An Improved SSD Model for Small Size Work-pieces Recognition in Automatic Production Line

Xiaoning Bo[1,2], *Zhiyuan Zhang*[3,4*], *Yipeng Wang*[5]

[1] *School of Electronic Engineering, Taiyuan Institute of Technology, China*
[2] *School of Information and Communication Engineering, North University of China, China*
[3] *Department of Electronic and Information Engineering, Beijing Jiaotong University, China*
[4] *Key Laboratory of Communication and Information Systems, Beijing Municipal Commission of Education, China*
[5] *Department of Network and Information Security, China Railway Information Technology Group Co. Ltd., China*
boxn@tit.edu.cn, zhangzhiyuan@bjtu.edu.cn, wangyipeng@sinorail.com

Abstract

Aiming at the problems of slow recognition speed and low recognition accuracy of arbitrarily placed workpiece by machine vision in traditional automated production lines, a workpiece recognition algorithm based on improved SSD is proposed. Firstly, the improved DarkNet53 is used to replace the backbone network in the original SSD network framework, and the network enhancement is used in the backbone network to solve the defect of small target missed detection. Then, channel attention module and deep semantic feature fusion module are added, in order to improve the recognition ability and detection accuracy of the small target features. Lastly, the loss function was optimized, and the problem caused by sample imbalance was solved by changing the weight distribution of positive and negative samples. In the experiment, image datasets of typical bolts, nuts, and connecting plates were constructed for the network training, the experimental results showed that, the recognition accuracy and speed have been optimized and meet the requirements of automatic work-piece detection in actual production, compared with traditional YOLOv4 and the original SSD algorithm in the work-piece recognition task.

Keywords: Deep learning, Automatic production line, Work-piece recognition, SSD, Feature fusion

1 Introduction

With the rapid development of intelligent manufacturing, automated production lines are widely used in the manufacturing industry. The application of computer vision technology in automated production lines can improve the accuracy of work-piece recognition, assembly, and defect detection, however, there are various types of work-pieces and their positions are diverse in real production, which requires higher accuracy and efficiency in the recognition. Therefore, how to use the computer vision to realize real-time, efficient, and high-precision recognition for work-piece types and positions in production, has become a research focus in intelligent manufacturing.

Traditional target recognition algorithms are mainly based on target feature extraction and feature matching, but the generalization ability of the algorithms are insufficient because they rely on the experience of designers heavily [1-2]. Recently, with the development of deep learning, the object detection algorithms based on deep learning are increasingly proposed, and they are mainly divided into two categories: (1) Two-stage object detection algorithms, such as R-CNN [3], Fast R-CNN [4], Faster R-CNN [5], etc., (2) One-stage object detection algorithms, such as YOLO [6], SSD (Single-Shot Multi-Box Detector) [7], etc. Different with the Two-stage object detection algorithms, which selecting the target candidate area firstly and then extracting feature in the deep network in target recognition, the One-stage algorithms consider the recognition as a regression problem, and reduce the region candidate boxes process which can improve the recognition speed. Therefore, the type of algorithm is more suitable for small work-pieces detection which requires the high real-time and efficiency in the recognition, but it also suffers from the low detection accuracy and missed detection issues for small target work-pieces [8]. It is necessary to improve the One-stage algorithm model according to the basic idea of the One-stage algorithm, to achieve the efficient and high-precision recognition of small target work-pieces while meeting real-time requirements.

At present, there is relatively little research on small work-piece recognition algorithms, and there is also a lack of the relevant literature. Khalid et al. [9] designed a new fully convolutional neural network structure and the recognition speed has been improved, in which the point cloud information of images has been divided, and the number and pose of work-pieces has been identified by checking whether the work-piece is in the work-piece area and non work-piece area. Bay et al. proposed a work-piece feature extraction algorithm, and the feature detection operators based on Hessian Matrix is used which can improved the image recognition speed, but the recognition speed cannot fully meet the recognition needs of real production lines, and the recognition accuracy is not high enough [10]. Schwinger proposed a linearized SURF feature detection classifier, in which the difference between feature points

*Corresponding Author: Zhiyuan Zhang; E-mail: zhangzhiyuan@bjtu.edu.cn

DOI: 10.53106/160792642024032502004

has been improved while ensuring the scale invariance, and the proposed classifier can improve the recognition speed compared with the original algorithm, but it reduces the image matching area [11]. Jian Xu et al. proposed an improved SSD algorithm for identifying dense work-pieces, the potential overfitting and optimization problems are improved in this algorithm, and it has reference significance for small work-piece recognition [12]. Some research applied attention mechanism in the recognition of work-pieces surface defects, the proposed algorithm improved the recognition accuracy by using image super-resolution reconstruction methods and multiple image feature extraction algorithms [13]. Reference [14] specialized in researching how to use the current small-scale datasets to achieve work-piece recognition, and they aim to achieve optimal recognition accuracy and efficiency.

In this paper, A new method is proposed based on the SSD target recognition algorithms, in order to meet the requirements of strong real-time and high accuracy for small and medium-sized work-pieces recognition in actual automated production. In the model, DarkNet53 [15] replace the original VGG-16 as the backbone network to reduce the risk of semantic feature loss in the original model. At the same time, the problem of imbalanced positive and negative samples during the training process is improved by redistributing the weight values of positive and negative samples based on the training strategy of SSD.

2 The Principle of SSD

The principle of SSD algorithm is using the convolutional neural networks to identify firstly, and then different predictive feature maps are generated. The front layer feature map is used to identify small objects due to its large amount of retained feature information, the back layer feature map is used for big objects recognition, the category and position information of the information in the image corresponding to pre-selected boxes with different sizes are set at different feature layers, and finally the prediction result is obtained through non-maximum suppression.

2.1 The SSD Model Architecture

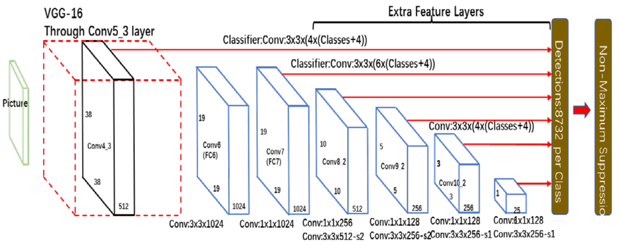


Figure 1. The SSD model architecture

In the SSD algorithm, the FC6 and FC7 layers of the backbone network VGG16 are changed to Conv6 and Conv7, and four additional layers are added: Conv_8. Conv_9. Conv_10. Conv_11. Then, the different feature maps are generated which the size are 10×10, 5×5, 3×3, 1×1, and these feature maps will be input into the subsequent classification and regression steps.

2.2 The SSD Algorithm

The training strategy of SSD is that the network model performs regression training on classification and position offset. There is information loss during the training process, which is measured using the loss function. The loss function can be divided into the position loss function and confidence loss function, and the overall loss function can be expressed as:

$$L(x,c,l,g) = \frac{1}{N}(L_{conf}(x,c)) + \alpha L_{loc}(x,l,g)\,. \tag{1}$$

$$L_{conf}(x,c) = -\sum_{i\ \ Pos} x_{ij}^{p}\log(\hat{c}_j^{p}) - \sum_{i\in Neg}\log(\hat{c}_i^{0})\,. \tag{2}$$

$$w\,L_{loc}(x,l,g) = \sum_{i\in Pos}^{N}\ \sum_{m\in(cx,cy,w,h)} x_{ij}^{k}\,smooth_{L1}(l_i^m - \hat{g}_j^m)\,. \tag{3}$$

$$smooth_{L1}(x) = \begin{cases} 0.5x^2 & if\,|x|<1 \\ |x|-0.5 & other \end{cases}. \tag{4}$$

All parameters in above formulas are shown in Table 1:

Table 1. Parameter description

Parameter	Explanation
N	Number of positive samples in the prior box
L_{loc}	The position loss function
L_{conf}	Confidence loss function
Pos	Number of sample prediction boxes
Neg	Negative sample prediction box
x_{ij}^{k}	Indicates that the *i* prediction box matches the *j* real box with respect to category *k*
l_i^m	Prediction box
$\hat{g}_j^m$	Realistic Box
x_{ij}^{p}	The *i* prior_ Box matches *j* gt_box, if the match is successful, the value is 1, otherwise it is 0
$\hat{c}_j^{p}$	The probability that the *j* sample is class *p*
x	Whether the default Box matches the real box successfully, $x\in(0,1)$
c	Classification confidence
l	Prediction box
g	Real label box
α	Balancing Confidence Loss Parameters and Position Loss Parameters
w	Default Box Width
h	Default Box Width
cx, cy	Default box center coordinates
$smooth_{L1}(l_i^m - g_i^m)$	Smooth $L1$ norm
$\hat{c}_i^{0}$	The *i* sample has a negative sample probability

The SSD algorithm matching strategy is to first determine the overlap rate IoU between the default boundary recognition box and the actual boundary box. If the overlap rate is greater than the default threshold, the two are considered to match.

If the overlap rate is less than the default threshold, it is considered that the two do not match.

3 The Design and Optimization of Improved SSD Model

The overall structure of the improved algorithm proposed in this paper is shown in Figure 2. The Channel Attention Module (CAM) and Deep Semantic Feature Fusion Module (DSF) are added into the SSD basic network. CAM module achieves attention intensity guidance for specific information by analyzing, calculating, quantifying the correlation of various network channels, and reweighting the importance of information. Therefore, the introduction of CAM can increase the proportion of important information in the channel and improve the recognition effect for small size parts by setting the small parts information. The DSF module can solve the problem that poor recognition performance for small target work-piece due to the insufficient shallow semantic information. From the framework diagram, it can be seen that: local and full information are fused after DSF extracts three scale features in parallel in order to ensure richer semantic information. In addition, the model overall structure also includes an improved backbone Darknet53 enhanced network, forward paths and reverse paths. The forward path follows the network structure from shallow features to deep features and the reverse path is opposite. The forward path can enhance the richness of detailed information extracted from the deep feature maps, while the reverse path can enhance the local semantic information of the shallow feature maps. In short, the advantages of the improved model proposed in this paper will mainly be reflected in the recognition accuracy and identity speed for small work-pieces.

Specifically, the three different scales feature maps are extracted firstly by the improved Darknet53 backbone enhancement network composed of F_3, F_4, F_5, and the deep feature maps containing rich detail information are extracted by the forward paths. Then, apply CAM to enhance the attention intensity (I_3, I_4, I_5) of the small target information in the three feature layers, in order to reduce the impact of image background. At the same time, I_6 is obtained through the DSF module by using the local feature P_7 and global feature P_8 of the deep feature map. Finally, perform multi-scale prediction on I_3, I_4, I_5, I_6, P_7, P_8, and obtain the final recognition results.

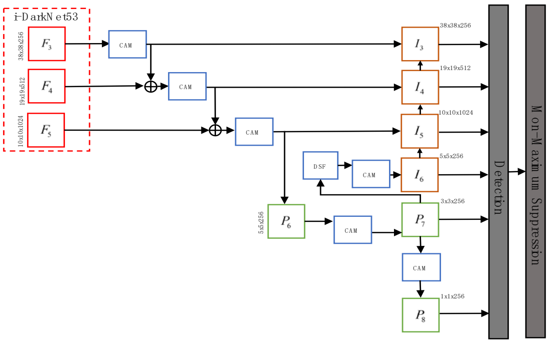
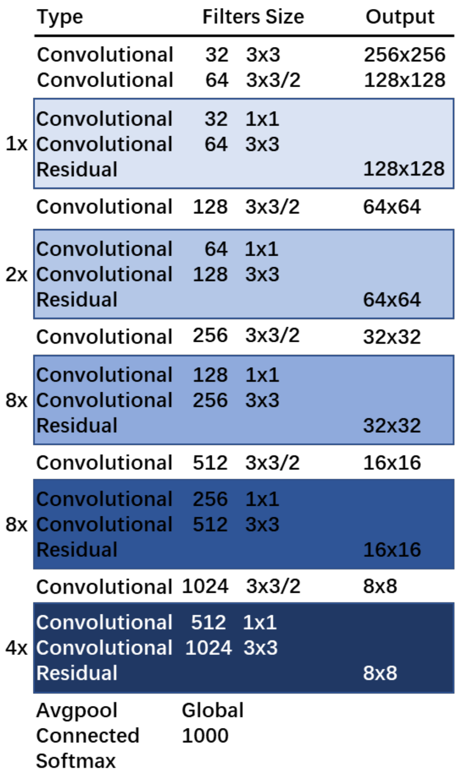


Figure 2. Improved SSD model architecture

3.1 Backbone Enhanced Network

The SSD algorithm adopts VGG-16 as the basic backbone network, but there are shortcomings such as the insufficient feature extraction ability and the missed detection for small target detection. Therefore, Darknet53 is used as the basic backbone network in this paper, which contains a large number of residual structures and can effectively enhance the depth of the network and provide higher level support for feature semantic information extraction. The network structure is shown in Figure 3. During backpropagation, residual structures can also transfer gradients to the shallower networks in the front-end, and weaken the chain reaction of reverse differentiation, and avoided the gradient disappearance and explosion. In addition, reduce the input channel of the residual structural unit by half using the 1×1 convolutional layer, and then preform 3×3 convolutional operation. These improved process can reduce the computational load and make the network run faster. Meanwhile, the down sampling process of the Darknet53 network is achieved through a convolution process with a side of 2, and the model involves 5 down sampling cycles, which effectively avoiding the problem of semantic loss in the pooling layer.



	Type	Filters	Size	Output
	Convolutional	32	3x3	256x256
	Convolutional	64	3x3/2	128x128
1x	Convolutional	32	1x1	
	Convolutional	64	3x3	
	Residual			128x128
	Convolutional	128	3x3/2	64x64
2x	Convolutional	64	1x1	
	Convolutional	128	3x3	
	Residual			64x64
	Convolutional	256	3x3/2	32x32
8x	Convolutional	128	1x1	
	Convolutional	256	3x3	
	Residual			32x32
	Convolutional	512	3x3/2	16x16
8x	Convolutional	256	1x1	
	Convolutional	512	3x3	
	Residual			16x16
	Convolutional	1024	3x3/2	8x8
4x	Convolutional	512	1x1	
	Convolutional	1024	3x3	
	Residual			8x8
	Avgpool		Global	
	Connected		1000	
	Softmax			

Figure 3. The architecture of Darknet53 network

In this paper, we designed an improved Darknet53 network as the backbone enhancement network. The network structure is dividing the convolutional layers of the original Darknet53 residual structure into t channel combinations, which are marked as c_1, c_2, …, c_t, based on the original convolution. Each channel in a single group has the same size, and the number of channels is t-tenth of the input feature maps. The improved structure is shown in Figure 4, the calculation for channel groups is as follows:

$$d_i = \begin{cases} c_i & i=1 \\ Conv_{3\times3}(c_i + d_{i-1}) & 1<i<t \end{cases}. \quad \textbf{(5)}$$

Where $Conv_{3\times3}$ is the convolutional kernel, t is the number of channel combinations, and the t value can be set according to the actual situation.

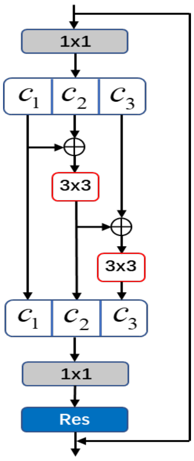


Figure 4. The convolution operation based on the improved residual structure

Compared to the original convolutional kernel, the improved residual structure expands the range of receptive fields and expands the network width of i-Darknet53, which can extract more global information. In addition, feature information from different feature channels can be interwoven and transmitted by establishing the connections through different feature channels in a unified convolutional layer, thus, the feature extraction ability of the basic network i-Darknet53 is improved, and small targets are identified through fine-grained features. In summary, the structure of this part is named the backbone enhanced network.

3.2 Design of the Channel Attention Module

The attention mechanism in deep learning draws on the selective attention mechanism of the human visual system, and the core goal is to select information that is more critical to the current task objective from numerous sources, and extract important features of sparse data quickly [16]. Jianfei Zhang et al. utilized the multi-head self attention mechanism to focus on the important information in different positions and representation subspaces of the input data, and learn the signal global features. The model proposed by Jianfei Zhang can reduce the complexity of the recognition model, make the training easily, and have higher damage identification accuracy and stronger noise resistance, as well as better identification ability for damage modes with similar damage characteristics [17]. Reference [18] formed a mixed attention mechanism module in both feature channels and space to suppress meaningless features, enhance meaningful features, and improve the segmentation accuracy of small-scale targets and target boundaries. Inspired by these ideas, for the problem of small targets being susceptible to interference from background information due to the unclear feature information, we add a channel attention module into the network architecture to enhance the efficiency of extracting effective features from small targets.

The channel attention module CAM (Channel Attention Module) designed in this paper, the input feature are sequentially processed through convolutional layers, global average pooling layers, and activation function operations. Then the obtained feature map is multiplied with the input features to output new features, and identify the presence of detection targets by assigning feature weights to different channels. The structure diagram of the channel attention module is shown in Figure 5, and the calculation formula is as follows:

$$F_C = \sigma(\mathrm{Re}lu(AvgPool(Conv(F)))) \otimes F. \quad \textbf{(6)}$$

Where σ represents the activation function, $\otimes$ represents multiplication.

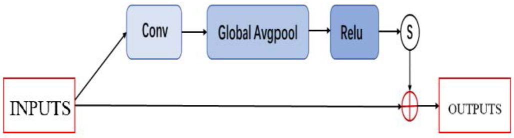


Figure 5. The attention module

3.3 Feature Fusion Module

The insufficient fusion of feature semantic information in the feature fusion module can lead to poor performance in small target object recognition. In order to address this issue, a deep semantic feature fusion module is used in this paper, which fuse and represent the local and global information of the three scale features extracted in parallel fully (seen in Figure 1). The specific structure of the deep feature fusion is shown in Figure 6, which $Conv_{3\times3}$ is used for extracting the local information and $DeConv_{3\times3}$ is used for enhancing the local information. P_8 represents an abstract global feature, it's size can be changed into the size of P6 by the Broadcast and lastly I_6 is obtained by fusing all feature information. Therefore, I_6 contains both local and global information, this can make the semantic information is more abundant, the semantic information of shallow feature maps can be enhanced in the reverse fusion, and thus improve the confidence and classification accuracy.

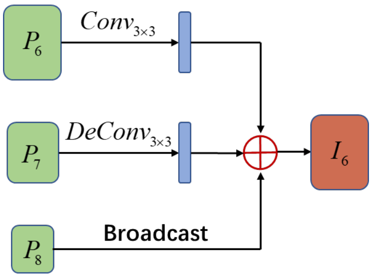


Figure 6. Feature fusion module

3.4 The Optimization of Loss Function

From the experience of the actual recognition, if there are a large number of negative samples in the background and a small number of positive samples in the target, it is an imbalanced distribution for the samples. At the same time, if the difficult samples between the small targets and background are difficult to classify, it can also lead to the problem of imbalanced positive and negative samples in the model [19]. If the loss of different samples is not considered, it may result in a number of the small target positive samples can not play a dominant role in the loss function, which will lead to the deviation during the training process, and reduce the small target recognition accuracy. In response to the aforementioned shortcomings, the loss function has been improved by optimizing the cross entropy loss, which is represented as follows:

$$CE(p,y)=\begin{cases}-\log(p) & if.y=1\\ -\log(1-p) & other\end{cases}. \tag{7}$$

Where $p\in[0,1]$ indicates the probability that the predicted sample belongs to a positive sample, $y = \{1, -1\}$ represents a sample category label, p_m is the tag probability, which can be represented as follows:

$$p_m=\begin{cases}p & if,y=1\\ 1-p & other\end{cases}. \tag{8}$$

Therefore, the cross entropy function is simplified as:

$$CE(p_t)=-\log(p_t). \tag{9}$$

In order to reduce the problem of large differences between the positive and negative sample, corresponding weights are set based on the contribution of the different sample for the loss, and shared weight value are introduced. When the shared weight value is small, the shared weight is denoted as β_t, and the improved loss function is:

$$CE(p_t)=-\beta_t\log(p_t). \tag{10}$$

For the samples that are difficult to classify, the dynamic balance adjustment coefficients $(1-p_t)^{\varepsilon}$ is used which can improve the loss proportion of the difficult classify samples. The final loss function calculation is expressed as:

$$FL(p_t)=-\beta_t(1-p_t)^{\varepsilon}\log(p_t). \tag{11}$$

Where $\varepsilon \geq 0$ represents the focusing parameter. In the real training, when $\beta = 0.3$, $\varepsilon = 0.25$, the recognition effect is best.

Table 2. Parameter settings

Parameter	Value
Iterations	300
Batch size	32
Learning rate	0.004
Momentum	0.9
Weight decay	0.0005

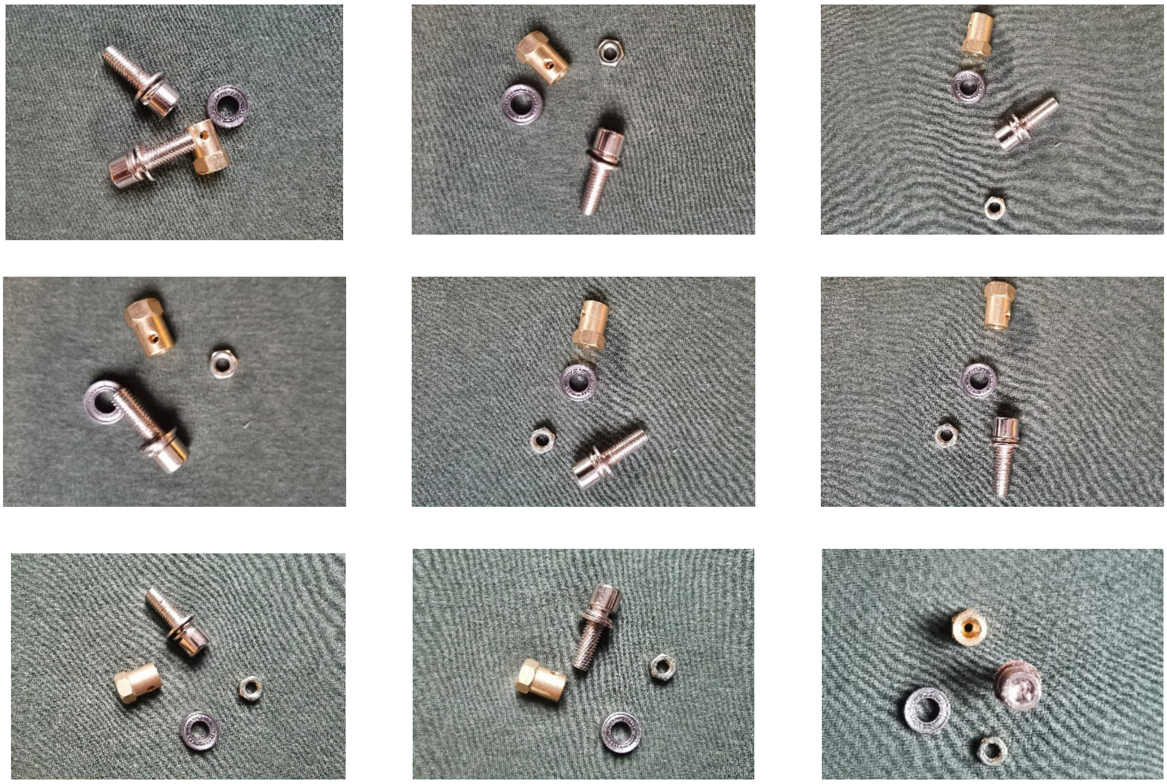


Figure 7. The three types of work-pieces to be identified

4 Experiment

4.1 Experimental Environment

In the experiment, the operating system is Ubuntu 18.04 LTS, CPU is Intel's Core i5-8500 with 32GB memory, GPU is Nvidia RTX 2080Ti with 22G memory, the industrial camera is MV-VA060-10GC, and the deep learning framework versions are Pytorch1.2 and Python3.6 respectively. In the model training, update and optimize the weights of the network model using the Stochastic Gradient Descent (SGD) algorithm [20]. Select three types of commonly used work-pieces on the production line as samples, namely hexagonal nuts, hexagonal bolts, and square connecting pieces, and the experimental images are shown in Figure 7. Annotate images by using Labeling software, and then expand the data through transformation methods such as flipping, mirroring and so on. The experiment data with 24000 pieces are divided into three groups: training set with 18000 pieces, validation set with 2000 pieces, testing set with 4000 pieces, and the parameter settings are shown in Table 2.

4.2 Training Results Analysis

During the training process, as the number of iterations continues to increase, the loss values tend to flatten out without the significant oscillations. The loss value curve can be seen in Figure 8, and the trend of the curve indicates that the convergence and training effect of the algorithm in this paper are good.

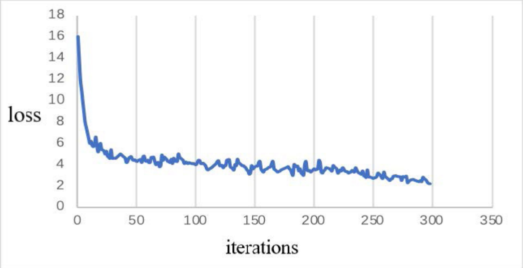
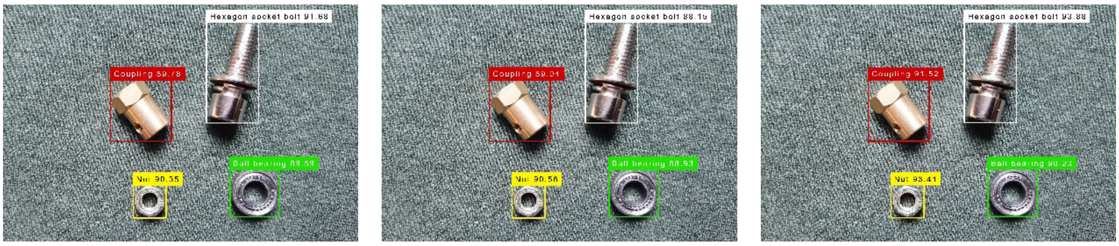


Figure 8. Model training and the results

4.3 Recognition Results Analysis

Two indicators of target recognition are used to evaluate the identification rate and accuracy of the model in this paper: (1) Select the frames per second (fps) to represent the model recognition speed to measure the number of images that can be identified per second. (2) Adopt the mean accuracy mAP to represent the recognition accuracy, which is the average value of recognition accuracy AP for all categories. The definition of AP is as follows:

$$AP = \int_0^1 p(r)dr\,. \tag{12}$$



(a) YOLOv4 (b) SSD (c) Our model

Figure 9. The recognition results

Where $p(r)$ is a curve with recall rate as the horizontal axis and accuracy as the vertical axis.

In order to analyze the recognition effect of the improved algorithm in this paper intuitively, the image size recognized by the industrial camera is 300×300, and different algorithms such as YOLOv4 and SSD are used to recognize the same image, the recognition results are shown in Figure 9.

In order to verify the advantages of the proposed algorithm for small work-pieces recognition in automated production lines, the algorithm described in this paper compared with other algorithms, and the experimental results are shown in Figure 10.

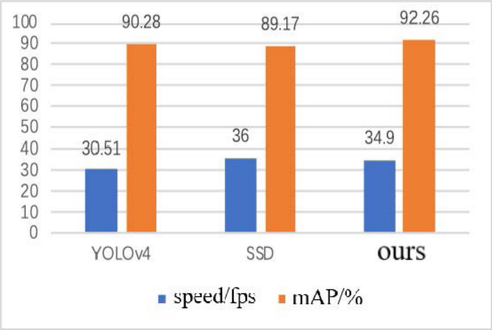


Figure 10. Algorithm comparison

In terms of the recognition accuracy, it can be seen that the algorithm proposed in this paper has significantly improved in the recognition speed and accuracy compared to other typical one-stage recognition algorithms. The recognition accuracy can reach 92.26%, which is 3.09% higher than the accuracy of the basic SSD algorithm, and 1.98% higher than YOLOv4 algorithm. The results indicate the improved SSD algorithm can meet the requirements of the production line. In terms of the recognition speed, due to the addition of the attention mechanisms and the reverse paths, the recognition speed of this paper's algorithm is lower than SSD algorithm, but it is significantly higher than YOLOv4 algorithm, and the detection speed can still meet the requirements of the production line. In a word, the improved algorithm which integrates attention mechanism, semantic information, positive and negative paths, makes the model structure become more complex and reduces the recognition speed of the small work-pieces, but it greatly improves the recognition accuracy to some extent.

5 Conclusion

In this paper, the One-stage SSD algorithm is improved. The purpose is to solve the low recognition accuracy problem caused by the lack of semantic information in the shallow feature maps of the original algorithm. Specifically, the first improvement is to add the channel attention machine mechanism into the original algorithm to increase the attention for the detail features, the second improvement is to eliminates the interference of background information using the semantic feature fusion module. The experimental result show that the improved algorithm has achieved good results in terms of recognition accuracy and speed for typical small work-pieces in automated production lines, such as hexagonal bolts, hexagonal nuts, and small connecting plates. In the future, more lightweight network model will be further studied to improve the recognition speed.

Acknowledgement

This work is supported by Beijing Nova Program from Beijing Municipal Science & Technology Commission (grant number Z211100002121120); Key Research and Development Program of Shan Xi (grant number 201803D121069); Scientific and Technological Innovation Programs of Higher Education Institutions in Shanxi (2022L534); Teaching Reform and Innovation Programs of Higher Education Institutions in Shanxi (J20221108).

References

[1] L. Sun, X. Zhang, W. Qin, Research on target recognition and tracking in mobile augmented reality assisted maintenance, *Computer Animation and Virtual Worlds*, Vol. 33, No. 6, pp. 1-14, November/December, 2022.

[2] H. J. Gu, S. Chen, A Tire Defect Detection System Based on Deep Learning, *Computer & Digital Engineering*, Vol. 50, No. 7, pp. 1463-1467, July, 2022.

[3] R. Girshick, J. Donahue, T. Darrell, J. Malik, Rich feature hierarchies for accurate object detection and semantic segmentation, *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR 2014)*, Columbus, OH, United States, 2014, pp. 580-587.

[4] R. Girshick, Fast R-CNN, *Proceedings of the IEEE international conference on computer vision (ICCV)*, Santiago, Chile, 2015, pp. 1440-1448.

[5] S. Ren, K. He, R. Girshick, J. Sun, Faster r-cnn: Towards real-time object detection with region proposal networks, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 39, No. 6, pp. 1137-1149, June, 2017.

[6] J. Redmon, S. Divvala, R. Girshick, A. Farhadi, You only look once: Unified, real-time object detection, *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, Las Vegas, NV, United States, 2016, pp. 779-788.

[7] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Fu, A. C. Berg, SSD: Single shot multibox detector, *14th European Conference on Computer Vision (EVVC)*, Amsterdam, The Netherlands, 2016, pp. 21-37.

[8] H. Xie, Y. Zhang, Z. Wu, An Improved Fabric Defect Detection Method Based on SSD, *AATCC Journal of Research*, Vol. 8, No. 1_suppl, pp. 181-190, September, 2021.

[9] M. Khalid, J. M. Hager, W. Kraus, M. F. Huber, M. Toussaint, *Deep workpiece region segmentation for Bin picking*, September, 2019. https://doi.org/10.48550/arXiv.1909.03462

[10] H. Bay, A. Ess, T. Tuytelaars, L. Van Gool, Speeded-up robust features (SURF), *Computer Vision & Image Understanding*, Vol. 110, No. 3, pp. 346-359, June, 2008.

[11] F. Schweiger, G. Schroth, R. Huitl, Y. Latif, E. Steinbach, Speeded-up SURF: design of an efficient multiscale feature detector, *IEEE International Conference on Image Processing*, Melbourne, Australia, 2013, pp. 3475-3478.

[12] J. Xu, Z. Lu, X. Liu, L. Huang, L. Han, H. Yan, Dense workpiece detection based on improved SSD, *Modular Machine Tool & Automatic Manufacturing Technique*, No. 11, pp. 70-74, November, 2022.

[13] L. Bao, Q. Chang, C. Jia, P. Xiong, Q. Sun, Worpiece detection hyper segmentation algorithm based on attention mechanism, *Chinese Science and Technology Information*, No. 21, pp. 126-129, November, 2022.

[14] P. Cong, K. Lv, H. Feng, J. Zhou, Improved YOLOv3 Model for Workpiece Stud Leakage Detection, *Electronics*, Vol. 11, No. 21, Article No. 3430, November, 2022.

[15] J. Redmon, A. Farhadi, Yolov3: An incremental improvement, April, 2018. https://arxiv.org/abs/1804.02767

[16] J. Hu, L. Shen, S. Albanie, G. Sun, E. Wu, Squeeze-and-excitation networks, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 42, No. 8, pp. 2011-2023, August, 2020.

[17] J. Zhang, C. Huang, Z. Wang, Research on structural damage identification based on multi-head self-attention and convolutional neural networks, *Journal of Vibration and Shock*, Vol. 41, No. 24, pp. 60-71, December, 2022.

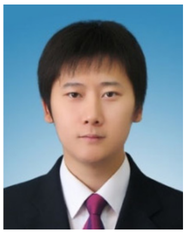
[18] E. Zhou, X. Xu, B. Xu, H. Wu, An enhancement model based on dense atrous and inception convolution for image semantic segmentation, *Applied Intelligence*, Vol. 53, No. 5, pp. 5519-5531, March, 2023.

[19] D. Yang, Y. Du, Y. Yan, H. Yao, L. Bao, Image semantic segmentation with hierarchical feature fusion based on deep neural network, *Connection Science*, Vol. 34, No. 1, pp. 1772-1784, 2022.

[20] X. Hua, X. Cui, X. Xu, S. Qiu, Y. Liang, X. Bao, Z. Li, Underwater object detection algorithm based on feature enhancement and progressive dynamic aggregation strategy, *Pattern Recognition*, Vol. 139, Article No. 109511, July, 2023.

Biographies

Xiaoning Bo received the M.S. degree in Beijing Jiaotong University. He is currently a PhD student in North University of China and working as an lecturer in Taiyuan Institute of Technology. His current research fields include computational imaging and image processing.



Zhiyuan Zhang received his PhD degree in Beijing Jiaotong University. Currently, he is an associate professor in the School of Electronics and Information Engineering at Beijing Jiaotong University. His current research fields include social network analysis, recommender system, machine learning, data mining and artificial intelligence.

Yipeng Wang received his PhD degree in Beijing Jiaotong University. Currently, he is a cyber security engineer in the Information center of China State Railway Group Co., Ltd. His current research fields include cyber defense, situation awareness, zero trust model and trusted computing.