Research on automobile bearing defect detection algorithm based on improved YOLOv7-tiny

Yu Xu[1, 2,a], Shan Hu[1, 2,*] , Rui Ming[1, 2,b,] TianZhi Zhang[1, 2,c]

[1]School of Automation and Electrical Engineering Tianjin University of Technology and Education, Tianjin, China

[2]Key Laboratory of Information Sensing &Intelligent Control, Tianjin, China

[a]1186502057@qq.com,[*]lakeshu@163.com,[b]1061106174@qq.com,[c]1360342013@qq.com

Abstract

As an important part of the auto parts, the quality of the automobile bearings directly affects the performance and safety of the automobile. Therefore, it is particularly important to test the automobile bearing defects. The traditional automobile bearing defect detection method mainly relies on manual visual inspection, which is not only inefficient, but also easily affected by human factors and fatigue degree. The existing defect detection algorithm is also difficult to balance the identification accuracy and the identification accuracy, especially for small target defects, which are missing and missing, so there are some limitations. In order to meet the requirements of the automation and intelligence of the modern automobile manufacturing industry, this paper adopts the YOLOv7-tiny algorithm with faster speed and flexible network structure. Based on this method, SiLU and SIoU are first used as the activation function and the border regression loss function to accelerate the fast convergence of the network and improve the localization accuracy. Secondly, the ACmix module integrating self-attention mechanism and convolution structure is added to the head layer to strengthen the ability of the network to capture feature information and improve the recognition accuracy of the network model. Meanwhile, in order to reduce the feature loss of key information, the ASFF adaptive spatial feature fusion network is introduced. The network integrates feature maps at different levels according to the learning weight parameters to enhance the accuracy of identifying hidden and overlapping small targets. Through experiments on the bearing defect data set, the improved YOLOv7-tiny-SSAA improves by 5.2% compared with the average detection accuracy (mAP) of the original algorithm, and the detection speed is 63 fps, which can meet the defect detection tasks in various scenarios.

Keywords

Surface defect detection; YOLO; attention mechanism; feature fusion.

1. Introduction

Under the background of the rapid development of science and technology, the automobile industry is facing unprecedented challenges and opportunities. As the core component of the automobile, the quality of the bearing is directly related to the safety, stability and service life of the automobile[1][2]. Therefore, the defect detection of automobile bearings is not only of far-reaching significance, but also an inevitable requirement of our times.With the continuous intensification of market competition, in order to meet the increasing demand of consumers for automobile performance, the standard of bearing precision and performance is also increasingly strict. Once the bearing is defective, it will not only affect the performance of the car, but also may lead to serious safety risks, threatening the safety of people's life and property. Therefore, the defect detection of bearings is an important means to ensure the product quality and safety, and it is also a key link for enterprises to win the market competition. At present, most parts manufacturers generally adopt the method of manual detection for product defect detection, which not only costs labor cost, but also is judged by the judgment of subjective consciousness, which is easy to miss and miss inspection, and has great limitations. Since the 20th century, thanks to the rapid development of sensor technology and computer hardware, there have been some non-contact detection, such as magnetic particle detection, eddy current detection and machine vision technology, which greatly improves the detection efficiency and saves the labor cost[3]. In recent years, with the rapid development of the Internet and big data and the improvement of hardware computing power, deep learning technology has been strongly promoted. Many achievements have been made in image recognition, image segmentation, target detection and tracking, and super resolution[4]. Deep learning can independently learn defect feature information of different kinds, sizes and different shapes by training data sets containing defect location and label information, etc. For the multi-scale, irregular defect detection task, it has a high detection rate[5]. At the present stage, most of the studies on defect detection tasks are developed around deep learning, mainly including the one-stage and two-stage defect detection algorithms. One stage of the YOLO algorithm has gone through many versions and is widely used[6][7][8]. At present, the detection speed, the detection accuracy of smaller targets and overlapping occlusion targets are still difficult points to be solved in the industrial defect detection task. For the actual automobile production line, it is necessary to consider whether the performance of the selected algorithm meets the requirements in different backgrounds and environments. So in order to solve the reality complex scenario real-time detection of small target and shielding area false inspection, missed problems, this paper puts forward a small target based on improved YOLOv7-tiny algorithm defect detection method, can achieve the purpose of speed up the quality inspection efficiency, improve the accuracy, improve auto parts manufacturing enterprise automation, intelligent degree, timely detection eliminate unqualified parts, further release the production potential, promote the continuous innovation and development of the automobile industry.

2. The algorithm improves the design

2.1. Optimization of the activation function

The activation function functions to convert the linear transformed output into a nonlinear form, thus giving the network more expressive power. Commonly used activation functions mainly include Sigmoid[9], Tanh[10], ReLU[11], etc. The original YOLOv7-tiny algorithm uses the Leaky ReLU activation function, which aims to solve the problem of neuronal death caused by ReLU. The Leaky ReLU function allows the introduction of a very small constant leak when the input is negative, so that the neurons can have a certain gradient even the negative input during the training process. In this paper, the Leaky ReLU activation function in the original algorithm is replaced with the SiLU activation function, because the training effect of the SiLU activation function on the deep model is better than the Leaky ReLU activation function, which can effectively inhibit the overfitting phenomenon. Both function images are shown in Figure 1. It can be seen that the SiLU function has no upper bound, lower bound, smooth and non-monotonic characteristics. Its derivatives exist in the whole range of real numbers, which helps to optimize the gradient descent more stably and avoid the saturation phenomenon caused by the gradient approaching 0 during the training period.

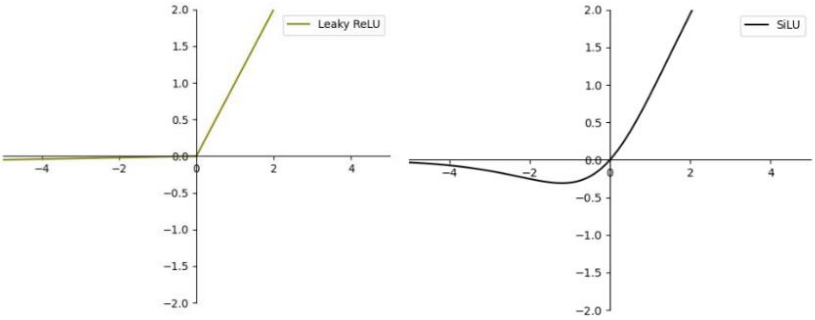


Figure 1. Activation function image

2.2. Optimization of the loss function

In the YOLOv7-tiny algorithm, the CIoU loss function was originally used. Traditional target detection loss functions (such as CIoU, GIoU[12], DIoU[13], etc.) mainly rely on the aggregation of bounding box regression indicators, such as the distance between prediction box and real box, overlapping region, and horizontal and vertical ratio. However, these methods do not take into account the orientation mismatch between the required prediction box and the real box, have some ambiguity, and ignore the balance between difficult samples, resulting in slow convergence and poor effect of the model. To solve the above problem, we propose a new loss function, namely SIoU (SCYLLA-IoU). SIoU builds on DIoU, considering the vector angle between the required regression, redefines the penalty metric and introduces directionality to make the network more suitable to the predictive box regression of the target. The SIoU loss function consists of four parts: angle loss, distance loss, shape loss, and IoU loss. The role of angular loss is to minimize the number of distance-related variables, and to bring the prediction box to the X or Y axis (whichever is closest), and then continue to approach along the correlation axis. A schematic representation of the angular loss calculation is shown in Figure 2.

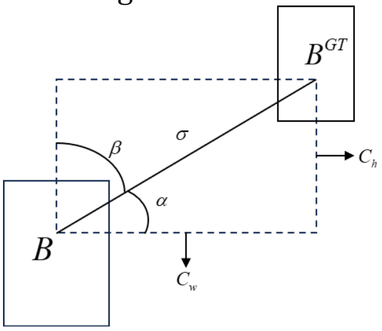


Figure 2. Schematic diagram of the angle loss

2.3. Introduce the ACmix attention mechanism

The attention mechanism is a key idea in deep learning that mimics the human visual system, inspired by the human brain's selective attention to information. Adding the attention mechanism to the image classification can help the model to pay more attention to the category-related areas in the image and improve the classification accuracy. In order to fully realize the different advantages of convolutional neural network and self-attention module, this paper proposes an ACmix algorithm module integrating two modular paradigms. This method skillfully combines the advantages of CNN and self-attention module, avoiding two complex projection operations and increasing the attention to small and occluded targets[14]. The ACmix module divides the network structure into two phases, as shown in Figure 3.

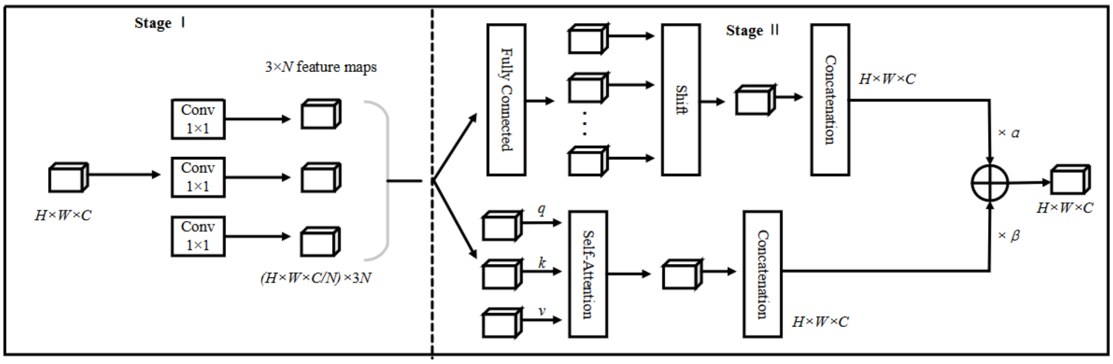


Figure 3. ACmix network structure

In the first stage, the input feature graph is projected by convolution through three convolutions to generate an intermediate feature set. In the second stage, it is divided into two branches: the upper branch is the convolution path with kernel size K. The N group of feature maps are generated through the full connection layer, and then translated first, and finally the sum of feature maps of different kernel positions is calculated through the aggregation operation. The lower branch is the self-attention path. After the multi-head self-attention model, the attention weight is calculated and the value matrix is multiplied, and finally the features are obtained through the displacement and aggregation processing. Eventually, the outputs of the upper and lower branches are summed and two coefficients are added to control the numerical size of the final output.

2.4. Adaptive feature fusion network

In order to solve the original algorithm each detection branch of each level to detect the target object using the way of direct splicing or add connection, resulting in the pyramid feature scale inconsistent problem, this paper puts forward a data-driven pyramid feature fusion strategy, called adaptive spatial feature fusion[15]. This method enables the network to learn autonomously and spatially filter all kinds of features at each levels, and is not related to the backbone model, which is very suitable for the feature pyramid structure of single detector. Its network structure is shown in Figure 4. The ASFF structure adopts the differential search optimal fusion operation, which is easy to learn the effective information of each level, and the implementation is simple to achieve and the calculation cost is not high.

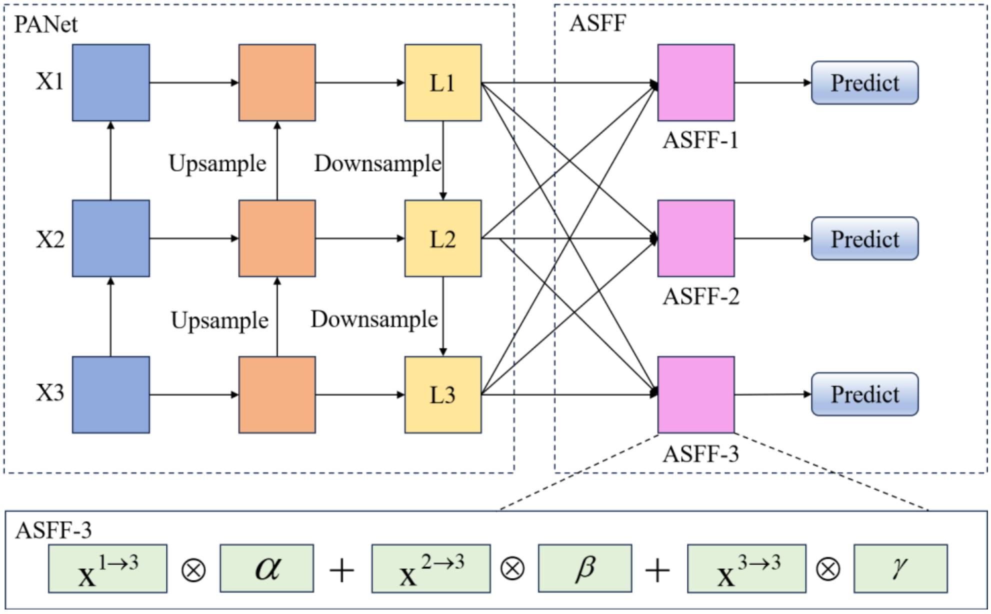


Figure 4. ASFF network structure

The first first step is feature scaling. Input X1, X2 and X3 as the last three feature layers of the backbone network, which first generate layers L1, L2 and L3, and then keep the number of channels consistent by the downsampling and pooling operations. The second step is the adaptive feature fusion. Take the fusion mode of ASFF-3 in the solid wire frame in Figure 4 as an example. Features from L1, L2 and L3 layers are multiplied by the corresponding weight value of each layer, and finally added up to obtain the new fusion features.

2.5. Improved network structure of the algorithm

In order to improve the detection performance of the YOLOv7-tiny algorithm model, this paper proposes the YOLOv7-tiny-SSAA target-detection algorithm. The improved network structure is shown in Figure Figure 5. In this paper, the CBS of the improved network structure replaces the activation function Leaky ReLU with the SiLU activation function, which helps to conduct the gradient descent optimization more stably and suppress the overfitting phenomenon. At the same time, the original YOLOv7-tiny uses CIoU as the boundary box regression loss function to accelerate the convergence of the model with the SIoU loss function; secondly, the ACmix module and ASFF module are applied to the head layer, which is more conducive to detect effective information of the feature maps at each levels, and can more effectively perform feature fusion.

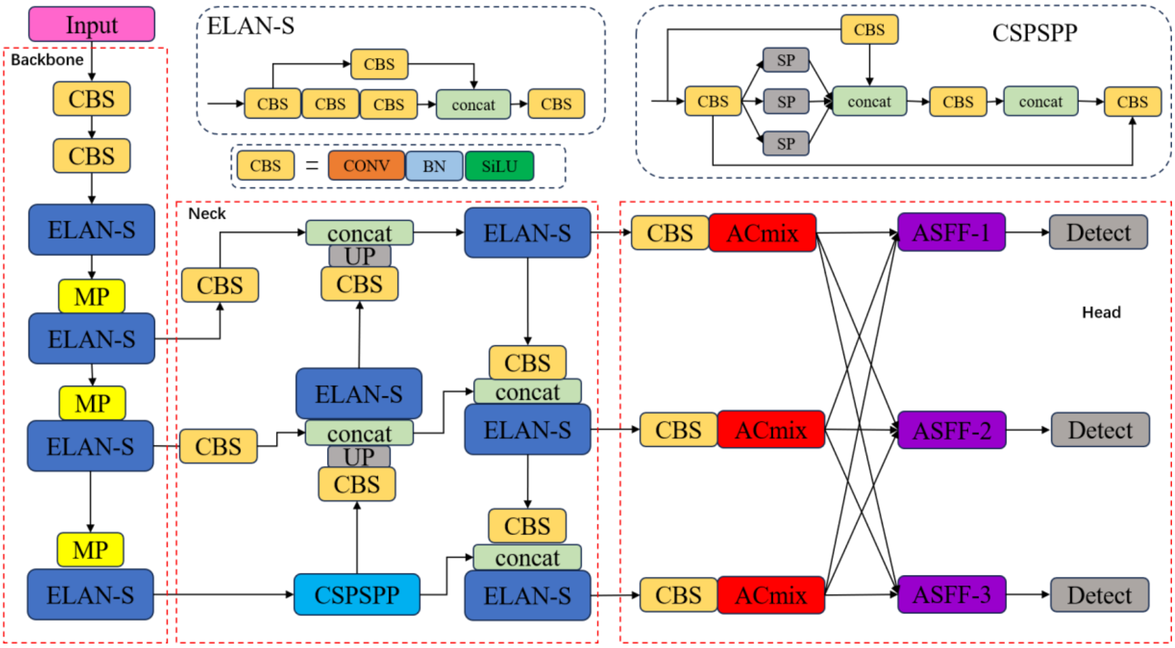


Figure 5. Structural diagram of the improved YOLOv7-tiny network

3. Experimental procedure and results analysis

3.1. Introduction to datasets

The bearing surface defects of this experimental study were collected from hub bearings on small vehicles. It mainly analyzes the damage of the bearing outer ring, and the defects are divided into three categories: scratches, scratches and grooves. The data set to 5800 images was expanded through image enhancement. After screening and sorting, it can be roughly divided into four categories: single defect target, hidden defect target, shallow defect target and overlapping and blocking small defect target. Some images are shown in Figure 6.

3.2. Model training process and results

The image input size was set to 640×640 and Batch size was set to 16, and Adam optimizer was used, which can quickly converge the network model parameters and speed up the model training. The initial learning rate of the model was set to 0.01, weight-day was set to 0.0005, warmup-epochs was set to 3, the learning rate was reduced, the training rounds were set to 200, and the network was trained until it converged.

In order to better verify the influence of each module on the model detection effect, the proposed improved method module was ablation, and the ACmix module and ASFF module were applied to the YOLOv7-tiny network one by one to verify the performance of the improved algorithm. The results of the ablation experiments are shown in Table 1.

Table 1. The YOLOv7-tiny ablation experiments

Test	ACmix	ASFF	SiLU	SIoU	P/%	R/%	mAP_0.5/%	Params/MB	FPS/f/s
1					83.1	80.2	86.4	6.2	75
2			√	√	82.4	81.8	87	10.7	67
3	√		√	√	84	81.8	87.5	11.7	66
4		√	√	√	85.1	83.4	89.6	13.4	64
5	√	√	√	√	85.9	85.4	91.6	14.4	63

Test 1 in Table 1 is the benchmark model YOLOv7-tiny, and Test 2 replaced the original model activation function Leaky ReLU with SiLU and the loss function with CIoU with SIoU. The loss function during training is shown in Figure 7.

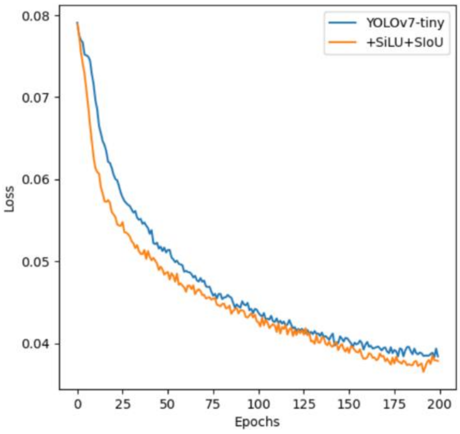


Figure 7. Training loss curve comparison plot

As can be seen from Figure 7, in the first 75 cycles, the improved model decreased faster than the original model Loss value, and the convergence effect was better. The original model and the improved model loss value finally converged to about 0.04 and 0.38, respectively, which proves that the proposed improvement strategy of replacement activation and loss function is feasible and effective. In the following Test 3-5, the improved modules were gradually added on the basis of Test 2. According to Table 1, each module improves the identification accuracy. The improved model combined with the four types of modules showed nearly 5.2% improvement in the average mean accuracy compared with the original model. The PR curves of both are shown in Figure 8. By comparison, the improved model increased the groove small target defects by 5.8%, the scratch target by 2.3%, and the scratch superficial target by 5.1%. Although the processing speed of the improved model decreased by 10f/s, it can effectively reduce the rate and misdetection, which is better than the original model.

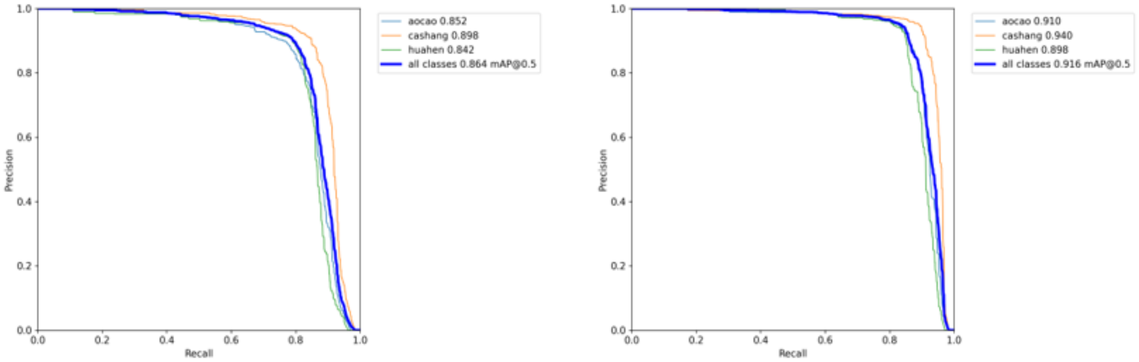
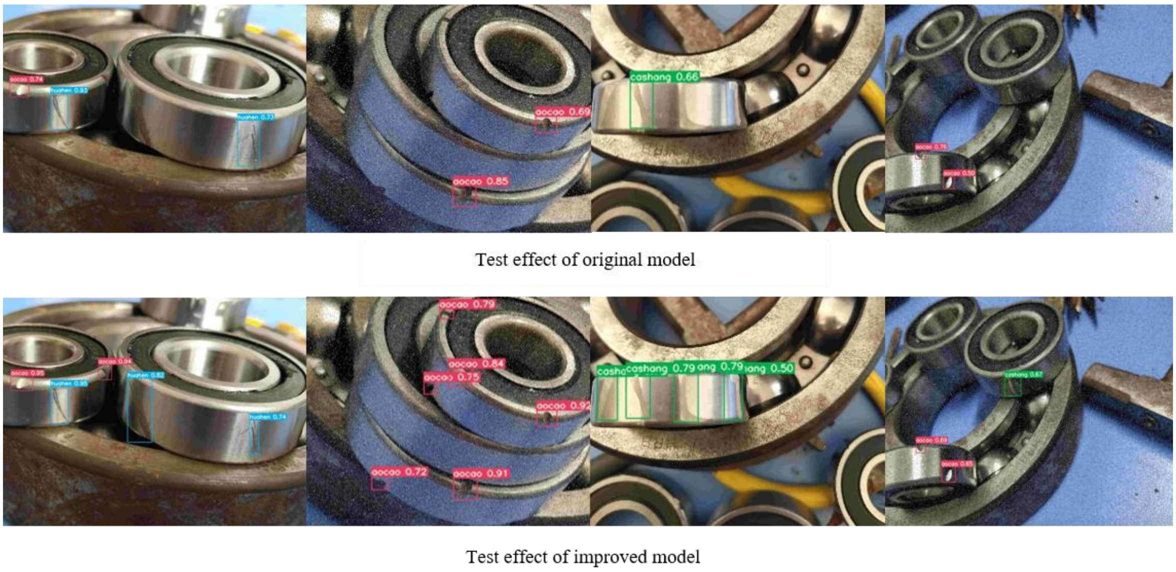


Figure 8. PR curve comparison

3.3. Detection effect display and analysis

In order to further verify the detection effect of the YOLOv7-tiny algorithm before and after the improvement in the detection of bearing surface defects, four representative images with shallow, hidden and overlapping occlusion targets were taken for detection, and the results are shown in Figure 9. Red boxes represent groove defects, green and blue represent abrasion and scratch defects.



Test effect of original model

Test effect of improved model

Figure 9. Results of detection before and after model improvement

Four groups can be observed from figure 9, the proposed improved ACmix module is more sensitive to hidden small target features, introducing the adaptive space fusion feature ASFF can strengthen the network features of global attention, makes the improved model robust and anti-interference improved, comparing the algorithm model successfully detected the original model of small target defects, in hidden, stacked and shallow complex scenarios for all kinds of defects have better recognition rate and robustness.

4. Conclusion

With the development of intelligent manufacturing technology, the automatic and intelligent surface defect detection system has become an important part of the production process of modern manufacturing enterprises. With the increasingly fierce competition in the automobile market, consumers' requirements for the overall quality and safety of automobiles are increasing day by day. The detection of surface defects of automobile parts is not only related to product quality and manufacturing efficiency, but also directly affects the competitiveness of the automobile industry and the satisfaction of consumers. At present, deep learning is gradually applied to defect detection tasks. This paper selects a stage YOLO target detection method, put forward a variety of optimization strategy for YOLOv7-tiny algorithm, and verified in the relevant data set, the results show that the performance of the algorithm has significantly improved, ensure the accuracy of hidden, overlapping targets, avoid missing detection phenomenon, and realize the good balance of identification accuracy and detection speed, can well complete the detection of surface defects. In the subsequent research, the algorithm can be deployed and tested in the actual production line.

References

[1] Guo Qichang. Brief analysis of the perfect quality management system of auto parts suppliers [J]. Times Automobile, 2019, (02): 163-164.

[2] Zhai Guoyong. Optimization and research of Quality management of Automotive Parts supply Chain [J]. Internal combustion engine and accessories, 2019, (16): 224-225.

[3] Li Honglei, Ouyang Zhihua. Application and development trend of nondestructive testing technology of auto parts [J]. Equipment Engineering of China, 2020, (06): 123-124.

[4] Luo Huilan, Chen Hongkun. Review of deep learning-based object detection studies [J]. Journal of Electronics, 2020,48 (06): 1230-1239.

[5] Zhou Feiyan, Jin Linpeng, Dong Jun. Review of Convolutional neural network research [J]. Journal of Computer Science, 2017,40 (06): 1229-1251.

[6] Yanan S, Hui Z, Li L, et al. Rail surface defect detection method based on YOLOv3 deep learning networks[C]//2018 chinese automation congress (CAC). IEEE, 2018: 1563-1568.

[7] Li X, Wang C, Ju H, et al. Surface defect detection model for aero-engine components based on improved YOLOv5[J]. Applied Sciences, 2022, 12(14): 7235.

[8] Wang K, Teng Z, Zou T. Metal defect detection based on yolov5[J]. Journal of Physics: Conference Series, 2022, 2218(1):012050.

[9] Kumar A, Singh S. Classification of data on stacked autoencoder using modified sigmoid activation function[J]. Journal of Intelligent & Fuzzy Systems, 2023, 44(1): 1-18.

[10] Zhang Q Q, Wang S Y, Lin D Y, et al. Robust Affine Projection Tanh Algorithm and Its Performance Analysis[J]. Signal Processing, 2023, 202: 108749.

[11] Gerlind P, Yannick R, Yurii K. Spline representation and redundancies of one-dimensional ReLU neural network models[J]. Analysis and Applications, 2023, 21(01).

[12] Liu X, Hu J, Wang H, et al. Gaussian-IoU loss: Better learning for bounding box regression on PCB component detection[J]. Expert Systems with Applications, 2022, 190: 116178.

[13] Zheng Z, Wang P, Liu W, et al. Distance-IoU loss: Faster and better learning for bounding box regression[C]//Proceedings of the AAAI conference on artificial intelligence. 2020, 34(07): 12993-13000.

[14] Pan X, Ge C, Lu R, et al. On the integration of self-attention and convolution[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022: 815-825.

[15] Liu S T, Huang D, Wang Y H. Learning Spatial Fusion for Single-Shot Object Detection [EB/OL]. arXiv: 1911.09516, 2019.