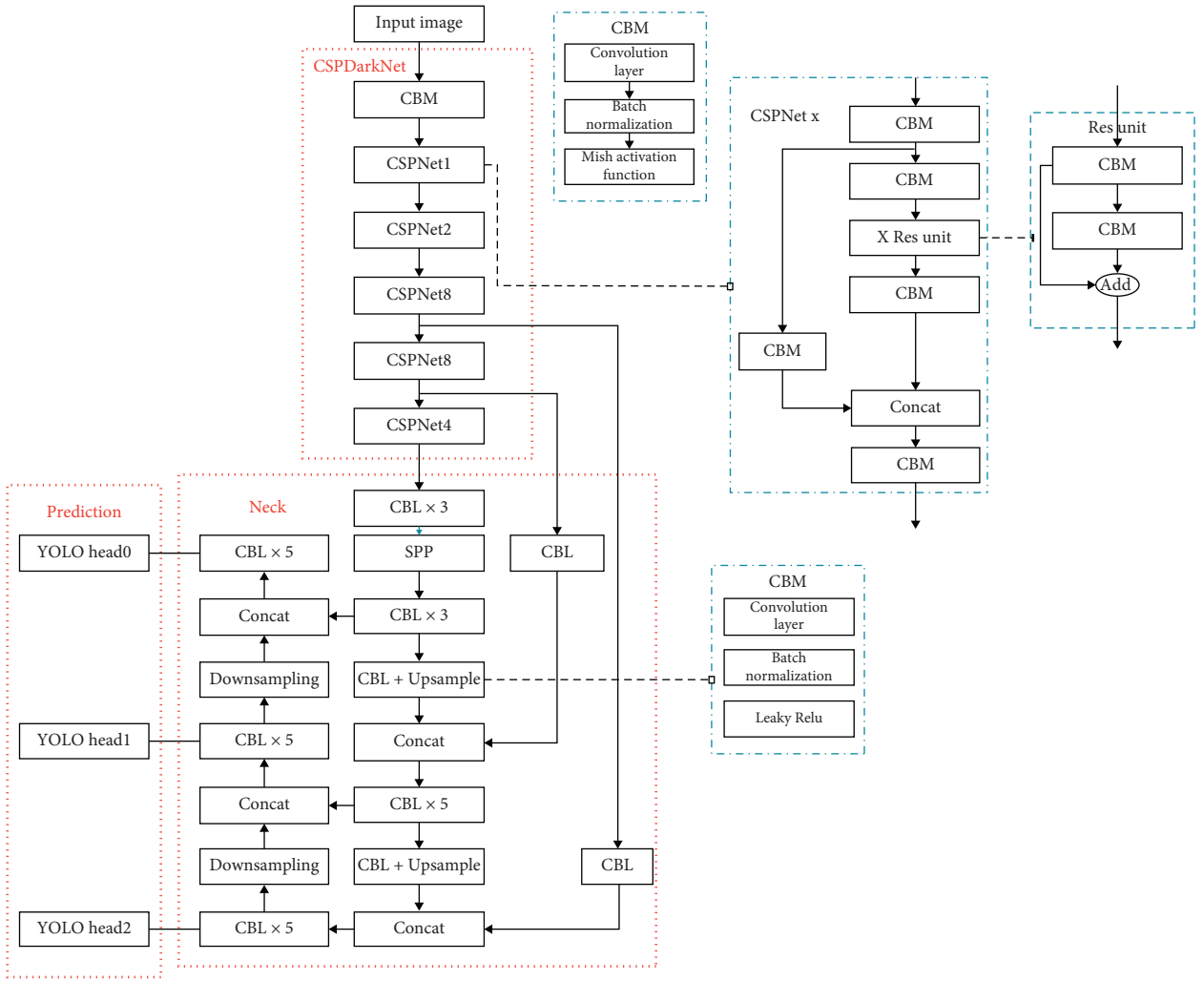


Figure 6: YOLOv4 framework.

3.2. Establishment of the Improved Model. Combined with the previous improvement method, the overall model of YOLOv4 is improved, the backbone network part of the model is improved accordingly, and the nonburr cylinder liner surface defect detection model of this study is designed. In this study, an attention mechanism is used in the backbone network part to enhance the feature extraction ability of the backbone network. The Bottleneck structure

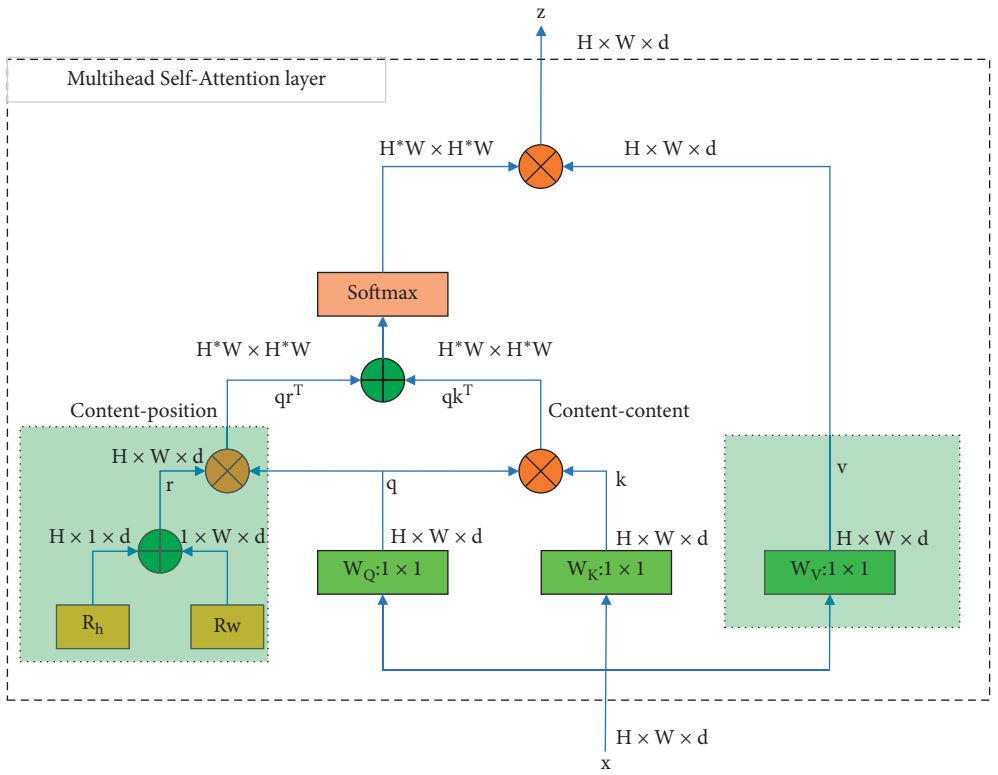


Figure 7: MHSA structure.

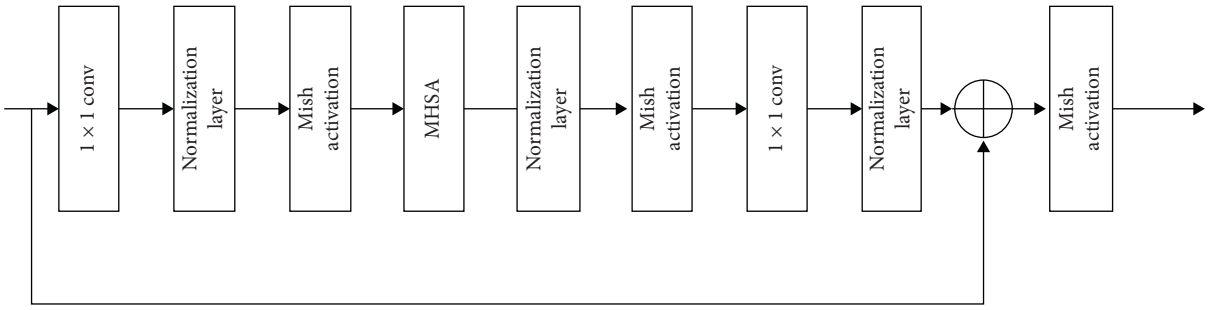


Figure 8: Bottleneck structure.

used in the backbone network is a self-attention mechanism. Its main function is to capture the global feature information. Its model structure is shown in Figure 10, and its red part is the part to modify the model.

4. Experiment

4.1. Image Dataset Enhancement. The subsequent training of the detection model needs a large number of data samples, and the amount of image data is often limited. In this study, 1000 images were collected through the image acquisition system, but more data is helpful to the network training and generalization ability. In order to further expand the image dataset, some image transformation is needed, such as image flipping and image rotation, which makes the model more robust.

Image flipping, also known as image mirroring, is generally divided into two types: mirror transformation in horizontal position and mirror transformation in vertical position. Both of them take the middle axis of the image as the center of transformation. The mathematical formula of the horizontal mirror is shown as follows:

$$\begin{bmatrix} t_x \\ t_y \\ 1 \end{bmatrix} = \begin{bmatrix} -1 & 0 & \text{width} \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix}, \quad (1)$$

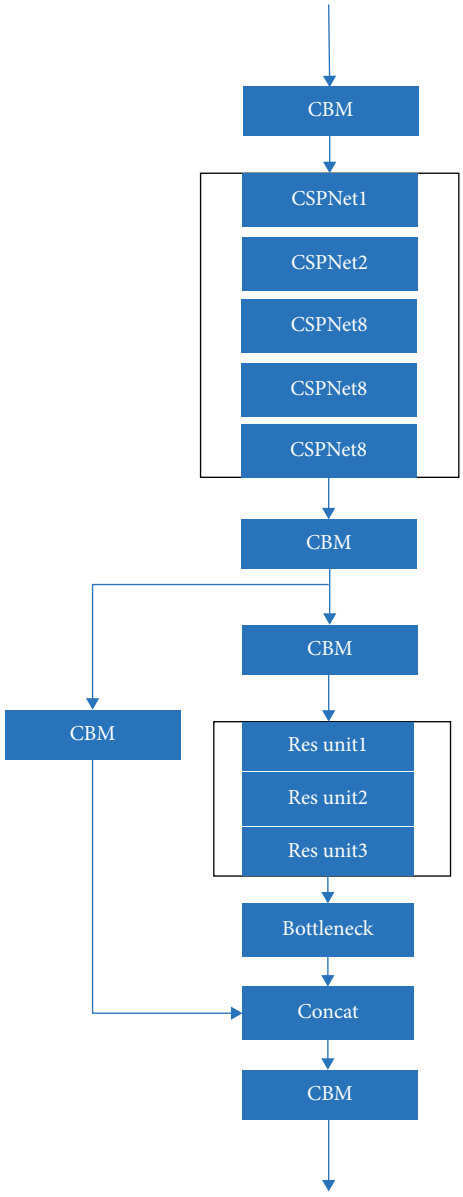


Figure 9: Improved backbone network.

where (x, y)represents a point in the original image, the width of the image is Width, the height is Height, (t_x, t_y)represents a point in the transformed image, and a point in the original image becomes $(\text{Width} - x, y)$after transformation. Similarly, in a vertical mirror image, a point in the original image becomes $(x, \text{Height} - y)$after transformation.

After image mirroring, the result is shown in Figure 11.

Image rotation is also a relatively complex geometric transformation. Different from image mirroring, image rotation takes the image center as the rotation origin, and all pixels on the image rotate at the same angle. At the same time, the size of the image will generally change after rotation. The mathematical formula is shown as follows:

$$\begin{bmatrix} t_x \\ t_y \\ 1 \end{bmatrix} = \begin{bmatrix} \cos\theta & \sin\theta & 0 \\ -\sin\theta & \cos\theta & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix}. \tag{2}$$

Set the rotation angle to 30 degrees and 90 degrees. After image rotation, the result of the cylinder liner surface image is shown in Figure 12.

After image enhancement, the defect detection dataset is constructed. In this paper, the image input size of the detection model is 416×416, so it is necessary to capture the surface image of the cylinder liner to obtain the image data of the corresponding size. This study mainly focuses on the detection of two main defects of the needleless cylinder liner, which are sunken and slag. The camera was used to collect 800 images of each category. After image data enhancement, the number of images of each category is doubled, and the training set and test set are divided according to the ratio of 7 : 3. In this study, the annotation format is Pascal VOC dataset annotation format, and the annotated file is saved in XML format. In Pascal VOC format, the location information of defects in the image is saved in the form of upper left corner coordinates and lower right corner coordinates. In this study, the position format of the real box of the model is in the form of the coordinates of the center point and the width and height of the box. Therefore, we need to make a certain transformation. The transformation process of the Pascal VOC format to the format of the model is shown as follows:

$$\begin{aligned} x_{\text{center}} &= \frac{x_{\max} - x_{\min}}{2}, \\ y_{\text{center}} &= \frac{y_{\max} - y_{\min}}{2}, \end{aligned} \tag{3}$$

$$\begin{aligned} w &= \frac{x_{\max} + x_{\min}}{2}, \\ h &= \frac{y_{\max} + y_{\min}}{2}. \end{aligned} \tag{4}$$

4.2. Software and Hardware Platforms. For the defect detection model designed in front, in order to verify its performance, this study built the corresponding experimental platform, which is composed of hardware platform and software platform. The detailed configuration is shown in Tables 1 and 2.

On the basis of the construction of the dataset and the construction of the experimental platform, the built defect detection model is trained. The specific hyperparameters are shown in Table 3. The batch size represents the training required for each iteration of the model. The number of images and the number of training (Epoch) represent that the model uses all the training image data for forward and backward propagation. The optimizer in this study chooses Adam to iteratively optimize the model, and the initial learning rate of the model is 1e-3.

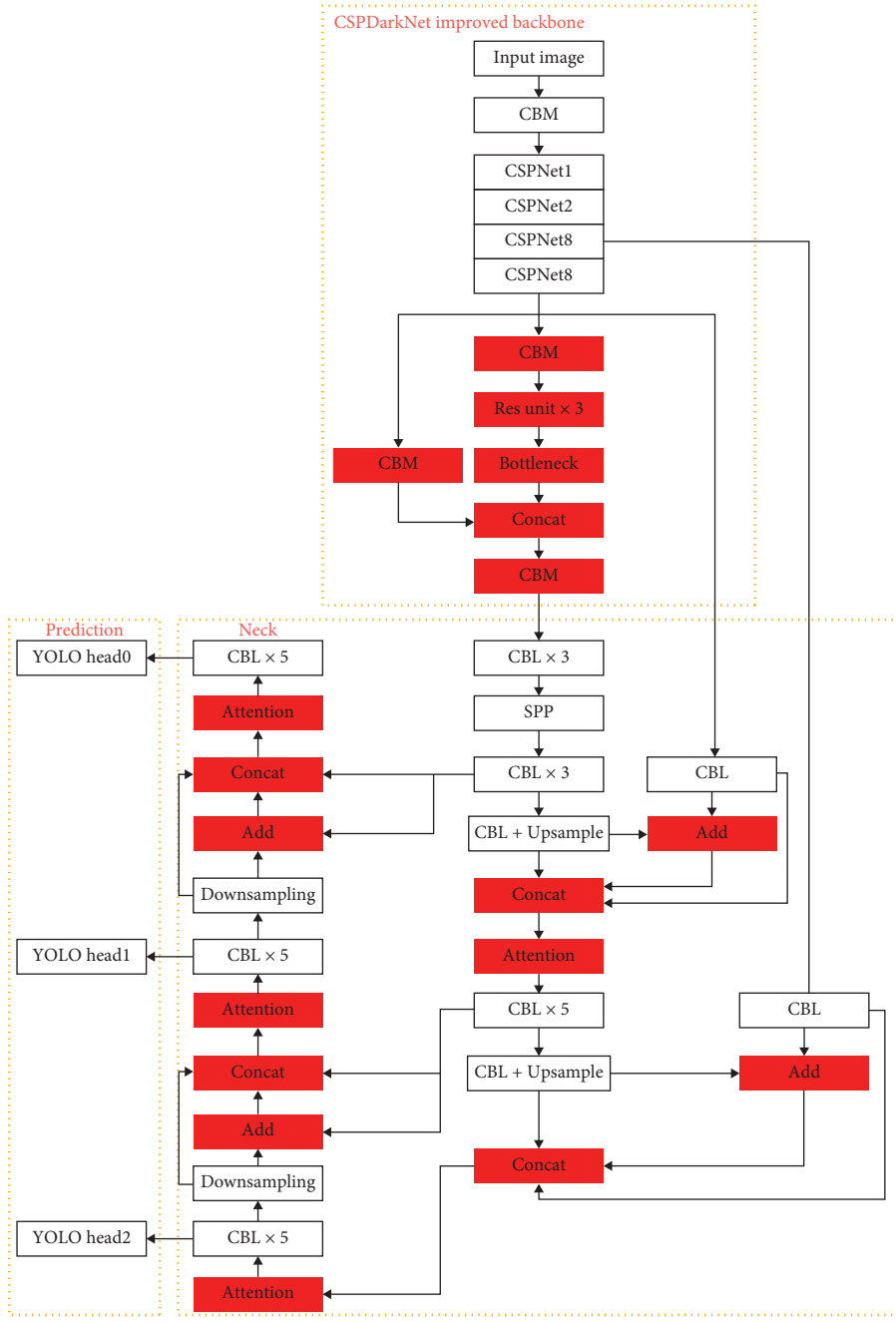


Figure 10: Improved model structure.

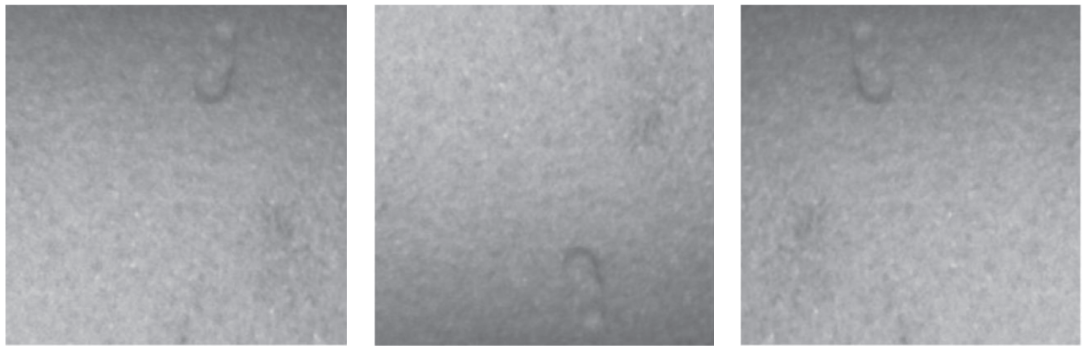
4.3. Evaluation Standard. After training, the model needs to be evaluated accordingly. The evaluation indicators of the target detection model mainly include precision, recall, and the mean average precision and the FPS (frame rate). Among them, several indicators are needed to measure the classification accuracy of the model, which can be represented by the confusion matrix in Table 4. The specific meanings are as follows:

(1) True positive (TP): the samples that are actually positive are correctly classified and predicted as positive examples, which can be understood as the number of detections frames whose IOU with the real frame is greater than the threshold.

(2) True negatives (TN): samples that are actually negative classes are correctly classified and predicted as negative examples.

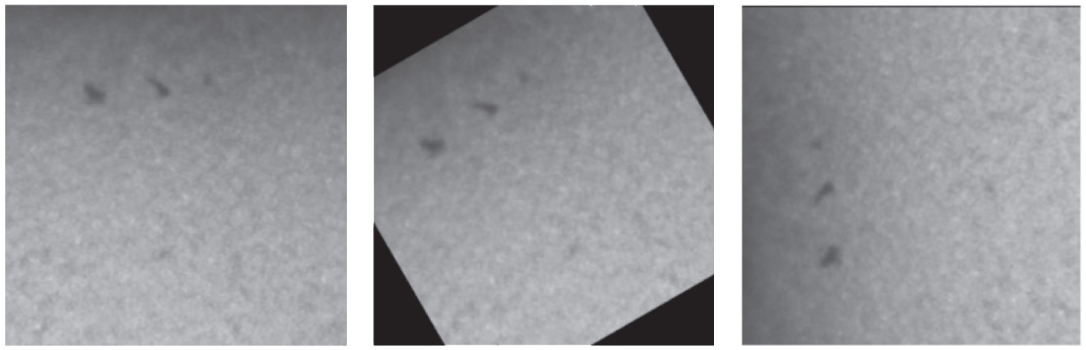
(3) False positives (FP): the number of positive samples that are actually negative but predicted by misclassification.

(4) False negatives (FN): actually, the number of positive samples predicted to be negative by misclassification,



(a) (b) (c)

Figure 11: Mirror effect example. (a) Origin image. (b) Vertical mirror. (c) Horizontal mirror.



(a) (b) (c)

Figure 12: Rotation effect example. (a) Origin image. (b) 30° rotation. (c) 90° rotation.

Table 1: Deep learning server configuration information.

Hardware	Type/num.
CPU	Intel Core i7-7700K/1
Mainboard	PRIME Z270-A/1
RAM	16G DDR4 2400 MHz/1
ROM	2 TB/1
GPU	GeForce GTX 1080Ti/1

Table 2: Software system information.

Software	Type
Operating system	Ubuntu 18.04
Programming platform	Pycharm Community, Anaconda3
Programming language	Python 3.7.0
Deep learning framework	Pytorch 1.4.0
Others	Numpy, math, PTL, pandas

Table 3: Model training hyperparameters.

Type	Data
Number of batches	8
Training times	300
Optimizer	Adam
Learning rate	1e-3

Table 4: Confusion matrix.

Truth	Prediction	
	Positive	Negative
True	TP	FN
False	FP	TN

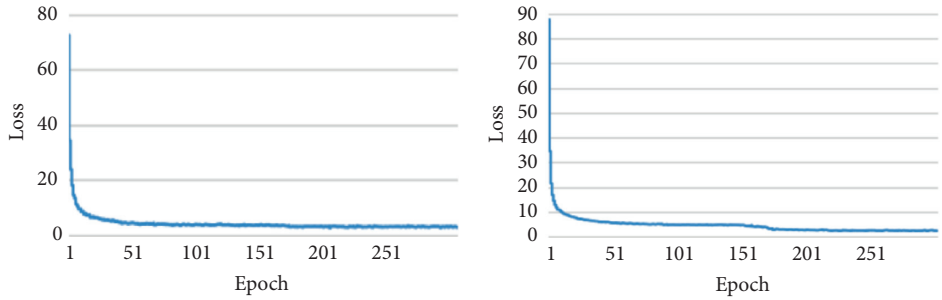


Figure 13: Changes in the training loss function of each model. (a) Basic model. (b) Improved backbone.

which can be understood as the number of true frames that have not been detected by the model.

According to the confusion matrix calculated from the detection results, the precision and recall of the model can be calculated. The calculations are shown in equations (5) and (6). Precision represents the proportion of the true class in the sample predicted to be the positive class in the prediction result; the focus is on the "precision" aspect. Recall indicates how many positive classes are detected in the sample, and the focus is on the "recall" aspect:

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \tag{5}$$

$$\text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \tag{6}$$

Under the condition of different confidence thresholds, different categories of precision and recall are obtained to form coordinate points, recall is taken as abscissa and precision as ordinate to draw the curve, the area under the curve is the average precision of the category, and the calculation process is shown in equation (7). The AP value of N categories can be calculated, and the final map value can be obtained after taking the average value of them. The calculation process is shown in equation (8):

$$\text{AP}_{\text{singleclass}} = \int_0^1 \text{precision}\,(\text{recall})\,d\,(\text{recall}), \tag{7}$$

$$m\text{AP} = \frac{1}{N}\sum_{i=0}^{N} \text{AP}_i. \tag{8}$$

FPS is a measure of the detection speed of the model. For industrial detection, the detection needs to meet real-time requirements. FPS indicates how many images the model can process in one second. Therefore, FPS is an important performance index for this study.

4.4. Experimental Results. In order to verify the role of each improved module in the model, on the basis of the image dataset of cylinder liner surface defects, two groups of experiments are compared, which are the basic detection model of YOLOv4 and the improved backbone network of YOLOv4. The two groups of experiments are carried out on the same experimental platform, and the super parameter settings are consistent. The change of loss curve in the training process of the model is shown in Figure 13. It can be found that the improved model can basically converge in the training process, and the convergence speed of the improved model is faster than that of the original model.

The above two models can basically converge after 300-epoch iterative training, so it is necessary to calculate and evaluate the corresponding evaluation indexes of the converged detection model. Firstly, the accuracy and recall of each model are calculated. The classification accuracy curve and recall rate curve of the two categories in the basic model under different thresholds are shown in Figures 14and 15. The classification accuracy curve and recall rate curve of the two categories of the improved model at different thresholds are shown in Figures 16and 17.

Because of the opposite relationship between precision and recall rate, although precision has been improved under a certain threshold, the recall rate shows a downward trend. It can be seen that it is difficult to make the final evaluation of the model according to the classification accuracy and recall curve of each model category. It is necessary to comprehensively consider the classification accuracy and recall rate to calculate the AP value of each category and the final map value of each model. The calculation results of the AP value are shown in Table 5.

The above experimental results are all the results of evaluating the model under the same conditions. Compared with the basic model, there are 5.98% and 6.07% improvements in the final model. With the gradual improvement of the model, the detection accuracy has been greatly improved. As shown in Figure 18, for the final evaluation index, map, the improved model brings 3.765% improvement.

The detection speed results of the two test models are shown in Table 6. It can be seen that the improvement of model accuracy does not sacrifice too much detection speed, which basically meets the requirements of real-time detection.

In order to make an intuitive comparison, the test results of each model are expanded, as shown in Figure 19. Different

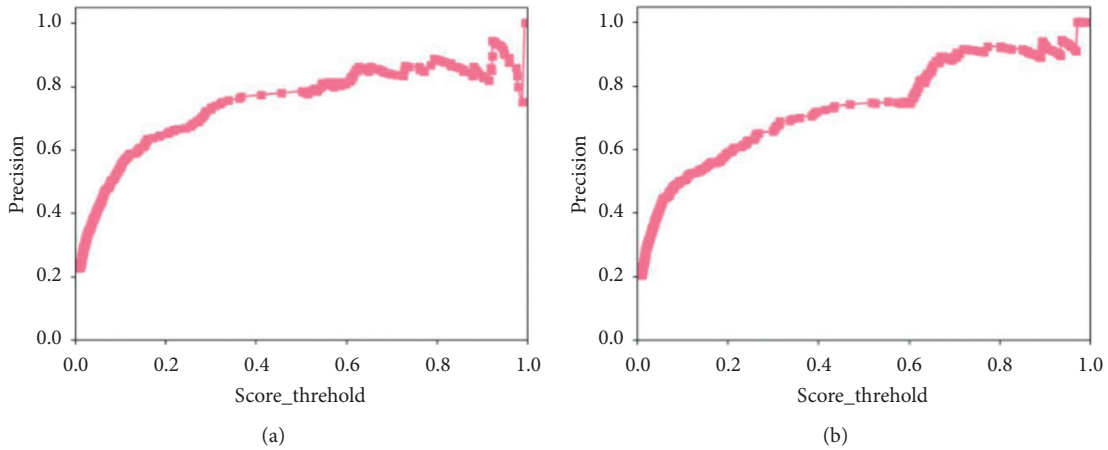


Figure 14: Basic model precision curve. (a) Slag. (b) Sunken.

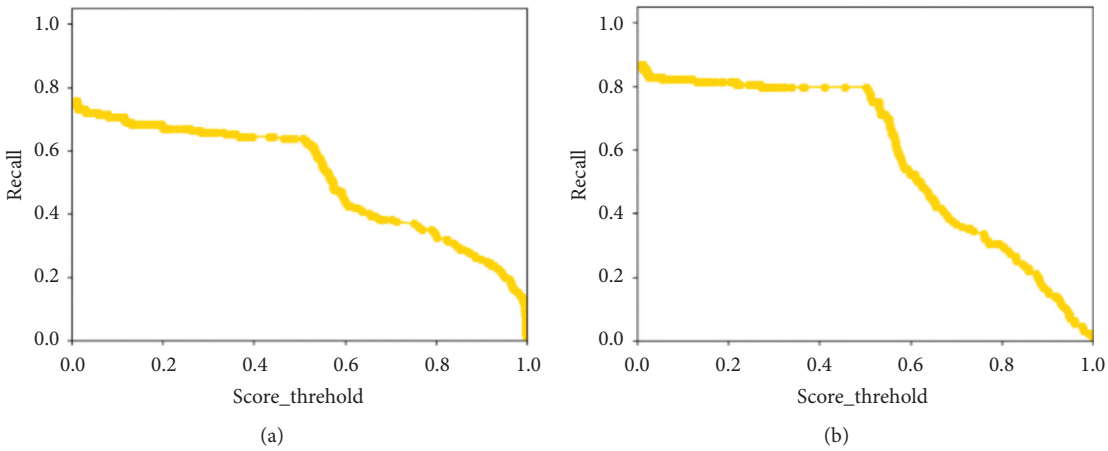


Figure 15: Basic model recall curve. (a) Slag. (b) Sunken.

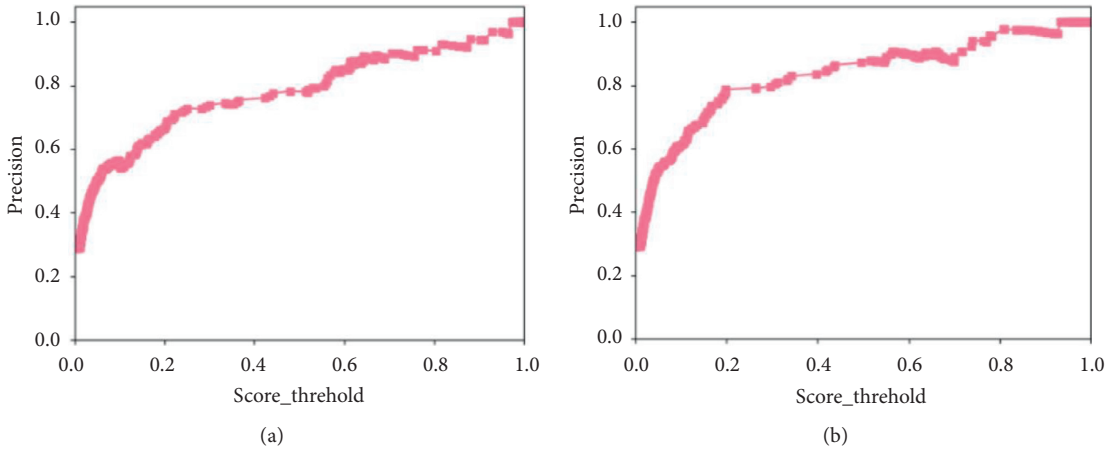


Figure 16: Improved backbone model precision curve. (a) Slag. (b) Sunken.

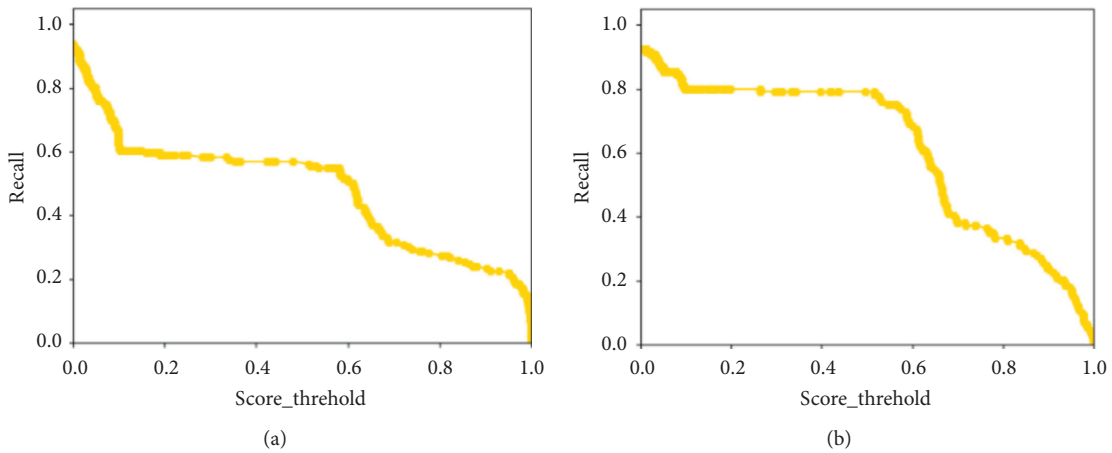


Figure 17: Improved backbone model recall curve. (a) Slag. (b) Sunken.

Table 5: AP index comparison.

Detect method	Input size	Slag (AP%)	Sunken (AP%)
Basic model	416 × 416	63.26	71.66
Improved backbone	416 × 416	69.24	77.73

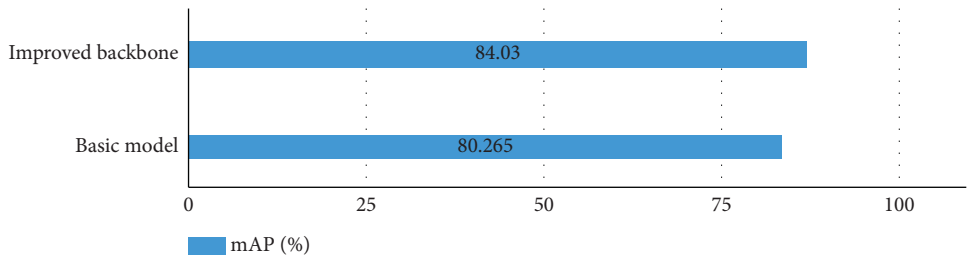


Figure 18: MAP results of each model.

Table 6: Detection speed comparison.

	Basic model	Improved backbone
FPS	40.53	38.056

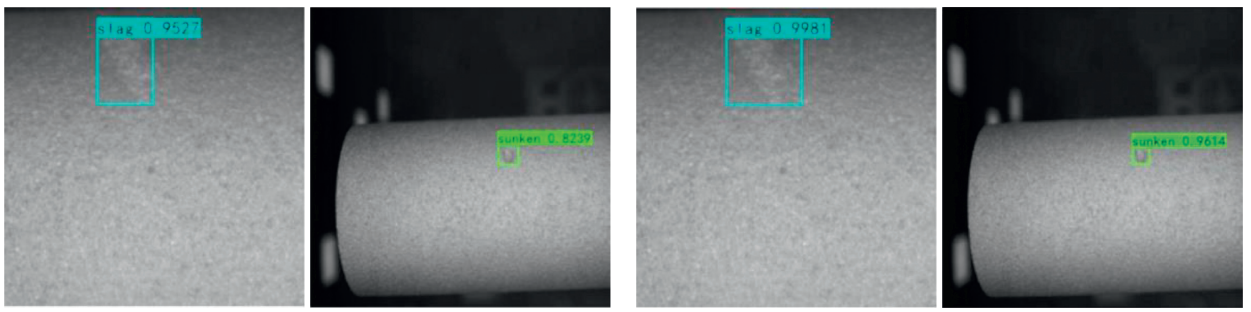


Figure 19: Comparison of test results of models. (a) Basic model. (b) Improved backbone.

defects are marked with different color detection boxes. On the whole, the detection effect of the improved YOLOv4 model of the backbone network in this study is the best.

5. Conclusion

This paper takes the nonburr cylinder liner surface defect detection as the research object and introduces the principle of the YOLOv4 detection model. Based on YOLOv4, the attention mechanism and feature fusion module are improved, and the improved algorithm model of this research is designed. The training optimization method and testing process of the algorithm model are explained. At the same time, the experimental platform environment of the detection model of this research is explained. The evaluation standard of the model is introduced, three sets of comparative experiments are carried out according to the improved module, and the model is evaluated according to the evaluation standard. The performance is evaluated, and the defect types such as slag and sunken surface of the cylinder liner are experimentally verified. Experimental results show that the method proposed in this paper can effectively improve the accuracy of surface defect detection of cylinder liners, and this method can be extended to other surface defect detection applications.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

[1] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 580–587, Columbus, OH, USA, June 2014.

[2] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 9, pp. 1904–1916, 2015.

[3] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Region-based convolutional networks for accurate object detection and segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 1, pp. 142–158, 2015.

[4] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-Cnn: towards real-time object detection with region proposal networks," 2015, https://arxiv.org/abs/1506.01497.

[5] J. R. Uijlings, K. E. Van De Sande, T. Gevers, and A. W. Smeulders, "Selective search for object recognition," *International Journal of Computer Vision*, vol. 104, no. 2, pp. 154–171, 2013.

[6] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask r-cnn," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2961–2969, Venice, Italy, December 2017.

[7] Z. Cai and N. Vasconcelos, "Cascade r-cnn: delving into high quality object detection," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6154–6162, Salt Lake City, UT, USA, December 2018.

[8] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: unified, real-time object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 779–788, Las Vegas, NV, USA, June 2016.

[9] J. Redmon and A. Farhadi, "YOLO9000: better, faster, stronger," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7263–7271, Honolulu, HI, USA, July 2017.

[10] W. Liu, D. Anguelov, D. Erhan et al., *Ssd: Single Shot Multibox Detector. European Conference on Computer Vision*, Springer, Berlin, Germany, 2016.

[11] J. Redmon and A. Farhadi, "Yolov3: an incremental improvement," 2018, https://arxiv.org/abs/1804.02767.

[12] H. Law and J. Deng, "Cornernet: detecting objects as paired keypoints," in *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 734–750, Munich, Germany, September 2018.

[13] K. Duan, S. Bai, L. Xie, H. Qi, Q. Huang, and Q. Tian, "Centernet: keypoint triplets for object detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision; 2019*, pp. 6569–6578, Seoul, South Korea, February 2019.

[14] A. Srinivas, T.-Y. Lin, N. Parmar, J. Shlens, P. Abbeel, and A. Vaswani, "Bottleneck transformers for visual recognition," 2021, https://arxiv.org/abs/2101.11605w.

[15] X. Xiao, D. Zhang, G. Hu, Y. Jiang, and S. Xia, "CNN-MHSA: a convolutional neural network and multi-head self-attention combined approach for detecting phishing websites," *Neural Networks*, vol. 125, pp. 303–312, 2020.