electronics

MDPI

Article

A Workpiece-Dense Scene Object Detection Method Based on Improved YOLOv5

Jiajia Liu [1], Shun Zhang [1], Zhongli Ma [1,*], Yuehan Zeng [1] and Xueyin Liu [2]

[1] College of Automation, Chengdu University of Information Technology, Chengdu 610255, China; liujj@cuit.edu.cn (J.L.); zchuangye.zs@gmail.com (S.Z.); cuitno3@gmail.com (Y.Z.)
[2] Sichuan Machinery Research and Design Institute, Chengdu 610063, China; liuxueyin@ccjys.com
* Correspondence: mazl@cuit.edu.cn

Abstract: Aiming at the problem of detection difficulties caused by the characteristics of high similarity and disorderly arrangement of workpieces in dense scenes of industrial production lines, this paper proposes a workpiece detection method based on improved YOLOv5, which embeds a coordinate attention mechanism in the feature extraction network to enhance the network's focus on important features and enhance the model's ability to pinpoint targets. The pooling structure of the space pyramid has been replaced, which reduces the amount of calculation and further improves the running speed. A weighted bidirectional feature pyramid is introduced in the feature fusion network to realize efficient bidirectional cross-scale connection and weighted feature fusion, and improve the detection ability of small targets and dense targets. The SIoU loss function is used to improve the training speed and further improve the detection performance of the model. The average accuracy of the improved model on the self-built artifact dataset is improved by 5% compared with the original model and the number of model parameters is 14.6MB, which is only 0.5MB higher than the original model. It is proved that the improved model has the characteristics of high detection accuracy, strong robustness and light weight.

Keywords: workpiece detection; YOLOv5s; attention mechanism feature fusion; loss function

Citation: Liu, J.; Zhang, S.; Ma, Z.; Zeng, Y.; Liu, X. A Workpiece-Dense Scene Object Detection Method Based on Improved YOLOv5. *Electronics* **2023**, *12*, 2966. https://doi.org/10.3390/electronics12132966

Academic Editor: Eva Cernadas

Received: 28 May 2023
Revised: 28 June 2023
Accepted: 4 July 2023
Published: 5 July 2023

1. Introduction

Workpiece sorting is a common task in manufacturing and industrial production. Due to its high repeatability, workpiece sorting has become one of the important application scenarios of industrial robots [1]. Traditional sorting robots are pre-programmed. Although they can carry out repetitive actions, such robots cannot be adjusted according to the actual situation and must strictly set the location of the sorting workpiece. Therefore, the robots lack the ability of independent identification and have low requirements for object detection technology, leading to an increase in the error rate and lower production efficiency [2]. Hence, in the automatic production line, improving the speed and accuracy of workpiece positioning and identification has important research significance for sorting robots. However, the problem of small and dense workpieces and workpieces blocking each other in industrial automation scenarios poses a greater challenge for workpiece inspection.

The disadvantages of the traditional object detection method are that it requires a large amount of manpower to extract effective features, the model lacks generalization ability when the target features change, and the detection algorithm of single feature or multiple features loses most of the feature information of the object, which cannot be applied to the actual industrial detection scene. Liu et al. [3] used the improved SURF-FREAK algorithm to recognize and grasp the workpiece. The algorithm adopted the improved SIFT for feature extraction but the experimental results of the algorithm were poor under complex illumination conditions. Jiang et al. [4] used the contour Hu moment invariant characteristic

to match and recognize the workpiece image but this algorithm needs to manually design the feature extraction algorithm, which has some shortcomings in universality. In recent years, deep learning has been widely applied in object detection. Luigi Bibbo et al. [5] have developed a facial expression recognition system based on the Ensemble AI model that could help improve healthcare. Wang et al. [6] proposed the Faster R-CNN algorithm for identification and classification of small automotive parts under complex working conditions, which can accurately detect scattered parts, but there is a problem of missed detection for mutually occluded parts. Gong et al. [7] applied the YOLOv3 algorithm to the part recognition model, which solved the problem that it was difficult for stacked board parts to identify the blocked parts and improved the detection accuracy. However, when the workpiece was densely stacked, false detection and missed detection would also be caused. In most production lines, there are many kinds and irregular quantities of workpieces and they are placed randomly, which requires that the designed object detection algorithm and network not only have good robustness for workpiece detection in complex situations, but also have strong detection ability for dense small targets, so as to reduce the rate of missed detection.

In this paper, aiming at the problem that it is difficult to identify a large number of small targets in the industrial production line due to the dense workpieces, we compare the widely used object detection algorithms at present, and choose to improve the YOLOv5 algorithm with both speed and precision. The detection effect of the original YOLOv5 model is slightly insufficient in the detection of small targets, which are easy to miss, and false detection. Moreover, it has a large positioning error when there are many targets and dense distribution. To solve the above problems, a workpiece detection method based on improved YOLOv5 is proposed in this paper. Firstly, the coordinate attention mechanism is embedded in the backbone feature extraction network to make the network pay more attention to the region of interest and increase the feature extraction capability of the network. Secondly, the space pyramid pool structure is replaced to reduce the computation and improve the running speed. Secondly, BiFPN is used as the feature fusion network to enhance the feature fusion capability of the network, so that the location information and semantic information are fully integrated. Finally, the SIoU loss function is used to replace the CIoU loss function in the original model to accelerate the training speed and increase the convergence of the network. The comparison of multiple existing object detection algorithms shows that the improved algorithm in this paper has a higher detection accuracy for dense workpieces and can achieve accurate detection effects.

2. Workpiece Detection Algorithm Based on the Improved YOLOv5

2.1. YOLOv5 Object Detection Algorithm

YOLOv5 has high detection accuracy and speed, more flexible network deployment, and is widely used in real-time object detection research [8].

The network structure of YOLOv5 is shown in Figure 1, which is composed of the backbone, neck and head. The backbone mainly performs feature extraction and is composed of structures such as the Focus layer, CBS layer, C3 layer, SPP layer [9] and Bottleneck layer. The neck network adopts the structure of FPN (Feature Pyramid Network) [10] + PAN (Path Aggregation Network) [11], which can fuse shallow position features and deep semantic features of images, and enhance the feature fusion capability of the network, and generate feature maps of different sizes. The head part obtains the feature maps extracted from the backbone or fused from the neck to obtain the location and class of the detected targets.

As shown in Figure 2, the recovery and reuse of workpieces has become increasingly important in industry [12], for example, automobile workpiece recycling, including screws, bearings, bolts, etc. These parts come from a variety of different automobile brands and models, and are in large and disorganized quantities, making the identification and classification of the parts in recycling extremely time-consuming and labor-intensive.

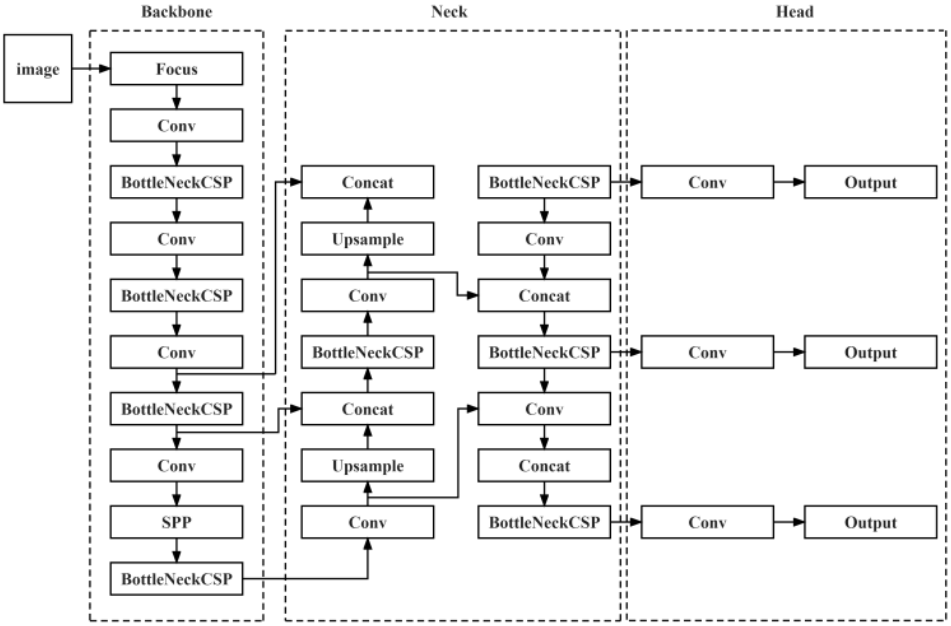


Figure 1. YOLOv5 network structure.

Figure 2. Workpiece sorting robot operation diagram.

In order to improve efficiency and accuracy, an efficient object detection algorithm is needed to identify and classify workpieces. However, the YOLOv5 model used has poor generalization for dense and small volume workpieces under different illumination; it is not suitable for practical application, so it has to be improved.

2.2. Improvements to the YOLOv5 Model

In order to solve the problem that it is difficult to accurately identify a large number of small targets formed by dense workpieces, the following improvements are made to the YOLOv5 model in this paper.

1. The coordinate attention mechanism is integrated into the backbone feature extraction network to increase the network's interest in important features and improve the feature extraction capability of the network.
2. The SPP in the original model is improved to SimSPPF, which reduces the computation and increases the running speed.
3. BiFPN structure is used for cross-layer feature fusion, which fully combines semantic information and location information to enhance the feature fusion capability of the network.

4. The CIoU loss function in the original model is improved to the SIoU loss function, and the direction matching between the real box and the predicted box is fully considered to improve the convergence performance of the model.

2.2.1. Coordinate Attention Mechanism

In view of the density and small size of some parts in industrial sorting, coordinate attention (CA) was introduced into the feature extraction network of YOLOv5 [13], which could effectively extract the feature information of small and dense targets of the workpiece and further improve the accuracy of detection.

Different from most attention mechanisms [14,15], which use maximum pooling or average pooling to process channels, the coordinate attention mechanism introduced in this paper adds location information to channel attention; the mobile network can participate in a larger area under the premise of avoiding a large number of calculations, so as to avoid the loss of location information. The introduced attention mechanism decomposes channel attention into two parallel one-dimensional feature coding processes, which aggregate features in two directions: one direction to obtain remote dependence, the other direction to retain accurate location information, and then encode the generated feature maps to form a pair of direction-aware and position-sensitive feature maps. The structure of the introduced CA module is shown in Figure 3, which uses coordinate information embedding and coordinate attention to generate the relationship between the channel and the position of the captured features.

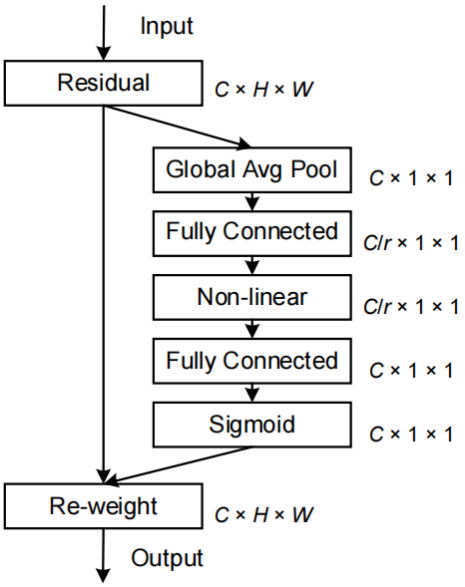


Figure 3. CA module structure.

In order to obtain the attention in the horizontal and vertical directions of the image and encode the exact position information, CA first divides the input feature graph x into horizontal $x_c(h,i)$ and vertical $x_c(i,w)$ directions for global averaging pooling. The two directions of the output $z_c^h(h)$ at the height h of channel c and the output $z_c^w(w)$ at the width w of channel are obtained.

$$z_c^h(h) = \frac{1}{W}\sum_{i=0}^{W} \mid x_c(h,i) \tag{1}$$

$$z_c^{\mathrm{w}}(\mathrm{w}) = \frac{1}{H}\sum_{j=0}^{H} \mid x_c(j,w) \tag{2}$$

Next, the horizontal feature graph z^h and vertical feature graph z^w obtained from the global receptive field are stitched together, and then they are sent into the shared 1×1 convolution transform F_1 to reduce their dimensions to the original c/r, and then the batch normalized feature graphs are sent into the nonlinear activation function δ to get the shaped $1 \times (W + H) \times c/r$ feature graph f.

$$f = \delta(F_1(|z^h, z^w|)) \tag{3}$$

Then, the feature graph f is divided into two feature vectors f^h and f^w according to the original horizontal and vertical directions, and two 1×1 convolution transform F_h and F_w, respectively, to get the feature graph with the same number of channels. After sigmoid activation function, the attention weight g^h of the feature graph in the horizontal direction and the attention weight g^w in the vertical direction are obtained.

$$g^h = \sigma(F_h()f^h) \tag{4}$$

$$g^w = \sigma(F_w()f^w) \tag{5}$$

Finally, the feature graph with attention weights in both horizontal and vertical directions will be obtained through multiplication weighting on the original feature graph.

$$y_c(i, j) = x_c(i, j) \times g_h^c \times g_c^w(j) \tag{6}$$

The CA module is a novel attention mechanism for mobile networks. It has the characteristics of being simple, flexible, and plug and play, which can improve the accuracy of the network without any extra computing overhead.

2.2.2. Simple and Fast Space Pyramid Pool

In traditional convolutional neural networks, the size of the input image must be fixed. However, in practical applications, the size of the input image is often uncertain, while the spatial pyramid pooling (SPP) [16] can flexibly obtain the output of any available dimension by increasing the number of layers of the feature pyramid or changing the size of the window. Its structure is shown in Figure 4. If the convolutional feature map of size (w, h) is input, the spatial pyramid of the first layer uses a 4×4 scale to divide the feature map into 16 pieces and the size of each piece is $(w/4, h/4)$. The second layer uses a 2×2 scale to divide the feature map into four blocks; the size of each is $(w/2, h/2)$. The third layer directly takes the whole feature map as a block, carries on the feature extraction operation and finally gets the feature vector of $21 = 16 + 4 + 1$ dimensions. SPP can not only solve the problem of inconsistent input image size, but also carry out multi-angle feature extraction and reaggregation of the feature map after convolution and pooling. SPP can significantly improve model performance and detection accuracy when used for target detection, while reducing the risk of over-fitting.

SimSPPF uses a cascade of multiple small-sized pooling kernels instead of a single large-sized pooling kernel in the SPP module while increasing the perceptual field of view. Specifically, it serial processing inputs through multiple maximum pooling layers of 5×5 size, replacing a 9×9 convolution operation with two 5×5 convolution operations and a 13×13 convolution operation with three 5×5 convolution operations. This design can not only retain the original function, but also reduce the amount of computation, improve the running speed and make the SimSPPF structure more efficient. The specific structure of SimSPPF is shown in Figure 5.

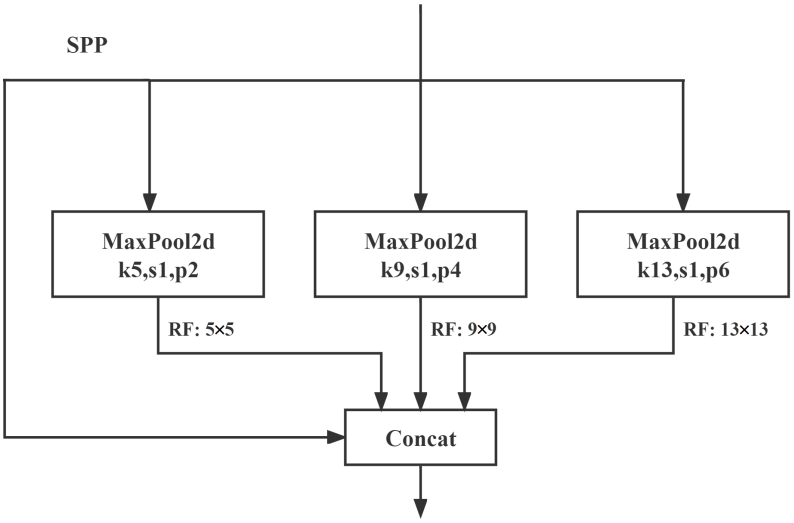


Figure 4. SPP structure diagram.

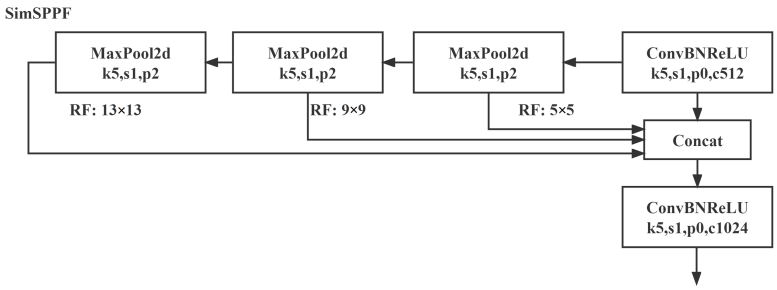


Figure 5. Structure diagram of SimSPPF.

2.2.3. Bidirectional Feature Pyramid

The purpose of the feature pyramid structure FPN is to fuse shallow position information and deep semantic information, as shown in Figure 6a. The original pyramid structure adopts the information fusion from top to bottom, which improves the information extraction ability of the network, but the fusion process will also lead to the loss of information. YOLOv5 adopts PANet structure, as shown in Figure 6b. Based on the idea of an FPN image feature pyramid, PANet not only carries out feature fusion from top to bottom but also adds feature fusion from bottom to top, so as to reduce information loss and achieve good detection results. However, the number of parameters in network training is increased. For workpiece detection, the original model has the problem of low detection accuracy due to the presence of more small target objects. Bidirectional Feature Pyramid Network (BiFPN) [17], as shown in Figure 6c, enhances the information extraction capability of the network, so that low-level position information can better combine with high-level semantic information, thus further improving the detection performance of the network for targets. The PANet structure of the original network is only stacked on the channel, while the BiFPN takes the weight information into account and implements bidirectional cross-scale feature fusion.

In this paper, BiFPN is integrated into the YOLOv5 structure to reduce the loss of feature information, improve the extraction efficiency of position information and enhance the detection ability of the network for small targets. Meanwhile, this improvement hardly increases the cost and has little impact on the size of model parameters.

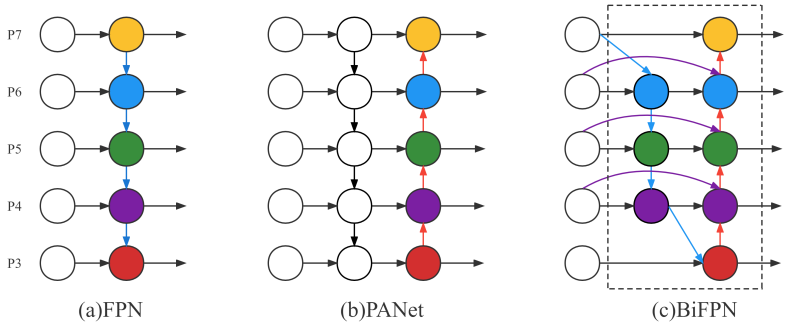


Figure 6. Schematic diagram of FPN, PANet and BiFPN structures.

2.2.4. SIoU Loss Function

The traditional object detection loss function relies on the aggregation of boundary box regression indicators [18], such as the distance between the predicted box and the real box, and the overlap area and the aspect ratio, but it ignores the direction of the mismatch between the desired real box and the predicted box. This deficiency leads to a slow convergence rate and correspondingly low efficiency of the model. For this SIoU loss function [19], the vector angle between the real box and the predicted box is introduced, and the angle, distance, shape and intersection ratio losses are redefined.

1. Angle cost
   The model first makes predictions on either the X or Y axes, and then approximates along the correlation axis. To achieve this, the convergence process will first attempt to minimize the angle, so the angle costing formula is introduced and defined.

$$\Lambda = 1 - 2 * \sin^2\left(\arcsin(x) - \frac{\Pi}{4}\right) \tag{7}$$

2. Distance cost
   Angle cost is introduced into distance cost and distance cost is redefined.

$$\Delta = \sum_{t=x}^{y}(1 - e^{-\gamma\rho t}) \tag{8}$$

3. Shape cost
   Shape cost is defined.

$$\Omega = \sum_{t=w}^{h}\left(1 - e^{-w_t}\right)^{\theta} \tag{9}$$

4. Cross and compare costs
   The crossover cost is defined.

$$L_{IoUCost} = 1 - IoU \tag{10}$$

   Finally, the SIoU loss function is defined.

$$L_{box} = 1 - IoU + \frac{\Delta + \Omega}{2} \tag{11}$$

3. Experimental Research and Result Analysis

3.1. Workpiece Dataset Establishment

3.1.1. Workpiece Data Acquisition

The sample types are common parts (screws, nuts, washers and wire screw sleeves) in industrial sorting. Taking into full account the interference brought by the external environment, different numbers and types of workpieces are randomly placed for collection

in the actual environment, so as to improve the robustness of the model. The size of the image is uniformly processed into 640 × 480; part of the image dataset is shown in Figure 7.

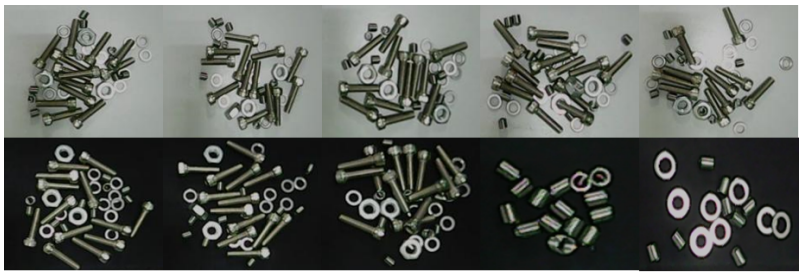


Figure 7. Part of image dataset sample.

In the training process of the deep learning network model, it is necessary to obtain the information of the target in the image accurately. In this paper, LabelImg is used to mark the image. After marking, the number, type and four vertex positions of the target can be obtained, and the corresponding .xml format tag file can be generated. The file contains the category, length, width and height information of the marked target, which is convenient for decoding and parsing.

3.1.2. Data Enhancement

A total of 1000 pictures were collected in this paper. In order to enrich the dataset, data enhancement strategies such as horizontal flip, vertical flip, cropping, affine transform, Gaussian blur, translation, adaptive Gaussian noise and brightness change were randomly introduced for the scenes of workpiece contamination, motion blur and brightness transformation in industrial sorting. And the above data enhancement strategies are randomly combined to process the training samples. After processing, the number of datasets increased to 18,000. Some random data enhancement samples are shown in Figure 8. The training set, test set and verification set were divided in a ratio of 8:1:1.

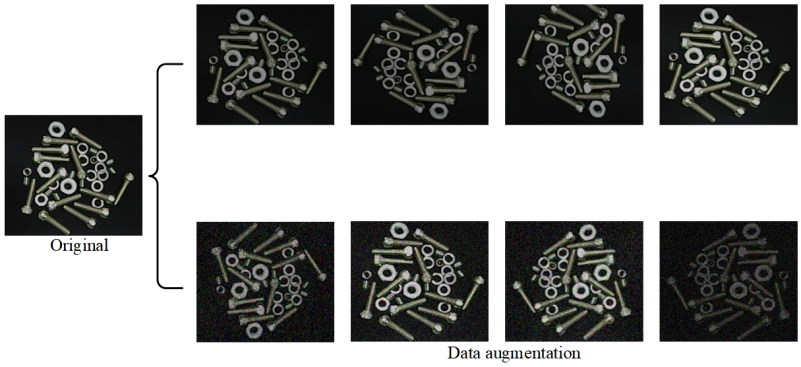


Figure 8. A sample of partial random data enhancement.

Figure 9 shows the visualization analysis results after dataset enhancement, where Figure 9a represents the distribution of object classes in the dataset, Figure 9b represents the distribution of object sizes, and horizontal and vertical represent the width and height of objects. It can be seen that the size distribution of small targets in the dataset is concentrated and occupies a large proportion.

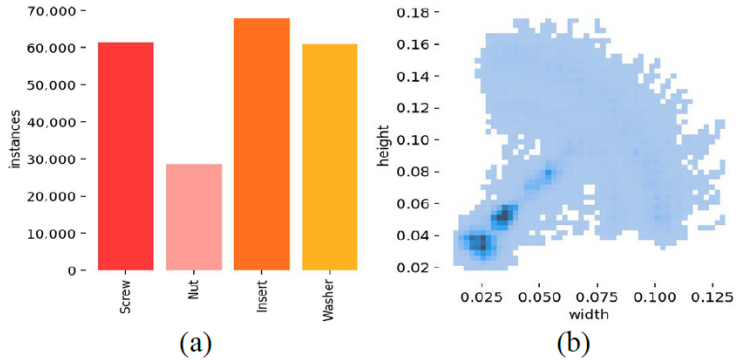


Figure 9. Data and analysis. (**a**) category distribution of objects in the data set; (**b**) size distribution of objects.

3.2. Experimental Environment and Evaluation Indicators

3.2.1. Setting the Experimental Environment and Parameters

The CPU model of the computer used in the experiment is i9-10900k, the GPU model is NVIDI A RTX3080 and the video memory size is 10 GB. The operating system is windows11 and the deep learning frame is Pytorch. In the comparison experiment of object detection algorithms, all algorithms are trained with the same dataset and the same parameter. Settings are at the same stage to ensure the comparability of experimental results. In the training process, the learning rate was set as 0.01, the momentum gradient descent algorithm was adopted for optimization, the momentum parameter was 0.937, the batch of each iteration was 16, the weight attenuation coefficient was 0.0005 and the number of iterations was uniformly set to 300.

3.2.2. Evaluation Index

In this paper, evaluation indexes such as recall, precision, AP (average precision) and mAP (mean average precision) were used to verify the accuracy of the model.

The precision rate refers to the probability that all predicted positive samples are actually positive samples, which can be calculated by

$$Precision = \frac{TP}{TP + FP} \tag{12}$$

The recall rate represents the probability of being predicted as a positive sample in the actual positive sample, calculated by

$$Recall = \frac{TP}{TP + FN} \tag{13}$$

The average accuracy refers to the area under a curve drawn with the recall rate as the axis and the accuracy rate as the axis, given by

$$AP = \int_0^1 p(r)dr \tag{14}$$

where p represents the accuracy rate, r represents the recall rate and the larger the area surrounded by the PR curve the higher the average accuracy.

The mean average precision represents the average precision of all categories in the dataset, which can reflect the accuracy and robustness of the model in target detection of different categories.

$$\mathrm{m}AP = \frac{1}{m}\sum_{i=1}^{m} AP_i \tag{15}$$

3.3. Experimental Research

3.3.1. Contrast Experiment

Comparison experiments are conducted to better demonstrate the advantages of the improved model. In this experiment, the performance of Ours-YOLOv5 model, Faster-RCNN [20], SSD [21] and YOLOv5s model was compared on the self-built workpiece dataset. Table 1 shows the comparison results of each model in mAP@0.5, weight size, parameter number and reasoning time. In addition, Figure 10 shows the curve comparison of mAP@0.5.

Table 1. Comparison of experimental results.

Model	Weight/MB	mAP@0.5/%	Params/10^6	Inference/ms
SSD	100.3	77.8	23.7	123
Faster-RCNN	159	84.6	136.0	207
YOLOv5s	14.1	89.3	7.0	12
Ours-YOLOv5	14.6	94.3 (↑ 5.0)	7.1	13

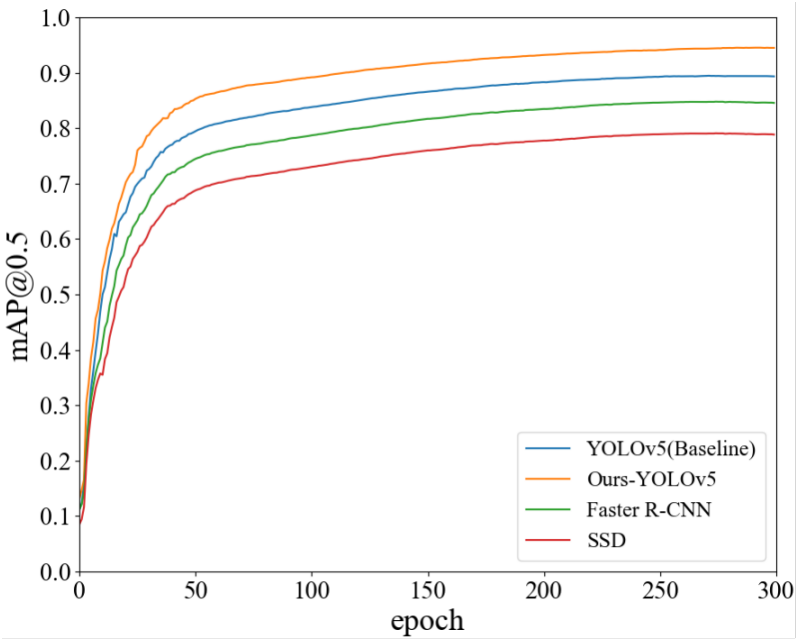
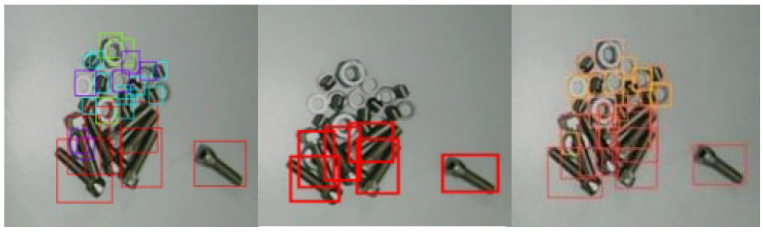


Figure 10. mAP@0.5 curve comparison.

By comparing the experimental results of different algorithm models in Table 1 and Figure 10, it can be seen that the model proposed in this paper has the highest detection accuracy compared with other mainstream models in the self-built dataset. Compared with the two-stage Faster-RCNN and first-stage SSD, the YOLOv5s model is a lightweight network model, while the improved model Ours-YOLOv5 proposed in this paper has a weight only 0.5 MB higher than that of YOLOv5s and 5.0% higher than that of YOLOv5s in mAP@0.5. Moreover, the reasoning speed is similar. The improved model in this paper has the highest detection accuracy while maintaining light weight and the original detection speed at the same time. Compared with Faster-RCNN, the average duration of reasoning video per frame is 194 ms faster and the overall performance is relatively outstanding, thus proving the superiority of the performance of Ours-YOLOv5 proposed in this paper.

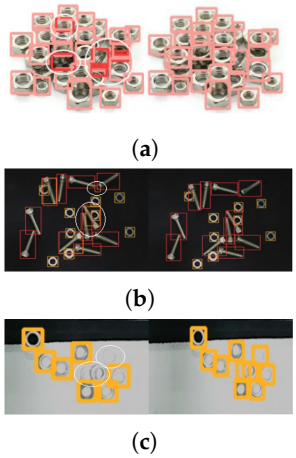
In order to more intuitively evaluate the performance of the improved model proposed in this paper, Figure 11 shows the comparison of the detection effect of the model of Faster-RCNN, SSD and Ours-YOLOv5 in the actual scene. As can be seen from the figure, the Ours-YOLOv5 model has the best detection effect without missing or false detection, while the Faster-RCNN and SSD models have multiple missing and false detections.



(a)Faster-RCNN (b)SSD (c)Ours-YOLOv5

Figure 11. Comparison of different model detection results.

Next, the detection effects of the YOLOv5s model and Ours-YOLOv5 model in different scenarios are compared, as shown in Figure 12a; the figure represents the detection of occluded targets; the left figure is the detection result of the YOLOv5s model; the white circle is the false detection target in the left figure; the right figure is the detection result of the Ours-YOLOv5 model. It can be seen that, in the white circle, parts of nuts were mistakenly detected in the left figure, while in the right figure they were successfully detected. Figure 12b shows the detection of cross-dense targets. In the left picture, when screws and nuts overlap, the original YOLOv5s model produces false detection, while, in the right picture, the improved model detects normally. Figure 12c shows the detection of small targets in a scene of strong illumination. The gasket was not identified in the left image due to the influence of illumination, while in the right image it was accurately identified successfully. To sum up, compared with the original YOLOv5s model, the improved Ours-YOLOv5 shows advantages in terms of performance, but the YOLOv5s model has poor performance in complex and diverse detection scenes, and there are cases of missing and false detection in the detection of small targets and dense targets. The Ours-YOLOv5 model has a better detection effect on small targets and dense targets, and has better robustness to different scenes, thus showing superior performance and more accurate positioning accuracy.



(**a**)

(**b**)

(**c**)

Figure 12. Comparison of test results before and after improvement. (**a**) Detection of occluded targets. (**b**) Detection of intersecting dense targets. (**c**) Detection of strong light scenes.

3.3.2. Ablation Experiment

The ablation experiment was conducted to verify the optimization effects of each improved module. The experimental results are shown in Table 2, where AAM represents adding an attention mechanism to the backbone network, RSP represents replacing the spatial pyramid pool structure, MFP represents modifying the feature pyramid structure and MTF represents modifying the loss function. Models 1 to 4 correspond to the addition of the AAM, RSP, MFP and MTF modules. Figure 13 shows the comparison of mAP@0.5 curves of the ablation experiment. All the improvements are combined into the model. The

improved model is 5% higher than the original model mAP@0.5, and the detection of small targets and dense targets is greatly improved.

Table 2. Ablation experiment comparison results.

Model	AAM	RSP	MFP	MTF	mAP@0.5/%
YOLOv5s	×	×	×	×	89.3
Model 1	√	×	×	×	91.2 (↑1.9)
Model 2	×	√	×	×	90.1(↑0.8)
Model 3	×	×	√	×	91.5 (↑2.2)
Model 4	×	×	×	√	91.0 (↑1.7)
Ours-YOLOv5	√	√	√	√	94.3 (↑5.0)

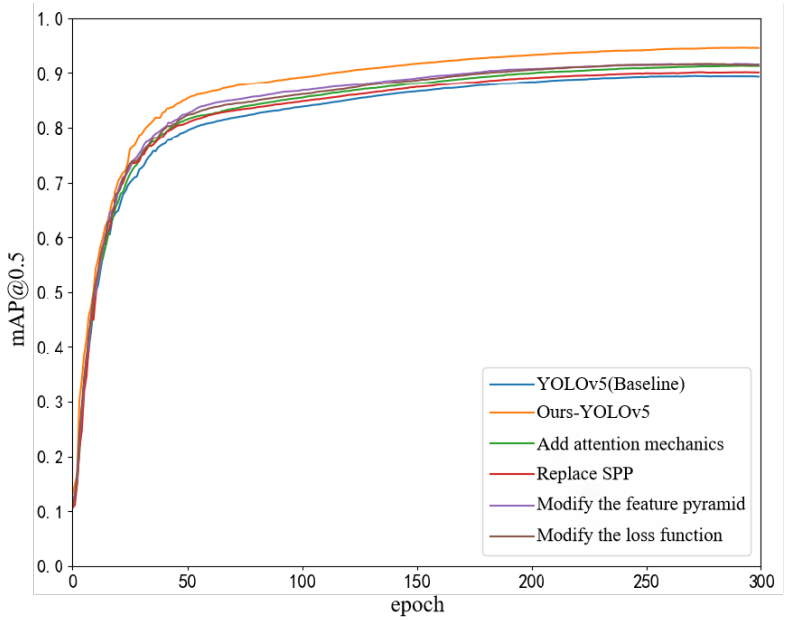


Figure 13. Ablation experiment mAP@0.5 curve comparison.

1. Analysis of the model test of increased attention mechanism
   In this paper, the CA module is added to the backbone network after feature extraction, so that it has clearer low-level contour information and coordinate information but also contains rich high-level semantic information. It can not only ensure the integrity of the feature information, but also improve the information expression ability of the feature map. According to the data in Table 2, it can be found that the index of mAP@0.5 of the model with the introduction of the attention mechanism is 1.9% higher than that of the original model, which indicates that adding the attention mechanism after the backbone network can effectively enhance the feature information.
   The accuracy rate–recall curve of the model introduced with the attention mechanism and the original YOLOv5s model on the self-built dataset is shown in Figure 14. In Figure 14a on the left, the area surrounded by the YOLOv5s blue curve is smaller than that surrounded by the axes, while in Figure 14b on the right, the area surrounded by the YOLOv5s-CA blue curve is larger than that surrounded by the axes, indicating that the classification performance of the model with the attention mechanism on the self-built dataset is improved compared with that of the YOLOv5s model.

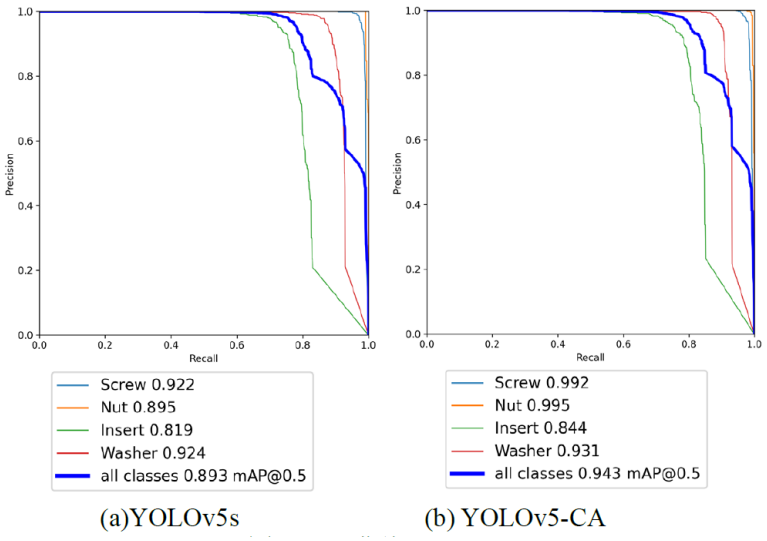


Figure 14. PR curve comparison.

Figure 15 shows the comparison of detection effects between the model with attention mechanism proposed in this paper and the YOLOv5s model. It can be seen that some wire sleeves are very small and dense, and the YOLOv5s model fails to correctly detect the targets and produces some false detections, while the YOLOv5-CA model can successfully detect these targets, indicating that it has become more accurate in the detection of small targets and dense targets after the introduction of the attention mechanism.

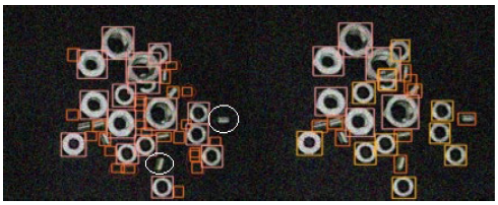


Figure 15. Test performance comparison.

2. Improved spatial pyramid pool model test analysis
In this paper, SPP in YOLOv5 was replaced by SimSPPF to increase the receptive field and uses multiple small-size pooling kernel cascades instead of a single large-size pooling kernel. Table 3 shows the comparison of the parameters of SPP and SimSPPF. Compared with SPP, the number of parameters and the amount of computation for SimSPPF decreased.

Table 3. Parameter comparison 1.

Model	Params/10^6	GFLOPs
SPP	7,225,885	16.5
SimSPPF	7,030,417	16.0

According to Table 2, the improved spatial pyramid pool model mAP@0.5 has an improvement of 0.8% over the original. The models configured with YOLOv5 and YOLOv5+simSPPF were, respectively, subjected to 50 times of reasoning and a comparison test of 100 images. The experimental comparison index was reasoning time, which could reflect the speed of image reasoning by the image processing module.

Figure 16 shows the curve comparison of reasoning time. It can be seen that the improved spatial pyramid pool model reasoning was faster than the original model. This proves that, while retaining the original function, SimSPPF reduces the amount of computation, further improving the speed and efficiency of operation.

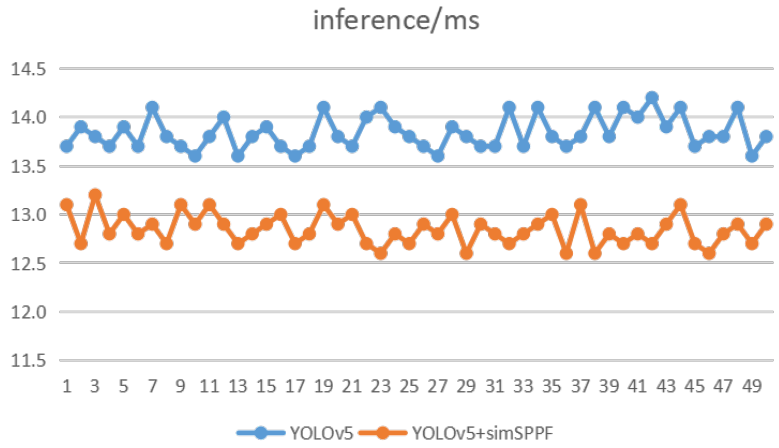


Figure 16. Inference time curve comparison plot.

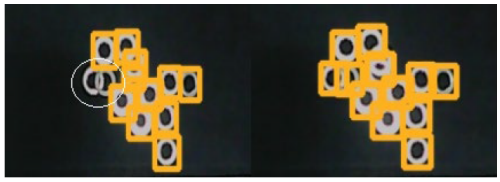
3. Improved feature pyramid model test analysis

In order to verify the performance of BiFPN added in this paper, the number of model parameters, model weight, and mAP@0.5 of FPN, PANet and BiFPN in the mainstream feature pyramid network are compared. The results are shown in Table 4. It can be seen that the detection accuracy of the FPN network in the top-down single-order direction is not high. Adding the bottom-up path on the basis of FPN improves the detection performance of the PANet network; adding the cross-layer BiFPN network on the basis of PANet has the best detection performance; mAP@0.5 increased by 2.2% compared with PANet. At the same time, the number of parameters and the weight of the BiFPN network do not increase greatly, which proves that it enhances the information extraction ability of the network, so that the low level of location information can better combine with the high level of semantic information.

Table 4. Parameter comparison 2.

Model	Params/10^6	Weight/MB	mAP@0.5/%
FPN	6.2	13.2	87.4
PANet	7.0	14.0	89.3
BiFPN	7.1	14.6	94.3

Figure 17 shows the comparison between the test effect of the improved feature pyramid model and the original model. It can be seen that the improved feature pyramid model has a better detection effect on small targets and less false detection. Therefore, it is proved that the BiFPN can extract position information more fully, reduce the loss of feature information and increase the ability of the network to detect small targets.



(a)Yolov5s (b)YOLOv5-Bi FPN

Figure 17. Comparison of test effects.

4. Improved loss function model test analysis
   In this paper, the SIoU loss function is used to replace the CIoU loss function in the original model. According to Table 2, after using the SIoU loss function, mAP@0.5 improves by 1.7% compared with using CIoU. Meanwhile, Figure 18 shows the comparison of loss curves before and after the improvement of the loss function. After the improvement, the convergence speed of the model is faster, the loss value is gradually reduced and the convergence ability is enhanced. This indicates that SIoU is used instead of CIoU in this paper to solve the problem of direction matching between the real box and the predicted box, and the convergence performance of the model is improved.

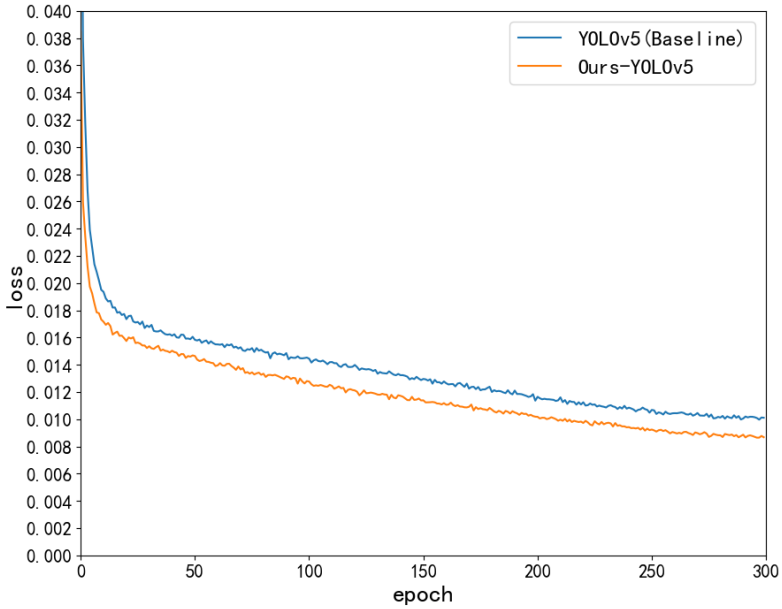


Figure 18. Improved before and after loss curve comparison.

4. Conclusions

In order to solve the problem of difficult identification caused by small and dense workpieces in industrial production lines, a workpiece detection method based on an improved YOLOv5 is proposed in this paper. Corresponding improvements are made in the backbone network, spatial pyramid pool structure, feature fusion network and loss function. The experimental results show that, compared with the current mainstream object detection algorithms, the improved model has the characteristics of small volume, high detection accuracy and fast reasoning speed, and can accurately detect the target and meet the real-time detection. Compared with the original YOLOv5s, the average accuracy of dense workpiece detection by the improved model is increased by 5% in the case of a small volume increase. In the industrial production line, sorting errors, missed inspection and other problems can cause great losses to the assets of the factory; increasing the accuracy by 5% can improve the assets of the factory, by providing a feasible method for actual workpiece detection.

Author Contributions: Methodology, Z.M.; Formal analysis, Z.M.; Resources, Y.Z.; Data curation, Y.Z. and X.L.; Writing—original draft, J.L.; Writing—review & editing, S.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the International Cooperation Project of Science and Technology Bureau of Chengdu OF FUNDER grant number No.2019-GH02-00051-HZ. Sichuan unmanned system and intelligent perception Engineering Laboratory Open Fund and Research Fund of Chengdu University of information engineering, under Grants (No.WRXT2020-001, No.WRXT2020-002, No.WRXT2021-002 and No.KYTZ202142) and the Sichuan Science and Technology Program China, under Grant (No.2022YFS0565). This paper is also supported by the Key R&D project of the Science and Technology Department of Sichuan Province, under Grants (2023YFG0196 and 2023YFN0077), the Science and Technology achievements transformation Project of Science and Technology Department of Sichuan Province, under Grant (2023JDZH0023) and the Sichuan Provincial Science and Technology Department, Youth Fund project, under Grant (2023NSFSC1429).

Data Availability Statement: Due to privacy or ethical restrictions, we are unable to disclose self-built datasets. We suggest that other datasets related to small target detection can be used.

Acknowledgments: The authors are grateful to the College of Automation, Chengdu University of Information Technology. This paper is supported by the International Cooperation Project of Science and Technology Bureau of Chengdu (No.2019-GH02-00051-HZ), Sichuan unmanned system and intelligent perception Engineering Laboratory Open Fund and Research Fund of Chengdu University of information engineering, under Grants (No.WRXT2020-001, No.WRXT2020-002, No.WRXT2021-002 and No.KYTZ202142), and the Sichuan Science and Technology Program China, under Grant (No.2022YFS0565). This paper is also supported by the Key R&D project of the Science and Technology Department of Sichuan Province, under Grants (2023YFG0196 and 2023YFN0077), the Science and Technology achievements transformation Project of Science and Technology Department of Sichuan Province, under Grant (2023JDZH0023) and the Sichuan Provincial Science and Technology Department, Youth Fund project, under Grant (2023NSFSC1429).

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. Zhu, Y. Development of Robotic Arm Sorting System Based on Deep Learning Object Detection. Master's Thesis, Zhejiang University, Hangzhou, China, 2019.
2. Dang, H.; Hou, J.; Qiang, H.; Zhang, C. SCARA robot based on visual guiding automatic assembly system. *J. Electron. Technol. Appl.* **2017**, *43*, 21–24. [CrossRef]
3. Liu, J.; Zhong, P.; Liu, M. Research on Workpiece Recognition and Grasping Method Based on Improved SURF_FREAK Algorithm. *Mach. Tool Hydraul.* **2019**, *47*, 52–55+82.
4. Jiang, B.; Xu, X.; Wu, G.; Zuo, Y. Contour Hu invariant moments of workpiece, image matching and recognition. *J. Comb. Mach. Tools Autom. Process. Technol.* **2020**, 104–107+111. [CrossRef]
5. Bibbo', L.; Cotroneo, F.; Vellasco, M. Emotional Health Detection in HAR: New Approach Using Ensemble SNN. *Appl. Sci.* **2023**, *13*, 3259. [CrossRef]
6. Wang, B. Positioning and Grasping Technology of Small Parts of Automobile Based on Visual Guidance. Master's Thesis, Yanshan University, Qinghuangdao, China, 2019. [CrossRef]
7. Gong, W.; Zhang, K.; Yang, C.; Yi, M.; Wu, J. Adaptive visual inspection method for transparent label defect detection of curved glass bottle. In Proceedings of the 2020 International Conference on Computer Vision, Image and Deep Learning (CVIDL), Chongqing, China, 10–12 July 2020; pp. 90–95.
8. Chen, Y.; Alifu, K.; Lin, W. CA-YOLOv5 for crowded pedestrian detection. *Comput. Eng. Appl.* **2022**, *1*, 1–10.
9. Tan, M.; Le, Q. Efficientnet: Rethinking model scaling for convolutional neural networks. In Proceedings of the International Conference on Machine Learning, Long Beach, CA, USA, 9–15 June 2019; pp. 6105–6114.
10. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2117–2125.
11. Liu, S.; Qi, L.; Qin, H.; Shi, J.; Jia, J. Path aggregation network for instance segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 8759–8768.
12. Shen, X. *Analysis of Automobile Recyclability*; Heilongjiang Science and Technology Information: Harbin, China, 2012; Volume 90.
13. Hou, Q.; Zhou, D.; Feng, J. Coordinate attention for efficient mobile network design. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Virtual, 19–25 June 2021; pp. 13713–13722.

14. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 7132–7141.
15. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. Cbam: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19.
16. He, K.; Zhang, X.; Ren, S.; Sun, J. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *37*, 1904–1916. [CrossRef] [PubMed]
17. Tan, M.; Pang, R.; Le, Q.V. Efficientdet: Scalable and efficient object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, hlSeattle, WA, USA, 13–19 June 2020; pp. 10781–10790.
18. Zheng, Z.; Wang, P.; Liu, W.; Li, J.; Ye, R.; Ren, D. Distance-IoU loss: Faster and better learning for bounding box regression. *Proc. AAAI Conf. Artif. Intell.* **2020**, *34*, 12993–13000. [CrossRef]
19. Gevorgyan, Z. SIoU loss: More powerful learning for bounding box regression. *arXiv* **2022**, arXiv:2205.12740.
20. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Trans. Pattern. Anal. Mach. Intell.* **2017**, *39*, 1137–1149. [CrossRef] [PubMed]
21. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. Ssd: Single shot multibox detector. In *Computer Vision–ECCV 2016, Proceedings of the 14th European Conference, Amsterdam, The Netherlands, 11–14 October 2016*; Springer International Publishing: Berlin/Heidelberg, Germany, 2016; pp. 21–37.