processes

MDPI

Article

Lightweight Yolov4 Target Detection Algorithm Fused with ECA Mechanism

Chunguang Wang [1,2], Yulin Zhou [1,*] and Junjie Li [3]

[1] School of Mechanical Engineering, Yanshan University, Qinhuangdao 066004, China; chunguang0@163.com
[2] School of Mechanical Engineering, Tianjin Sino-German University of Applied Sciences, Tianjin 300350, China
[3] School of Mechanical Engineering, Tianjin University of Science and Technology, Tianjin 300202, China; jaysenlee@163.com
* Correspondence: zhouyl61@163.com

Abstract: For the task of garbage classification, to overcome the main disadvantages of the Yolov4 target detection algorithm, such as the large network model and lower detection accuracy for small objects, a lightweight Yolov4 target detection network based on the EfficientNet-B0 fusion ECA mechanism is presented. The lightweight EfficientNet was used to replace the original backbone network, which reduces the parameters of the network model and improves the detection accuracy. Moreover, a deep detachable convolution block replaced the common convolution block in the original network, which further reduced the number of parameters in the model. In the feature pyramid model PANet, a lightweight ECA attention mechanism was introduced to realize the weight analysis of the importance of different channel feature maps through cross-channel interaction, allowing the network to extract more obvious features with which to distinguish categories. Finally, a Soft-NMS algorithm was introduced in the post-processing stage of the detection frame to reduce the missed target detection rate in dense areas, which can improve the detection accuracy of the network and detection efficiency. As shown in the results, the size of the model was only 48 MB, and the mAP was 91.09%. Compared with the original Yolov4 network, the mAP was increased by 5.77% based on the 80% reduction in the model size. The recognition of small targets was also improved, which proved the effectiveness and robustness of the improved algorithm.

Keywords: Yolov4; EfficientNet; depth-wise separable convolution; ECA; Soft-NMS

check for updates

Citation: Wang, C.; Zhou, Y.; Li, J. Lightweight Yolov4 Target Detection Algorithm Fused with ECA Mechanism. *Processes* **2022**, *10*, 1285. https://doi.org/10.3390/pr10071285

Academic Editors: Qiang Gao, Xiaoming Yuan, Chuan Wang, Bo Hu and Yong Zhu

Received: 18 May 2022
Accepted: 20 June 2022
Published: 29 June 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

1. Introduction

With the development of society, the generation of garbage is inevitable and garbage disposal has become a hot topic. Landfills are currently the main way to dispose of garbage in China, but they not only occupy a large amount of land resources, but non-naturally degraded garbage can also seriously pollute oil and water resources after landfilling [1]. Therefore, it is particularly important to classify the waste before it enters the landfill. The recyclable waste is sorted out, recycled, and reused, and the non-degradable waste is treated accordingly. This not only improves the utilization rate of waste but also reduces environmental pollution. At present, the garbage sorting method occurs mainly through manual sorting. Because the total daily amount of garbage is too large and the manual sorting energy is limited, the problems of low sorting efficiency and low recovery rate result. Moreover, working in this environment for a long time will also cause harm to the human body [2], so automation technology has become the development trend. With the development of computer vision, automatic classification of garbage is realized according to its characteristics. The traditional computer vision uses SIFT, SURF [3–5], and other algorithms to extract edge features. However, with the increase in garbage types, the features will increase, and the classification effect will not be guaranteed.

With the wide application of deep learning in target detection, the deep convolutional neural network has shown good detection results in various complex environments. The general target detection network is mainly composed of two parts. One is the backbone network used to extract the feature information of the target object (Backbone). The other part is the head used to predict the classification (Head) [6–8]. With the improvement of target detection performance requirements, some classic backbone networks have emerged, such as VGG and GooleNet proposed by ILSVRC in 2014 [9,10], ResNet proposed in 2015 [11], and MobileNet proposed in 2017 [12]. The head of the target detection network can be divided into single-stage target detection and two-stage target detection. The two stages are image segmentation feature extraction through selective search or RPN. The detection speed is slow, but the accuracy is high. The common models are Fast R-CNN [13] and Faster R-CNN [14]. The other stage is a single-stage target detection represented by Yolo and SSD, which directly uses CNN convolution feature extraction. Thus, it has the characteristics of faster speed, but the accuracy is slightly worse. The price of improving the target detection performance is the increase in the number of layers of convolution, which leads to an increase in the number of parameters and computation of the model. This is not conducive to deployment on embedded devices, so the efficiency and storage issues are addressed to enable the network model to work better in practice. For such problems, channel pruning, quantization, knowledge distillation, and other methods can be used to reduce the complexity of the model and improve its inference speed. Jingdong Lin et al. [15] summarized the optimization techniques of convolutional neural networks in four aspects: pruning and sparsification, tensor decomposition, knowledge migration, and fine module; Pingwei Shao et al. [16] proposed using MobileNet as the backbone extraction network of Yolov3, and the model was reduced by 90% compared to Yolov3, with basically no loss of detection accuracy.

Referring to the above methods, this paper proposes a lightweight Yolov4 target detection network based on EfficientNet-B0 fused with an ECA mechanism. The backbone extraction network uses EfficientNet-B0 to replace the original CSPDarkNet53 network and deep separable convolution instead of normal convolution in the feature pyramid network PANet. Embedding a lightweight ECA attention mechanism after feature pyramid upsampling and downsampling completes the network with no loss of detection accuracy, and the computational and parametric quantities of the model are substantially reduced, which improves the utility of the model.

2. Detection Model Based on Yolov4

Yolov4 is a single-stage target detection algorithm based on deep convolutional neural networks that consist of three main components: the backbone network CSPDarknet53, the feature fusion structure SPP + PANet, and the feature prediction Yolo-Head [17,18]; the network model is shown in Figure 1. Yolov4 has undergone a large number of improvements compared with Yolov3 [19], including an improved backbone network CSPDarknet53, Mosaic data enhancement, a loss function CIoU (Complete-Interp over Union) with better effect, and a Focal-loss function to balance positive and negative samples, an added attention mechanism, etc. These improvements greatly enhance detection accuracy and speed.

The backbone network CSPDarknet53 is based on the improvement of Darknet53. The original residual module was changed to the CSP structure shown in Figure 2 [20]. The CSP structure changes the original residual module into two parts: one part goes through the residual network, and the other part is directly connected to the residual. The output of the poor network is merged. This method can maintain high accuracy while reducing the number of parameters and amount of calculation.

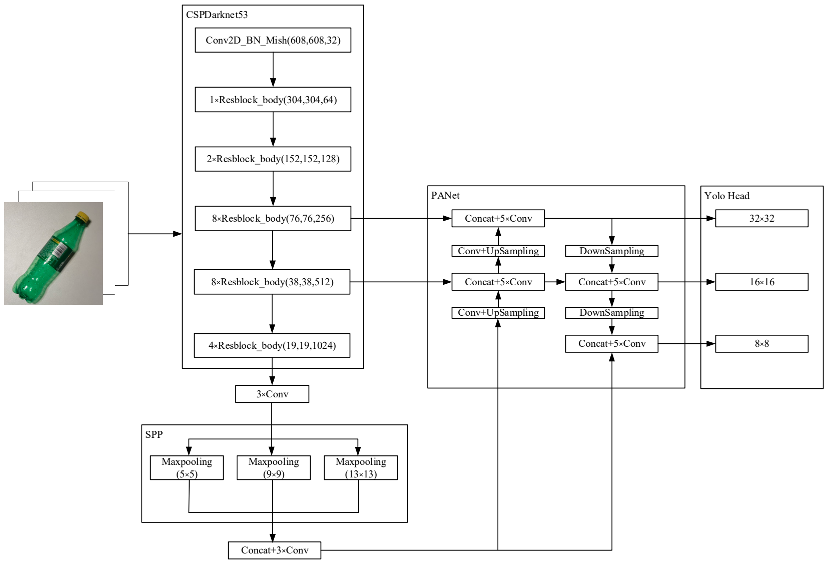


Figure 1. Yolov4 network structure.

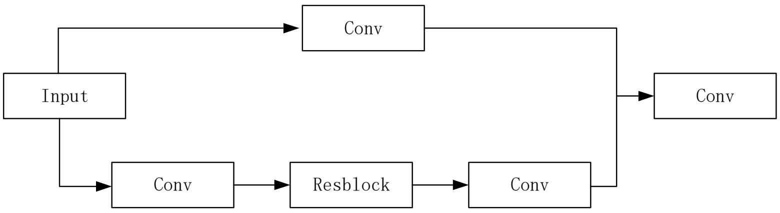


Figure 2. CSP structure block.

The SPP module uses pooling kernels of different sizes to perform maximum pooling on the convolution results [21], which can effectively separate features and increase the receptive field. The PANet network is used to build a feature pyramid, analyze semantic features from top to bottom, and reduce the loss of underlying feature information in the propagation process of the FPN algorithm. It does this by adding bottom-up feature fusion paths to the feature layer results obtained by downsampling and upsampling [22–24], which increases the richness of the feature map and improves the detection accuracy of the algorithm. The feature map after feature fusion is predicted by Yolo-head in three different sizes. The prediction results include offset x, y, detection frame size w, h, and confidence c of target information. Finally, through the intersection ratio loss IoU (intersection over union) and NMS (non-maximum suppression), redundant detection is eliminated to complete the target prediction. The Yolov4 algorithm mainly includes regression and classification networks. During the training process, the regression network calculates the regression loss through the regression loss function. The loss function consists of the intersection ratio loss IoU, confidence loss conf, and classification loss class. The total loss function is defined as follows:

$$loss = loss_{\mathrm{CIoU}} + loss_{conf} + loss_{class} \tag{1}$$

The calculation formula of the intersection ratio loss function is as follows:

$$loss_{\mathrm{CIoU}} = 1 - \mathrm{IoU} + \frac{\rho^2\left(c_{pre}, c_{gt}\right)}{d^2} + av \tag{2}$$

$$a = \frac{v}{1 - \mathrm{IoU} + v} \tag{3}$$

$$v = \frac{4}{\pi^2}(\arctan\frac{w^{gt}}{h^{gt}} - \arctan\frac{w}{h})^2 \tag{4}$$

Among them c_{pre}, c_{gt} represent the center positions of the prediction frame and the real frame, respectively; ρ represents the Euclidean distance between the two center points; d represents the diagonal length of the overlapping area between the prediction frame and the real frame; a and v represent the trade-off parameters and measure aspect ratio consistency parameters; w^{gt}, h^{gt} are the width and height of the real box; w, h are the width and height of the predicted box, respectively. The CIoU loss function also considers the intersection ratio, center point distance, and aspect ratio, which effectively improves the regression speed and prediction accuracy.

The evaluation Indicators of the network model mainly analyze the performance of accuracy and real-time performance. The indicators to measure the accuracy include Precision, Recall, AP, mAP, etc. To analyze network performance, a confusion matrix should be introduced, as shown in Table 1:

Table 1. Confusion matrix.

		Predictive Value	
		Positive Sample	**Negative Sample**
actual value	positive sample	TP (True Positive)	FN (False Negative)
	negative sample	FP (False Positive)	TN (True Negative)

Precision represents the ratio of the value of the positive sample predicted to be a positive sample to all predicted positive samples.

$$Precision = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{5}$$

Recall represents the proportion of the correct number of targets among all the targets identified, and the expression is as follows:

$$Recall = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{6}$$

AP is the area enclosed by the P-R curve (the abscissa is the curve of Recall and the ordinate is the Precision) and the coordinate axis. The larger the AP, the better the detector effect. The expression is as follows:

$$\mathrm{AP} = \int_0^{1\int} Precision\ d\ Recall \tag{7}$$

mAP represents the average of all categories of AP, so mAP can represent the average accuracy of all categories:

$$\mathrm{mAP} = \frac{\sum_{i=1}^{n} \mathrm{AP}}{n} \tag{8}$$

3. Model Optimization

Although the target detection based on Yolov4 satisfies the classification and detection of garbage, there are still the following problems: the number of model parameters and the calculation amount are too large, which is not conducive to deployment on embedded devices; the detection accuracy of small objects such as cigarette butts is not high. Therefore, to improve the robustness of the model, the following improvements were made to the original network to enhance the practicability of the detection model.

3.1. Experiment Method

The main extraction network of Yolov4 is CSPDarknet53. Although the calculation of the model is optimized based on Darknet53, including 23 CSP modules, the CSPDarknet53 network is still too complex for tasks such as garbage sorting. With the increase in the number of CSP modules and channels, the number of parameters and amount of calculation increase directly. Excessive parameters and calculations will lead to an increase in training time and a decrease in detection speed. Therefore, to improve the detection efficiency, we considered replacing the main network with a more lightweight network. EfficientNet [25] is a lightweight network structure improved based on MobileNetV3 [26]. In order to balance the correlation between network width, depth, and input image resolution, EfficientNet uses a fixed ratio to perform compound scaling on depth, width, and image size; by setting different scaling coefficients φ, the optimal image depth and network width are calculated, and the scale factor of the resolution: α, β, γ, so as to obtain 8 different scales of EfficientNetB0-B7. The input image size of EfficientNetB0 is 224 × 224, while the input image size of EfficientNetB7 is increased to 600 × 600. With the increase in the input image size, the extraction ability of image features is enhanced, but the amount of computation increases. As the training time increases, the memory occupied will also increase sharply. Therefore, in order to highlight the lightweight characteristics, this paper adopts EfficientNetB0 with the least computational load as the backbone feature extraction network, and introduces the ECA attention mechanism after the three output layers of the backbone network to enhance the feature extraction ability, thereby improving the detection accuracy.

EfficientNet follows the inverted residual structure block MBConv of MobileNetV3, as shown in Figure 3. The structure block first expands the channel through a 1 × 1 convolution kernel to enrich the feature information, and then uses the depth-wise separable convolution to carry out features extraction and introduce the SE channel attention mechanism; at the same time, it filters out the importance of different channel features of the feature map, and finally compresses the channel through a 1 × 1 convolution kernel.

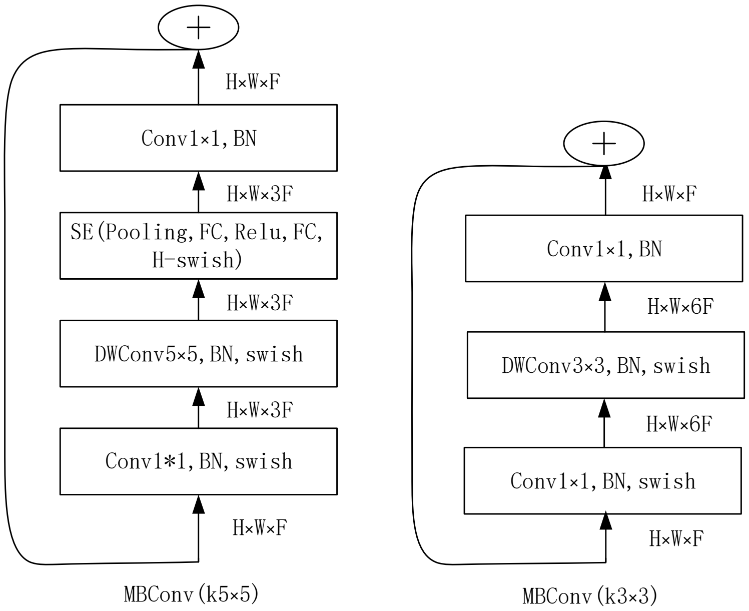


Figure 3. Inverted residual block.

EfficientNet-B0 has a total of 16 MBConv modules. To better integrate with the PANet input layer and follow the multi-scale detection of Yolov4 [27], this paper added three cascades at the end of the backbone network to realize the characteristic pyramid structure of PANet, as shown in Figure 4. The output layer of the fifth MBConv module is the P3 layer, the output layer of the 11 MBConv modules is the P4 layer, and the output layer of the 16 MBConv modules is the P5 layer.

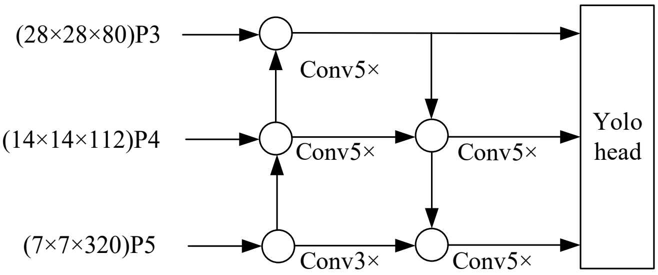


Figure 4. PANet featurized image pyramid.

In the inverted residual structure block, the ordinary convolution is replaced by the depth-wise separable convolution. Compared with the ordinary convolution, the depth-wise separable convolution first performs the channel-by-channel convolution of feature extraction and then performs the point-by-point volume of the expansion channel. The product [28] is shown in Figure 5; under the premise of ensuring the number of feature map channels, the number of parameters and amount of computation are greatly reduced.

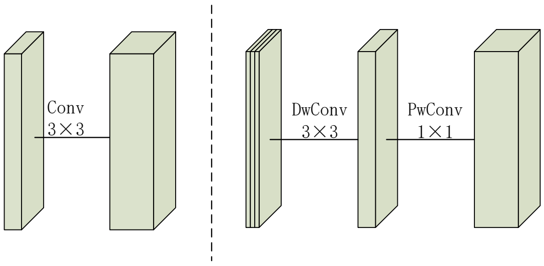


Figure 5. Depth-wise separable convolution.

Taking the input image as size $D_f \times D_f \times M$, and the convolution kernel size as size $D_k \times D_k \times M$, for example, the number of output channels is N, and the calculation amount of ordinary convolution is:

$$D_k \times D_k \times M \times D_f \times D_f \times N$$

The computation of deep separable convolution is:

$$D_k \times D_k \times M \times D_f \times D_f + M \times D_f \times D_f \times N$$

The parameters of ordinary convolution are:

$$(D_k \times D_k \times M) \times N$$

The parameter number of deep separable convolution is:

$$(D_k \times D_k \times 1) \times M + (1 \times 1 \times M) \times N$$

When the step size is 1, the number of parameters and the amount of calculation of the deep separable convolution are $\frac{1}{N} + \frac{1}{D_k^2}$ times that of the ordinary convolution. Therefore, the number of parameters and amount of calculations in the network model are greatly reduced. With the deepening of the number of network layers, the number of channels is larger and the reduction in the number of parameters and calculations is greater, which theoretically greatly improves the model's performance and detection speed. EfficientNet-B0 has a total of 16 residual blocks, seven 3 × 3 convolution kernels, and nine 5 × 5 convolution kernels. After the replacement of the backbone extraction network, Yolov4-EfficientNet-B0, compared with Yolov4, has a 38% reduction in the original network parameters and a 36% reduction in model size.

3.2. PANet Network Optimization

In the structure diagram of Yolov4 in Figure 1, it can be seen that there are 20 ordinary convolutions in the feature pyramid of PANet, of which 8 are upsampling and downsampling with a 3 × 3 convolution kernel, resulting in too many parameters and too long a detection time. Therefore, to achieve the lightweight purpose, following the idea of EfficientNet, the depth-wise separable convolution was replaced by ordinary convolution. Compared with the original network, the final parameter quantity is reduced by 80%, and the model size is also reduced by 80%. It can be seen that the overall model has been greatly reduced, enabling the deployment of embedded devices.

The Yolov4 network model has the problem of low detection accuracy for small objects such as cigarette butts. The main reason is that with the deepening of the network layers, the features of small objects are lost during downsampling, but downsampling is a key step to reduce the amount of computation. Therefore, retaining important features is the key to improving the accurate detection of small objects [29–31]. Inspired by the attention mechanism in the EfficientNet network, the attention mechanism was embedded into the feature pyramid PANet [32,33] to enhance the feature extraction ability of the network. SENet [34] is used in EfficientNet. Firstly, through the Squeeze operation of the channel, the dimension is reduced along the spatial dimension. Secondly, the excitation operation generates the weight of a channel. Finally, the importance of the channel is analyzed by reweighting, as shown in Figure 6.

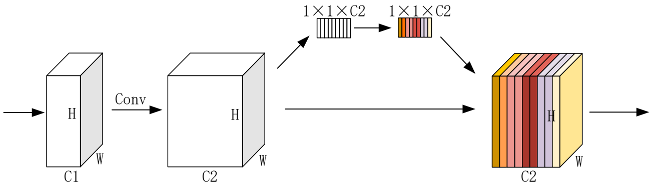


Figure 6. SENet structure chart.

Although the full connection dimension reduction operation contained in SENet can reduce the complexity of the network model, it destroys the direct correspondence between the weights and channels, which will affect the correlation prediction. At the same time, the correlation analysis of all channels will also reduce the computational efficiency of the network. Considering the number of network parameters and the amount of calculation, the latest lightweight attention mechanism ECANet [35–37] network was selected. As shown in Figure 7, compared with SENet, the dimension reduction operation is canceled. When the network completes the global average pooling, the local cross-channel connection is directly carried out, which reduces the number of model parameters and improves the classification accuracy.

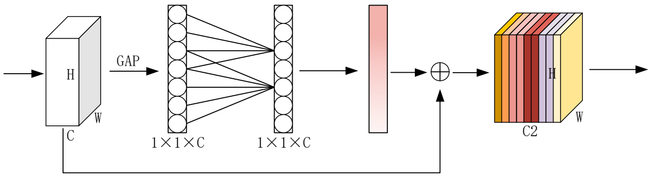


Figure 7. ECANet structure chart.

The improved PANet network is shown in Figure 8: (1) The feature layers of the main network with three output channels of 40, 112, and 320 are input into the ECA module. The end-to-end learning of the channel weight in each ECA module is carried out by a convolutional neural network, so the network can abandon the redundant features, reduce the amount of network calculation, accelerate the convergence speed, and improve the classification accuracy of the model. (2) The ECA module is embedded after upsampling and downsampling in the feature pyramid to aggregate the global context information of the feature map to provide important global semantic information.

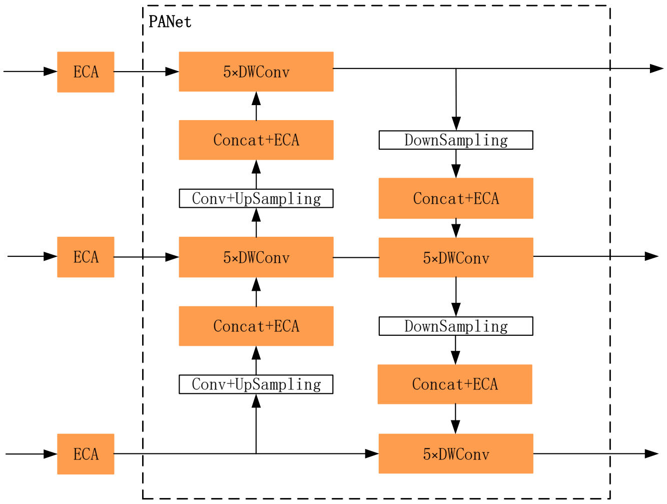


Figure 8. Improved PANet model.

3.3. NMS Algorithm Optimization

Yolov4 will generate a lot of redundant candidate boxes in the prediction stage. To remove redundant overlapping candidate boxes, the non-maximum suppression algorithm (NMS) is used to filter out the detection box with the highest score. The main idea is to retain the extreme score by searching the prediction box for a large value, suppress the non-maximum value, extract the prediction frame with higher confidence [38], and finally retain the optimal detection. The expression of the NMS algorithm is shown in Formula (9).

$$s_i = \begin{cases} s_i, iou(M, b_i) \geq N_t \\ 0, iou(M, b_i) \geq N_t \end{cases}. \tag{9}$$

s_i represents the score of the current prediction box, b_i represents the prediction box corresponding to the score, M represents the prediction box with the largest score, and N_t represents the threshold of the prediction box. The principle of the NMS algorithm is to sort the list of preselected boxes by the confidence of the detection boxes, remove the detection boxes M with the highest confidence from the list, and finally remove all detection boxes in B whose overlap rate threshold is greater than N_t.

This paper introduces the Soft-NMS algorithm (softening non-maximum suppression) [18], which no longer directly removes all candidate boxes whose overlap rate is greater than the threshold but adds a weighted penalty item, a confidence suppression function, and the greater the overlap rate, the higher the penalty. The higher the coefficient, the smaller the corresponding score s_i, which is conducive to the detection of overlapping targets, thereby reducing missed detections. There are two weighting methods of the Soft-NMS algorithm, linear weighting, and Gaussian weighting. The expressions are as follows:

(1) Linear weighting:

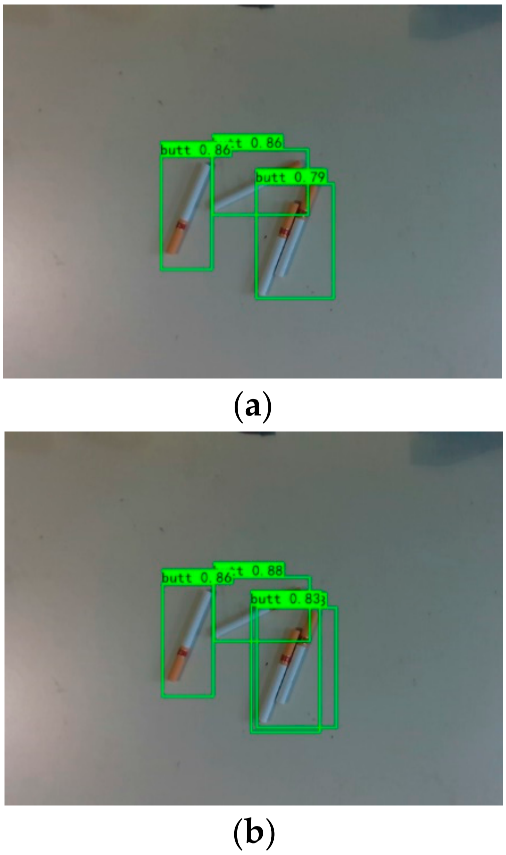
$$s_i = \begin{cases} s_i, & iou(M, b_i) < N_t \\ s_i \times f(iou(M, b_i)), & iou(M, b_i) \geq N_t \end{cases} \tag{10}$$

(2) Gaussian weighting:

$$s_i = \begin{cases} s_i, & iou(M, b_i) < N_t \\ s_i e^{-\frac{iou(M, b_i)^2}{\sigma}}, & iou(M, b_i) \geq N_t \end{cases} \tag{11}$$

where A is the suppression function and B is the Gaussian coefficient.

The detection effect of the introduction of Soft-NMS is shown in Figure 9. It can be seen that when the cigarette butts overlap, the NMS algorithm only outputs one detection frame, while the Soft-NMS algorithm outputs two adjacent target detection frames, reducing the missing detection of small models. Compared with the traditional NMS algorithm, the Soft-NMS algorithm has roughly the same detection process. The computational complexity is C, and the Soft-NMS uses an IoU-based penalty weighting term to further adjust the candidate frame. Therefore, in this experiment, the Soft-NMS algorithm is used in the post-processing of the detection frame of the Yolov4 algorithm to obtain a better detection effect.



(**a**)

(**b**)

Figure 9. Comparison between NMS and Soft-NMS. (**a**) The detection efficiency of the NMS algorithm, (**b**) Detection effect of the Soft-NMS algorithm.

4. Experiment and Analysis

4.1. Data Set and Experiment Configuration

The data set used in the experiment in this paper contained 6 categories, each with 500 pieces of data. To prevent overfitting caused by too few data, this paper used Gaussian noise and random color transformation for data enhancement. The final data quantity was 3000 sheets.

The training and testing environments of this article were run on a Windows 10 system, an NVIDIA GeForce RTX 2080TI GPU with 16G of video memory. A data set of 3000 images was divided, 90% of which were used for training and 10% for validation. We trained 100 epochs and froze the last 50 epochs, and the initial value of the learning rate for the first 50 epochs was 0.01 and the batch size was 32; the initial value of the learning rate for the last 100 epochs was 0.001, and the batch size was 16. The Adam optimizer was used during training, and a cosine annealing learning rate decay strategy was adopted.

4.2. Experimental Results and Analysis

It can be seen from the global loss function curve in Figure 10 that the loss curve decreased rapidly at the beginning of the iteration, indicating that the model was rapidly fitting, and the learning efficiency of the model was high. When epochs reached 40, the network model gradually tended to be stable. After 100 iterations, the final loss function of the model converged to 1.50. In addition, by observing the consistency of the loss curve of the training set and the verification set, we noted that the generalization ability of the model reached the best state.

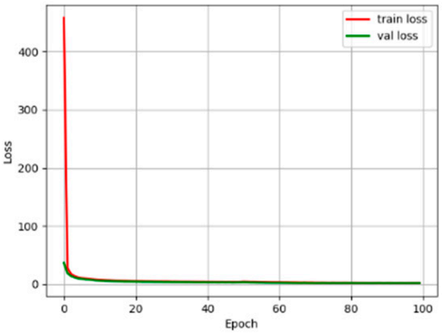
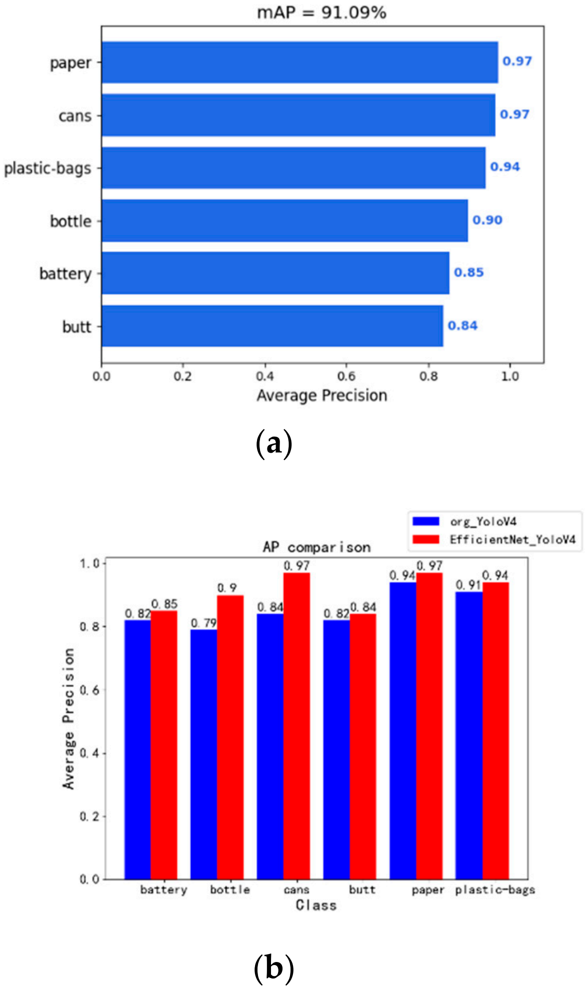


Figure 10. Loss value curve.

The detection accuracy of the improved model is shown in Figure 11, where Figure 11a is the average detection accuracy (mAP) and Figure 11b is the AP value comparison of various targets before and after improvement. It can be seen that all categories of detection accuracy were improved by the improved model in this paper. Among them, the detection accuracy of large objects such as paper chips, plastic bottles, and cans was high, and the detection accuracy of small objects such as batteries and cigarette ends was slightly lower.

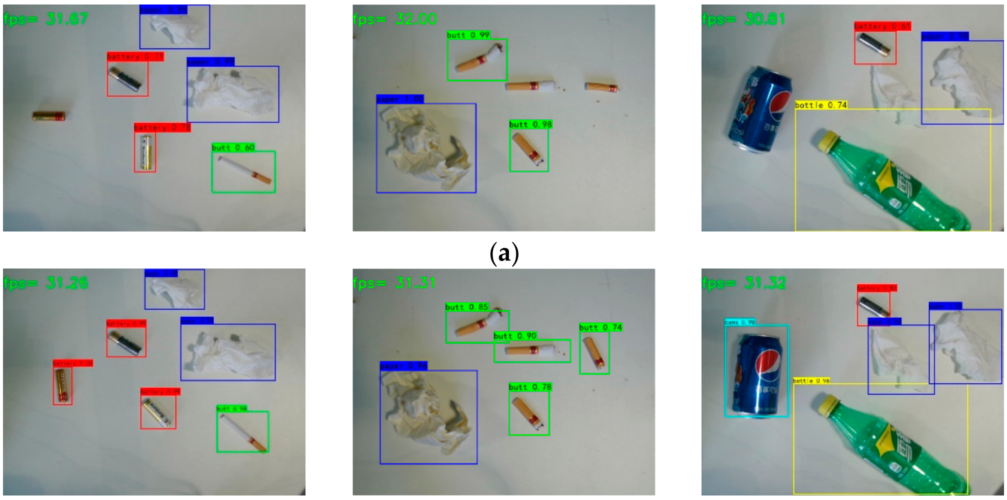


(**a**)

(**b**)

Figure 11. Comparison of model performance before and after improvement.(**a**) Improved model map, (**b**) Comparison of various types of AP.

The detection results of the improved model and the original Yolov4 are shown in Figure 12. Figure 12a shows the detection effect of the original Yolov4. It can be observed that there was a serious omission in the identification of such targets as cigarette ends. Figure 12b is the detection effect of the improved Yolov4 model. We noted the improved network model accurately identified all cigarette heads and paper scraps without missing a detection, and the FPS reached about 31. It can be seen from the above results that the detection accuracy and recall rate of the improved Yolov4 for the target were better than those of the original Yolov4 model, and the detection efficiency for small targets was higher and more practical.



(**a**)

(**b**)

Figure 12. Comparison of target detection results. (**a**) Original Yolov4 test results. (**b**) Improved Yolov4 detection results.

We compared the original model with the improved Yolov4 model in this paper. (1) In terms of model size, EfficientNet-Yolov4 reduced the parameters by about 80% compared with Yolov4 without sacrificing the detection accuracy. (2) In terms of detection accuracy and speed, compared with the original Yolov4 model, the mAP was increased by 5.77% and the detection speed of a single image was increased by 7 milliseconds. The improved model is more suitable for this experimental task.

5. Conclusions

Based on the Yolov4 algorithm, the lightweight Yolov4 recognition algorithm proposed in this paper combined with the ECA mechanism has a certain practicability. The main extraction network and feature pyramid were optimized and improved. The lightweight EfficientNet was used as the main feature extraction network. Based on the basically unchanged accuracy, the model was lighter, and the portability of the network was improved. At the same time, in the feature pyramid structure PANet network, the lightweight ECA mechanism was introduced to further improve the feature extraction ability of the network. Through cross-channel interaction, the weight analysis of the importance of different channel feature maps was realized, allowing the network to extract more obvious features with which to distinguish categories. Finally, the Soft-NMS algorithm was introduced in the post-processing stage of the detection box to reduce the missed target detection rate in dense regions and improve the detection performance. Compared with the original Yolov4 algorithm, the size of the improved Yolov4 model was reduced by 80%, and the mAP value was increased by 5.77%. We determined that the improvement of the Yolov4 detection algorithm based on EfficientNet has a high practicability and can be successfully applied to garbage sorting tasks. The optimized YoloV4 model in this paper can obtain the position and attitude information of garbage, but the detection effect will be affected by light, so the follow-up research focus is to improve the calculation accuracy of the pose and further improve the algorithm accuracy.

Author Contributions: Conceptualization, Y.Z.; investigation, C.W.; methodology, J.L.; writing—original draft, C.W. All authors will be informed about each step of manuscript processing including submission, revision, revision reminder, etc. via emails from our system or assigned assistant editor. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Tianjin special fund for technological innovation guidance grant number 21YDTPJC00260.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data in this paper are available from the corresponding authors upon request.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Zhang, J.; Zhang, Z.; Zhang, J.; Fan, G.; Wu, D. A Quantitative Study on the Benefit of Various Waste Classifications. *Adv. Civ. Eng.* **2021**, *2021*, 6660927. [CrossRef]
2. Malta, V.J.; Viegas, S.; Sabino, R.; Viegas, C. Fungal and Microbial Volatile Organic Compounds Exposure Assessment in a Waste Sorting Plant. *J. Toxicol. Environ. Health Part A* **2012**, *75*, 22–23.
3. Low, D.G. Distinctive image features from scale-invariant keypoint. *Int. J. Comput. Vis.* **2001**, *60*, 91–110. [CrossRef]
4. Bay, H.; Ess, A.; Tuytelaars, T.; Van Gool, L. SURF: Speeded-up robust features. *Comput. Vis. Image Underst.* **2008**, *110*, 346–359. [CrossRef]
5. Divya, S.V.; Sourabh, P.; Umesh, C.P. Structure tensor-based SIFT algorithm for SAR image registration. *IET Image Process.* **2020**, *14*, 929–938.
6. Alexey, B.; Wang, C.Y.; Liao, H.Y. YOLOv4: Optimal speed and accuracy of object detection. *arXiv* **2020**.
7. Tang, S.N.; Zhu, Y.; Yuan, S.Q. Intelligent Fault Identification of Hydraulic Pump Using Deep Adaptive Normalized CNN and Synchrosqueezed Wavelet Transform. *Reliab. Eng. Syst. Saf.* **2022**, *224*, 108560. [CrossRef]
8. Tang, S.N.; Zhu, Y.; Yuan, S.Q. Intelligent Fault Diagnosis of Hydraulic Piston Pump Based on Deep Learning and Bayesian Optimization. *ISA Trans.* **2022**. [CrossRef]
9. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv* **2014**, arXiv:1409.1556.
10. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going Deeper with Convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014.
11. Huang, G.; Liu, Z.; van der Maaten, L.; Weinberger, K.Q. Densely Connected Convolutional Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016.
12. Howard, A.G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; Adam, H. MobileNets: Efficient convolutional neural networks for mobile vision applications, Computer Vision and Pattern Recognition. *arXiv* **2017**, arXiv:1704.04861.
13. Girshick, R.B. Fast R-CNN. In Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, 7–13 December 2015; pp. 1440–1448.
14. Shaoqing, R.; Kaiming, H.; 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.
15. Dong, L.J.; Yi, W.X.; Yi, C.; Peng, Y.H. Structure Optimization of Convolutional Neural Networks: A Survey. *Acta Autom. Sin.* **2020**, *46*, 24–37.
16. Shao, W.P.; Wang, X.; Cao, Z.R.; Bai, F. Design of lightweight convolutional neural network based on MobileNet and YOLOv3. *J. Comput. Appl.* **2020**, *40*, 8–13.
17. Tang, S.N.; Zhu, Y.; Yuan, S.Q. A Novel Adaptive Convolutional Neural Network for Fault Diagnosis of Hydraulic Piston Pump with Acoustic Images. *Adv. Eng. Inform.* **2022**, *52*, 101554. [CrossRef]
18. Bodla, N.; Singh, B.; Chellappa, R.; Davis, L.S. Soft-NMS improving object detection with one line of code. In Proceedings of the 2017 IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 5562–5570.
19. Redmon, J.; Farhadi, A. Yolov3: An incremental improvement. *arXiv* **2018**, arXiv:1804.02767.
20. Wang, C.Y.; Liao, H.Y.M.; Wu, Y.H.; Chen, P.Y.; Hsieh, J.W.; Yeh, I.H. CSPNet: A new backbone that can enhance learning capability of CNN. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops, Seattle, WA, USA, 14–19 June 2020; pp. 390–391.
21. 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]

22. Tang, S.N.; Zhu, Y.; Yuan, S.Q. An adaptive deep learning model towards fault diagnosis of hydraulic piston pump using pressure signal. *Eng. Fail. Anal.* **2022**, *138*, 106300. [CrossRef]
23. Mao, L.; Ren, F.Z.; Yang, D.W.; Zhang, R.B. Two-way feature pyramid network for panoptic segmentation. *J. Jilin Univ.* **2022**, *52*, 657–665.
24. Zhu, Y.; Li, G.; Tang, S.; Wang, R.; Su, H.; Wang, C. Acoustic Signal-based Fault Detection of Hydraulic Piston Pump using a Particle Swarm Optimization Enhancement CNN. *Appl. Acoust.* **2022**, *192*, 108718. [CrossRef]
25. Tan, M.; Le, Q.V. EfficientNet: Rethinking model scaling for convolutional neural networks. *arXiv* **2019**, arXiv:1905.11946.
26. Howard, A.; Sandler, M.; Chu, G.; Chen, L.; Chen, B.; Tan, M.; Wang, W.; Zhu, Y.; Pang, R.; Vasudevan, V.; et al. Searching for MobileNetV3. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Korea, 27 October–2 November 2020.
27. Li, X.; Qin, Y.; Wang, F.; Guo, F.; Yeow, J.T. Pitaya detection in orchards using the MobileNet-YOLO model. In Proceedings of the 2020 39th Chinese Control Conference (CCC), Shenyang, China, 27–30 June 2020; pp. 6274–6278.
28. Teng, Y.D.; Gao, P.X. Generative Robotic Grasping Using Depthwise Separable Convolution. *Comput. Electr. Eng.* **2021**, *94*, 107318. [CrossRef]
29. Liu, C.A.; Feng, X.L.; Sun, C.H.; Zhao, L.J. Maximum 2-D entropy image segmentation method based on improved sparrow algorithm. *Laster Technol.* **2022**, *46*, 274.
30. Wei, F.; Wang, L.; Ren, P.M. Tinier-YOLO: A real-time object detection method for constrained environments. *IEEE Access* **2019**, *8*, 1935–1944.
31. Tan, M.X.; Pang, R.M.; Le, Q.V. Efficientdet: Scalable and efficient object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Piscataway, NJ, USA, 19–24 June 2020; pp. 10781–10790.
32. Lyn, J.; Yan, S. Image super-resolution reconstruction based on attention mechanism and feature fusion. *arXiv* **2020**, arXiv:2004.03939.
33. Hu, J.; Shen, L.; Sun, G. Squeeze-and-Excitation Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141.
34. Hou, Q.; Zhou, D.; Feng, J. Coordinate attention for efficient mobile network design. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 13713–13722.
35. Wang, Q.; Wu, B.; Zhu, P.; Li, P.; Zuo, W.; Hu, Q. ECA-Net: Efficient Channel Attention for Deep Convolutional Neural Networks. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020.
36. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-CAM: Visual explanations from deep networks via gradient-based localization. *Int. J. Comput. Vis.* **2020**, *128*, 336–359. [CrossRef]
37. 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.
38. Wu, L.; Ma, J.; Zhao, Y.; Liu, H. Apple Detection in Complex Scene Using the Improved YOLOv4 Model. *Agronomy* **2021**, *11*, 476. [CrossRef]