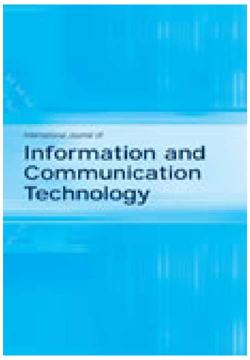


International Journal of Information and Communication Technology

ISSN online: 1741-8070 - ISSN print: 1466-6642
https://www.inderscience.com/ijict

Apple surface defect detection based on lightweight improved YOLOv5s

Lijun Lv, Yaermaimaiti Yilihamu, Yalin Ye

Article History:

Received:	04 March 2024
Last revised:	07 May 2024
Accepted:	07 May 2024
Published online:	13 June 2024

Apple surface defect detection based on lightweight improved YOLOv5s

Lijun Lv, Yaermaimaiti Yilihamu* and Yalin Ye

School of Electrical Engineering,
Xinjiang University,
Urumqi, 830017, China
Email: lv_lijun@foxmail.com
Email: xjgxy2001@sina.com
Email: 1764635209@qq.com
*Corresponding author

Abstract: Aiming at the current Apple surface defect detection algorithms with a large quantity of parameters, and poor real-time detection, a defect detection model to improve YOLOv5s is proposed. With the YOLOv5s model serving as a foundation, the EfficientNetv2 structure takes the role of the YOLOv5s model's backbone network. Second, by including the EMA attention mechanism in the neck component, the model's ability to extract important characteristics can be improved. Finally, using Alpha-IoU to optimise the IoU loss function can successfully raise the precision of the prediction box. The experimental findings demonstrate that the model size of the improved YOLOv5s model in this paper have been reduced by 20%, the recognition speed has been increased by 39.3%, and the mAP has been improved by 1.4%. In contrast to the initial model, the improved model has a smaller model size and a faster detection speed, while guaranteeing the detection accuracy.

Keywords: YOLOv5s; EfficientNetv2; EMA; defect detection.

Reference to this paper should be made as follows: Lv, L., Yilihamu, Y. and Ye, Y. (2024) 'Apple surface defect detection based on lightweight improved YOLOv5s', *Int. J. Information and Communication Technology*, Vol. 24, No. 7, pp.113–128.

Biographical notes: Lijun Lv is a graduate student at Xinjiang University. His main research areas are object detection and image processing.

Yaermaimaiti Yilihamu is a Graduate Advisor and Professor. His main research areas are artificial intelligence, pattern recognition, face recognition, target tracking and detection.

Yalin Ye is a graduate student at Xinjiang University. His main research areas are artificial intelligence and image processing.

1 Introduction

With the growth of the social economy and the ongoing raising of living standards, customers are becoming more picky about fruit quality. Apples have high quality in terms of taste and nutritional value, but various defects may occur in the process of growth, picking, sorting and transport. These defects will not only affect the appearance and taste of apples, but may also have a serious impact on consumers' health. Therefore, effective surface defect detection of apples is the key to ensure the quality of apples.

The conventional approach of identifying surface flaws in apples mostly depends on manual visual examination, which is labour-intensive, time-consuming, and subject to subjectivity. It also finds it challenging to fulfil the demands of the market and mass production. With the advancement of computer vision technology in recent years, it is now possible to detect flaws on the surface of fruits using image processing technology. By collecting the image information of the fruit surface and using deep learning algorithms to analyse and process it, the rapid identification and classification of fruit surface defects can be achieved.

Krizhevsky et al. proposed an approach for image classification using CNN networks, which uses convolutional and pooling layers to extract features of an image, and then maps these features to different classes through a fully connected layer. This method performed well in the image classification task and won the ImageNet competition in 2012; Redmon et al. proposed YOLO algorithm in 2016. Compared with the two-stage detection algorithm, it utilises direct regression, which greatly reduces the amount of computation and improves the operation speed. However, there are only 2 prediction frames per grid, which is not effective when detecting dense objects or small target objects; Xing et al. (2020) proposed a lightweight YOLOv3 network-based apple detection method for fast and accurate detection of apples in complex backgrounds for an plucking apple robot. According to the experimental results, the Light-YOLOv3 network-based apple detection method greatly increases both the speed and accuracy of detection; on computer workstations and embedded processors, detection speeds can reach 116.96FPS and 7.59FPS, respectively, with an average accuracy of 94.69%. An apple defect detection algorithm based on the FCM and NPGA combined with multivariate image analysis was suggested by Zhang et al. (2020) to address the issue that the surface defects of apples are small and the features of the defective region are difficult to extract. The overall detection accuracy of the method is 98%. Liang et al. (2022) graded defective apples in real-time using a semantic segmentation combined and pruned YOLOv4 network with an average accuracy of 92.42%. It has great potential for application in commercial grading and sorting machines. For the real-time identification of apple leaf diseases, Jiang et al. (2019) suggested a deep learning technique based on enhanced convolutional neural networks. By introducing the GoogLeNet Inception structure and Rainbow crosstalk, a new deep CNN-based apple leaf disease detection model was proposed. According to the testing results, this revised algorithm can achieve a mAP of 78.80%. Wu et al. (2021) suggested an improved apple identification approach for complex scenes for the YOLOv4 model, which reduces model size by replacing the YOLOv4 model's backbone network, Cross Stage Partial Darknet53, with EfficientNet. The final average values of Recall, Precision and F1 reached 97.43%, 95.52% and 96.54%, respectively. Using YOLOv4 as the foundation and MobileNetv3 as the feature extraction network, Wang and colleagues (2021) suggested a lightweight real-time apple detection approach. By adding depth separable convolution to the feature fusion network, the computational complexity of the network is decreased and the improved model's detection speed is significantly increased; Tu et al. (2021) suggested modified Faster R-CNN algorithm for passion fruit target detection, using ResNet (He et al., 2016) network combined feature pyramid network (FPN) (Lin et al., 2017) for multi-scale feature extraction of passion fruit, to achieve the automatic detection of four types of scenes: unobstructed, occluded, overlapped and background, and the average accuracy rate of detection under four types of situations reached 87.98%; For the purpose of detecting apple leaf diseases in real time, Li et al. (2022) introduced the Apple-YOLO lightweight disease detection model. With a mAP of 96.04%, and a size of just 5.33M, the dual-branch Apple-CSP module and the upgraded FDSA module among them effectively reduce the quantity of model parameters.

Despite the gradual increase in the use of deep learning algorithms for defect detection in apples as well as round fruits and vegetables, the attention they have received is still limited overall. In addition, research in this field started late in China. As a large agricultural country, in-depth research in this field is not only of great practical significance, but also of great economic value. It can not only help the majority of farmers to reduce costs and increase income, but also provide diversified solutions for assembly line inspection in fruit factories. At the same time, it is also in line with our current national situation and helps to promote our rapid progress towards a modernised agricultural country.

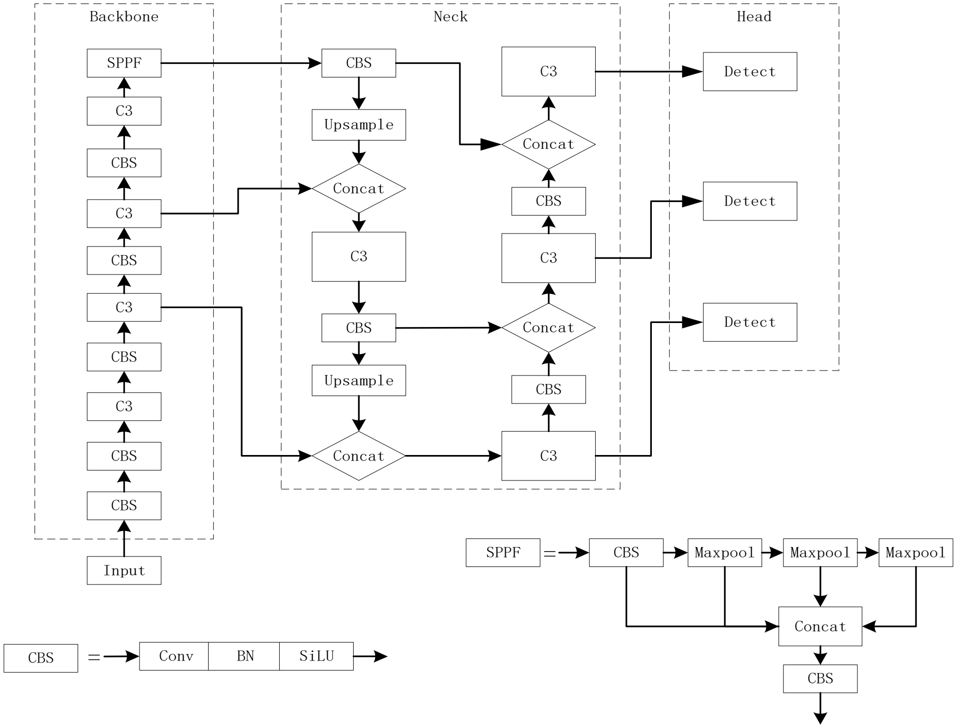
Previous apple surface defect detection algorithms had issues with a lot of parameters, poor real-time detection, and not easy to be deployed on the mobile side. In view of this, the algorithm proposed in this article focuses on solving these problems. We use the EfficientNetV2 network to replace the backbone network of the YOLOv5s model and incorporate the EMA attention mechanism, and then optimise the bounding box regression loss function using Alpha-IoU. The enhanced model increases recognition accuracy and speed while simultaneously decreasing the number of parameters.

2 YOLOv5 network profile

YOLOv5 is a single-stage target detection algorithm suggested by Glenn Jocher in 2020. The features of YOLOv5 include great accuracy and quickness. The Input, Neck, Head, and Backbone networks make up the four components of YOLOv5. Figure 1 depicts its network architecture.

YOLOv5 uses Mosaic data improvement on the input aspect. Mosaic data improvement randomly uses four pictures that are spliced in a random expanding, random cutting, and random organising way, which greatly expands the target detection dataset, especially the random expanding, which adds a lot of tiny goals, which significantly improves the network's robustness. The model can directly compute data from four images while training with Mosaic augmentation, significantly lowering the GPU workload and consequently GPU consumption. Mosaic data augmentation crops the target object, which allows the model to recognise objects based on local features and enhances the precision of obscured object recognition; The Backbone part is mainly responsible for extracting image features, which is the core part of the whole network. Backbone adopts the structure of CSPDarknet53, which is an efficient network structure that allows features to be fused by connecting them across stages, drastically reducing the quantity of parameters, while maintaining a high accuracy rate. One of the C3 modules can expand the depth and sensory field of the algorithm to boost the feature extraction capability, thus improving the accuracy of target detection; Path aggregation networks (PANs) and FPNs make up the majority of the neck section. FPN is a top-down structure that passes the top feature information down through the backbone network to the bottom layer; PAN introduces a bottom-up structure based on FPN, which enables the location information of the bottom layer to be passed to the top layer as well through bottom-up feature fusion, enhancing the localisation capability on multiple scales. The neck part uses a combination of PAN and FPN, which uses the form of pyramid to connect feature maps with different scales, and fuses high-level features with low-level features, which can effectively enrich feature information; the head section performs the final regression prediction.

Figure 1 YOLOv5 network structure diagram



In contrast to other target detection algorithms, such as R-CNN, SSD and Faster R-CNN, the advantages of YOLOv5 are in some aspects:

1. Fast. YOLOv5 uses a single neural network to simultaneously predict different objects and their locations in multiple categories in the image, and the whole process can be completed with only one forward propagation, which greatly reduces the computational complexity compared with other target detection algorithms that require multiple candidate frame generation, classification and location regression operations.

2. High accuracy. While maintaining speed, YOLOv5 increases accuracy while avoiding coupling between multiple modules and error accumulation. It can also directly output the object's class and location data without the need for SVM, bounding boxes regression, or other target detection techniques.

3 Strong applicability. YOLOv5 can be applied to target detection of objects of various sizes and shapes, and is not limited by the number of objects, so it can be a good solution to the common problem of small target detection. Consequently, YOLOv5 has received a lot of attention and has been used in real-world scenarios.

3 Improved YOLOv5s network model

3.1 Introduction of the lightweight module EfficientNetV2

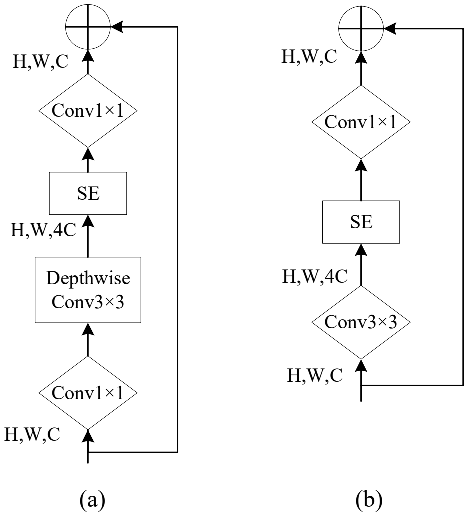
The C3 structure adopted by the YOLOv5 backbone feature extraction network results in a greater number of parameters, slower detection performance, and a smaller range of applications. Firstly, because the model is too large, it will face the issue of insufficient memory or require higher performance hardware equipment. Secondly, many scenarios require fast response speed and low latency, such as real-time monitoring systems or real-time factory assembly line detection systems, which put forward higher requirements for speed. So we have to pursue a defect detection model that is fast, has a small number of parameters, and can guarantee a certain degree of accuracy. In this research, we attempt to further increase velocity and precision of defect identification by substituting a lighter EfficientNetV2 (Tan and Le, 2021) network for the backbone feature extraction network, resulting in a lightweight network model.

Compared with EfficientNetV1 (Tan and Le, 2019), more attention is paid to model training speed in EfficientNetV2. EfficientNetV2 introduces the Fused_MBConv module, which can make better use of the accelerator on the mobile side or server side. And EfficientNetV2 introduces a progressive learning strategy, which can dynamically adjust the regularisation methods (dropout, rand augment, Mixup) according to the size of the training image, which could improve the training speed and increase the accuracy by a small margin. Fused_MBConv compared with MBConv is to replace the 1×1 convolutional kernel dw convolution with a 3×3 convolution. Depth and width are scaled equally in each stage of EfficientNetV1, but scaling each stage equally is suboptimal. Each stage does not contribute equally to the training speed of the network and to the quantity of parameters, so it is not reasonable to use equal scaling directly. In contrast, EfficientNetV2 scales the model using a more sensible non-uniform scaling technique.

Table 1 EfficientNetV2 network structure

Stage	*Module*	*Stride*	*Expand ratio*	*Channel*	*Layers*
0	Conv 3×3	2	1	24	1
1	Fused-MBConv	1	1	24	3
2	Fused-MBConv	2	4	48	5
3	Fused-MBConv	2	4	80	5
4	MBConv	2	4	160	7
5	MBConv	1	6	176	14
6	MBConv	2	6	304	18
7	MBConv	1	6	512	5

Figure 2 MBConv and fused-MBConv structural diagram, (a) MB Conv (b) fused-MB conv



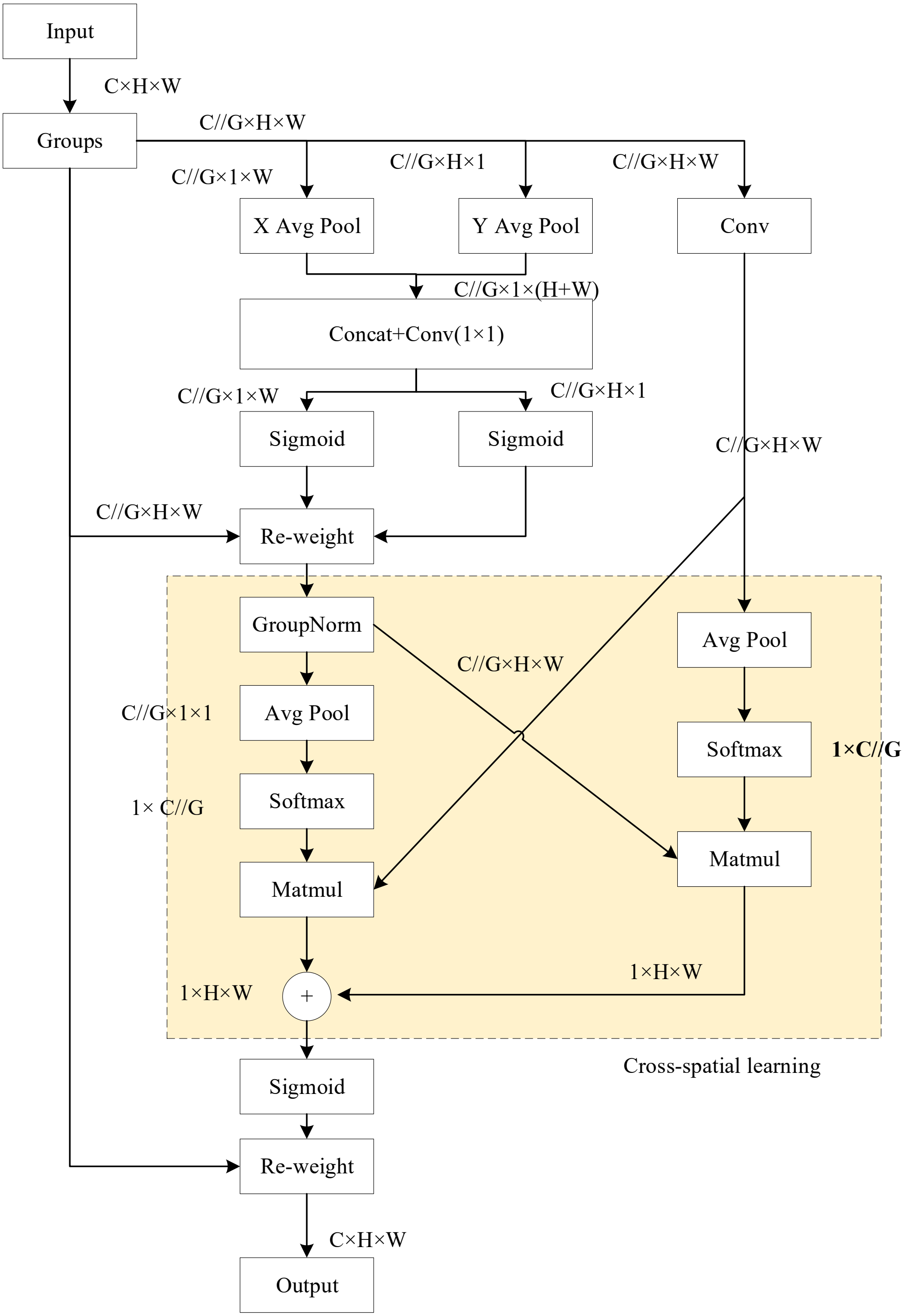
3.2 Add EMA attention mechanism

In a variety of computer vision tasks, channel or spatial attention mechanisms have achieved significant effectiveness in producing more discriminative feature representations. However, modelling cross-channel relationships through channel dimensionality reduction may have side effects on extracting deep visual representations. The CA (Hou et al., 2021) attention mechanism, for example, has some improvement in accuracy but requires additional computation leading to resource consumption as it needs to compute the attention weights for the entire feature map, in addition to failing to capture long distance dependencies between channels.

June 2023 saw the proposal of a new, effective multi-scale attention EMA (Ouyang et al., 2023) module by Aerospace Science and Technology in Shenzhen. To prevent dimensionality reduction through generalised convolution, portions of the channel's dimensions are first reshaped into bulk dimensions; then local cross-channel interactions are constructed in each parallel subnetwork, and the resultant map attributes of the two concurrent subnetworks are fused using a novel cross-spatial learning technique, and to design a multi-scale parallel sub-network to establish short and long dependencies. The network structure diagram of the EMA attention mechanism is displayed in Figure 3.

The EMA attention mechanism can focus the limited attention on the key features we want to pay attention to, which can effectively save computational resources and quickly locate the features we pay most attention to. To achieve better outcomes, we introduce the EMA attention mechanism in this study after the small target layer, medium target layer, and large target layer in the Neck part of YOLOv5.

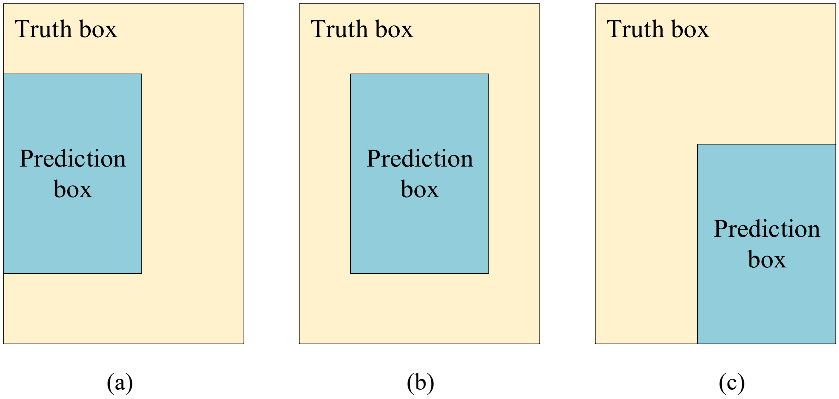
Figure 3 EMA attention mechanism (see online version for colours)



3.3 Improvement of the bounding box loss function

In the original YOLOv5 network, the bounding box loss function uses GIoU (Rezatofighi et al., 2019). Due to GIoU's pursuit of the least outer rectangle for every truth box and prediction box, computation speed and convergence are constrained. If there is a situation where the prediction box is contained inside the truth box, such as Figure 4, then at this time, no matter where the prediction box is located inside the truth box, the value of GIoU is equal to the value of IoU, which is 0.45, which is obviously unreasonable. At this time, GIoU can not truly reflect the position change of the prediction box, so it can be seen that there are some shortcomings in GIoU.

Figure 4 The truth box contains the prediction box, (a) $L_{GIoU} = L_{IoU} = 0.45$ (b) $L_{GIoU} = L_{IoU} = 0.45$ (c) $L_{GIoU} = L_{IoU} = 0.45$ (see online version for colours)



In view of the shortcomings of the above GIoU in use, this paper changes to use Alpha-IoU (He et al., 2021) with more penalty terms, Alpha-IoU unifies the existing loss of IoU and expresses it in the form of a power exponent, and can satisfy different levels of bounding box regression precision by introducing adjustable hyperparameter α. Experiments show that Alpha-IoU works best when α = 3 is taken, and the loss function of Alpha-IoU can effectively improve the recall and detection precision.

Equation (1) defines the Alpha-IoU loss function, and most of the IoUs in the existing losses can be derived by adjusting the hyperparameter alpha.

$$\alpha - IoU = \frac{1 - IoU^{\alpha}}{\alpha}, \alpha > 0 \tag{1}$$

As in equation (2), by including a penalty term, the aforementioned Alpha-IoU loss function can be expanded to a more generic form, and the current IoU-based losses based on the alpha value can be made more widely applicable.

$$Loss_{\alpha - IoU} = 1 - IoU^{\alpha_1} + \rho^{\alpha_2}\left(B, B^{gt}\right) \tag{2}$$

Based on the above general form of Alpha-IoU can be generalised to the commonly used IoU loss, which is adopted in this paper as shown in equation (3); where the consistency of the ratio of height to width is indicated by the letter v and the letter *c* stands for the distance from the diagonally of the least enclosed region that may contain simultaneously the truth box and the prediction box. To balance v, β is the trade-off parameter, and ρ is the Euclidean distance between the truth box and prediction box. $\frac{\omega^{\mathrm{gt}}}{h^{\mathrm{gt}}}$ and $\frac{\omega}{h}$ are the aspect ratios of the truth box and prediction box, respectively. The specific expressions are shown in equations (4) and (5).

$$Loss_{\alpha-\mathrm{CIoU}} = 1 - IoU^{\alpha} + \frac{\rho^{2\alpha}\left(B, B^{\mathrm{gt}}\right)}{c^{2\alpha}} + (\beta v)^{\alpha} \tag{3}$$

$$\beta = \frac{v}{(1 - IoU) + v} \tag{4}$$

$$v = \frac{4}{\pi^2}\left(\arctan\frac{\omega^{\mathrm{gt}}}{h^{\mathrm{gt}}} - \arctan\frac{\omega}{h}\right)^2 \tag{5}$$

4 Experiment and analysis of results

4.1 Dataset

The dataset for this research was taken in an apple orchard in Yuncheng City, Shanxi Province, with apple varieties such as Red Fuji, Yellow Marshal, Red Star, Guo guang, Gala, etc. It contains five types of detection, including good apple, wrinkled apple, rot, mechanical damage, and wormhole, with 1,000 sheets of each type and a total of 5,000 sheets. The five detection types are shown in Figure 5 below. In this article, five methods of data enhancement, namely Gaussian noise, salt and pepper noise, brightening, darkening and low pixel, are used to augment the dataset as shown in Figure 6. This can effectively simulate various unfavourable environments encountered during the shooting process, thus enhancing the anti-interference capability of the algorithm. The data enhancement resulted in a total dataset of 30,000 sheets. According to the ratio of 8:1:1 randomly divided into training set, validation set and test set.

Figure 5 Five types of detection, (a) good apple (b) wrinkled apple (c) rot (d) mechanical damage (e) wormhole (see online version for colours)

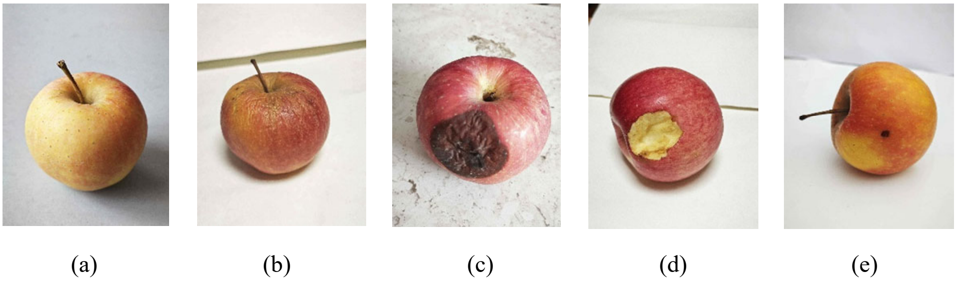
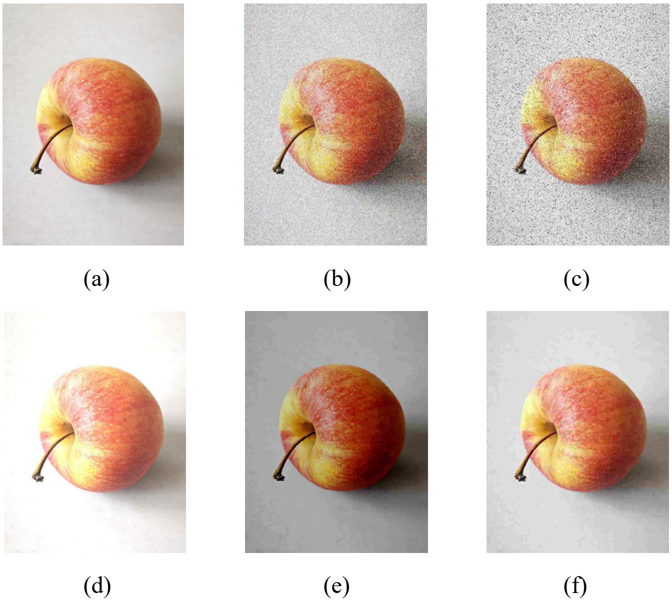


Figure 6 Five data enhancement methods, (a) original image (b) Gaussian noise (c) salt and pepper noise (d) brightening (e) darkening (f) low pixel (see online version for colours)



4.2 Experimental environment

In this study, the Ubuntu operating system is used in an experimental environment built around the AutoDL cloud server. PyTorch 1.11.0 framework, CUDA version 11.3, Python version 3.8, GPU is RTXA5000 with 24GB of video memory, CPU is AMD EPYC 7543 with 30GB of RAM.

4.3 Training parameter setting

Table 2 Training parameters

Parameter	*Parameter value*
Epoch	150
Weight decay	0.0005
Momentum	0.937
Learn rate	0.01
Batch size	64

4.4 Indicators for model assessment

The assessment metrics for the algorithm selected for this work are mAP, Recall, and Precision. Recall (R) indicates the proportion of positive category identified as correct to the amount of actual positive category, and Precision (P) indicates the ratio of positive category identified as correct to the amount of positive category identified as positive as in the following equations (6) and (7). N_{FN} indicates the amount of samples labelled as positive category identified as negative s category; N_{TP} indicates the amount of positive category correctly identified as positive category; and N_{FP} indicates the amount of negative category incorrectly identified as samples in the positive category.

$$P = \frac{N_{\mathrm{TP}}}{N_{\mathrm{TP}} + N_{\mathrm{FP}}} \tag{6}$$

$$R = \frac{N_{\mathrm{TP}}}{N_{\mathrm{TP}} + N_{\mathrm{FN}}} \tag{7}$$

The mAP is calculated as follows:

$$AP = \int_0^1 P(R)d(R) = \sum_{k=1}^{N} P(k)\Delta R(k) \tag{8}$$

$$mAP = \frac{\sum_{i=1}^{n} AP(i)}{n} \tag{9}$$

4.5 Experimental results and analysis of the improved YOLOv5s

4.5.1 Comparative experiment on lightweight modules

We compared the EfficientNetV2 module with a number of widely-used lightweighting modules in order to confirm its efficacy. According to Table 3, after the lightweighting module is added, the model dimensions and quantity of parameters decreased to varying degrees, but the mAPs are all decreased as well, with ShuffleNetV2 (Ma et al., 2018) having the largest reduction in mAP, which decreased by 10.8%. The mAP of the algorithm after adding the MobileNetV3 (Howard et al., 2019) module is 88.5%, while the algorithm after adding the GhostNetV2 (Tang et al., 2022) module is 87.9%. The mAP of EfficientNetV2 is reduced the least, and is the closest to that of the original model, and the quantity of parameters and the model size are reduced by 20%. Taken together, we chose the EfficientNetV2 module in order to balance lightweight and accuracy.

Table 3 Comparison of lightweight modules

Model	*mAP_0.5/%*	*P/%*	*R/%*	*Parameters/M*	*Model size/MB*	*GFLOPs*
YOLOv5s	92.2	93.2	87.9	7.0	14.4	15.8
YOLOv5s+EfficientNetV2	90.4	89.2	86.4	5.6	11.5	5.6
YOLOv5s+ShuffleNetV2	81.4	80.6	83.3	0.8	2.0	1.8
YOLOv5s+MobileNetV3	88.5	85.9	84.7	1.4	3.1	2.3
YOLOv5s+GhostNetV2	87.9	86.1	82.8	2.8	6.2	4.5

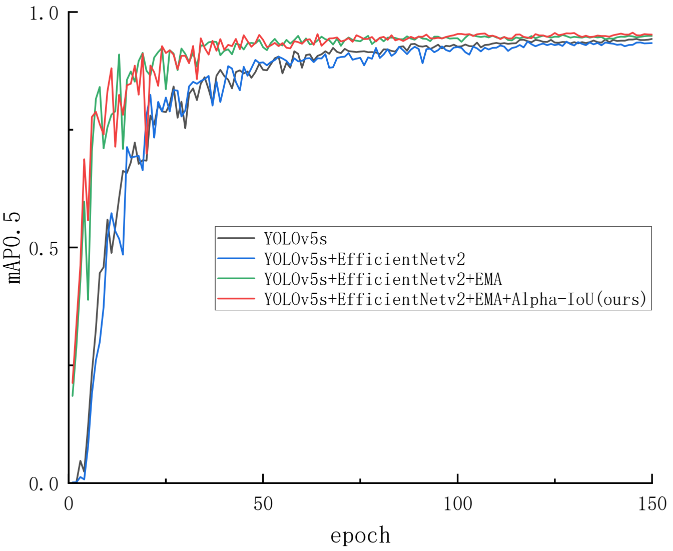
Table 4 Ablation experiments

Model	*mAP_0.5/%*	*P/%*	*R/%*	*Parameters/M*	*Model size/MB*	*FPS*	*GFLOPs*
YOLOv5s	92.2	93.2	87.9	7.0	14.4	62.6	15.8
YOLOv5s+EfficientNetv2	90.4	89.2	86.4	5.6	11.5	86.4	5.6
YOLOv5s+EfficientNetv2+EMA	92.9	92.3	92.4	5.6	11.5	86.9	5.6
YOLOv5s+EfficientNetv2+EMA+Alpha-IoU	93.6	93.4	93.7	5.6	11.5	87.2	5.6

4.5.2 Ablation experiment

Ablation experiments are carried out to confirm the outcome of the enhanced method suggested in this article on the YOLOv5s method. The outcomes are displayed below. It can be seen that with the addition of EfficientNetV2, the model size and the quantity of parameters are downsized by 20%, the FPS is improved, and there is an improvement in lightness as well as in detection speed. The mAP increases by 2.5%, the recall increases by 6%, and the precision increases by 3.1% with the addition of the EMA attention mechanism once more. These improvements are impressive. After adding Alpha-IoU again, mAP is further improved by 0.7%. The experimental findings demonstrate that the model size and the quantity of parameters of the enhanced YOLOv5s in this article have been reduced by 20%, the GFLOPs have been reduced by 64.6%, the recognition speed has been improved by 39.3%, and the mAP of the enhanced YOLOv5s in this article has been improved by 1.4% compared with the initial model. The enhanced YOLOv5s model has a smaller number of parameters, faster detection speed, and higher detection accuracy.

Figure 7 Results of ablation experiments with mAP0.5 (see online version for colours)



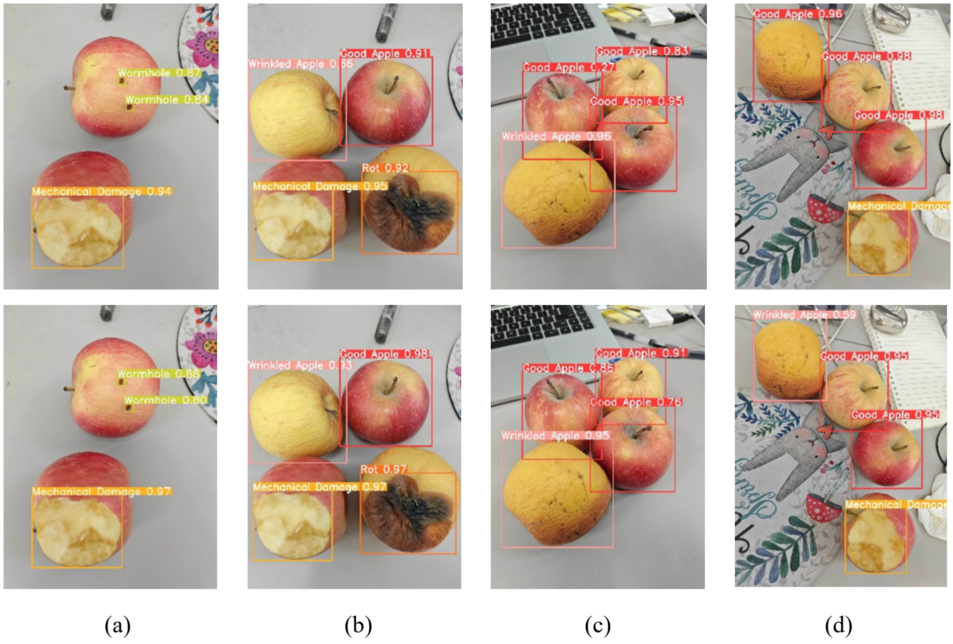
4.5.3 Comparative experiments of target detection algorithms

For the purpose of evaluate the effectiveness of the enhanced YOLOv5s in this article, it is compared with some of the current commonly used target detection algorithms, and the outcomes are displayed in Table 5. YOLOv5n has a smaller number of parameters as well as faster detection, but the mAP is only 89.9%, which is not enough to meet the actual demand. The mAP of YOLOv5m reaches 93.2%, but the model size is 3.66 times larger than the enhanced YOLOv5s, and the detection speed is only 37.7% of the algorithm in this paper, which does not satisfy the demand of real-time detection. The SSD algorithm has a relatively large gap with other algorithms in terms of detection accuracy, detection speed and quantity of parameters. YOLOv7-tiny detection speed is only 52.2% of the enhanced YOLOv5s, and its detection speed is relatively slow. Taken together, the enhanced YOLOv5s has smaller parameters and model size, as well as faster detection speed under the premise of guaranteeing detection accuracy, and is able to better perform the detection of apple surface defects.

Table 5 Comparison of object detection algorithms

Model	*mAP_0.5/%*	*P/%*	*R/%*	*Parameters/M*	*Model size/MB*	*FPS*
YOLOv5s	92.2	93.2	87.9	7.0	14.4	62.6
YOLOv5n	89.9	86.6	84.6	1.8	3.9	147.4
YOLOv5m	93.2	94.9	89.4	20.9	42.2	23.6
SSD	88.3	82.7	85.8	26.2	91.5	30.2
YOLOv7-tiny	92.4	93.6	88.0	6.2	12.3	32.7
YOLOv5s+Alpha-IoU+ EfficientNetV2+EMA(ours)	93.6	93.4	93.7	5.6	11.5	87.2

Figure 8 Comparison of test results, (a) small target (b) multiple targets (c) overlap (d) complex background (see online version for colours)



4.6 Comparison of real-life applications

For the purpose of further evaluate the effectiveness of the enhanced YOLOv5s in this article, we select various types of images to conduct experiments respectively, and Figure 8 displays the outcomes of the experiment. We use the initial YOLOv5s model and the enhanced YOLOv5s model to test and compare, the first line of images are the detection results of the YOLOv5s source code and the second line of images are the detection results of the enhanced YOLOv5s. Observation shows that both YOLOv5s source code and this paper's algorithm are good at detecting small-target defects, multi-target defects, and apple overlap cases, but overall this paper's algorithm has a bit higher confidence. In complex background situations, YOLOv5s source code incorrectly recognises the wrinkled apple in the upper left corner as a healthy apple, and the algorithm in this article detects the wrinkled apple well. In summary, the algorithm presented in this article achieves a lower quantity of parameters and a faster detection time while maintaining detection accuracy.

5 Conclusions

The apple surface defect detection algorithm now in use has a lot of parameters, has poor real-time detection, and is difficult to implement on the mobile side. To address these issues, we present an improved YOLOv5s-based approach in this work. For the purpose to reduce the quantity of parameters and increase the speed of model recognition, the EfficientNetV2 module is first employed in place of the YOLOv5s model's backbone network. Secondly, the EMA attention mechanism has been incorporated into the neck section to enhance the model's capability in extracting key features, thereby leading to an improvement in detection accuracy. Finally, in order to improve the predicted box's accuracy, the bounding box regression loss function is optimised using Alpha-IoU. The experimental findings demonstrate that the speed of this paper's algorithm is improved by 39.3%, the quantity of parameters and model size are decreased by 20%, the GFLOPs are reduced by 64.6%, and the mAP is increased by 1.4% over the source code of YOLOv5s. In comparison to the initial network model, the enhanced model offers quick recognition times, high detection precision, and little memory usage, which can better detect the apple surface defects and can detect more types of defects, which has good practical application value.

Acknowledgements

The authors want to acknowledge the financial support from the National Natural Science Foundation of China (Project No.:62362063.61866037).

References

He, J., Erfani, S., Ma, X. et al. (2021) 'Alpha-IoU: a family of power intersection over union losses for bounding box regression', *Advances in Neural Information Processing Systems*, Vol. 34, pp.20230–20242.

He, K., Zhang, X., Ren, S. et al. (2016) 'Deep residual learning for image recognition', *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp.770–778.

Hou, Q., Zhou, D. and Feng, J. (2021) 'Coordinate attention for efficient mobile network design', *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp.13713–13722.

Howard, A., Sandler, M., Chu, G. et al. (2019) 'Searching for mobilenetv3', *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp.1314–1324.

Jiang, P., Chen, Y., Liu, B. et al. (2019) 'Real-time detection of apple leaf diseases using deep learning approach based on improved convolutional neural networks', *IEEE Access*, Vol. 7, pp.59069–59080.

Krizhevsky, A., Sutskever, I. and Hinton, G.E. (2012) 'Imagenet classification with deep convolutional neural networks', *Advances in Neural Information Processing Systems*, Vol. 25, pp.1097–1105.

Li, J., Zhu, X., Jia, R. et al. (2022) 'Apple-YOLO: A novel mobile terminal detector based on YOLOv5 for early apple leaf diseases', *2022 IEEE 46th Annual Computers, Software, and Applications Conference (COMPSAC)*, IEEE, pp.352–361.

Liang, X., Jia, X., Huang, W. et al. (2022) 'Real-time grading of defect apples using semantic segmentation combination with a pruned YOLO V4 network', *Foods*, Vol. 11, No. 19, p.3150.

Lin, T.Y., Dollár, P., Girshick, R. et al. (2017) 'Feature pyramid networks for object detection', *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp.2117–2125.

Ma, N., Zhang, X., Zheng, H.T. et al. (2018) 'Shufflenet v2: Practical guidelines for efficient cnn architecture design', *Proceedings of the European Conference on Computer Vision (ECCV)*, pp.116–131.

Ouyang, D., He, S., Zhang, G. et al. (2023) 'Efficient multi-scale attention module with cross-spatial learning', *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, pp.1–5.

Redmon, J., Divvala, S., Girshick, R. et al. (2016) 'You only look once: unified, real-time object detection', *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp.779–788.

Rezatofighi, H., Tsoi, N., Gwak, J.Y. et al. (2019) 'Generalized intersection over union: a metric and a loss for bounding box regression', *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp.658–666.

Tan, M. and Le, Q. (2019) 'Efficientnet: rethinking model scaling for convolutional neural networks', *International Conference on Machine Learning*, PMLR, pp.6105–6114.

Tan, M. and Le, Q. (2021) 'Efficientnetv2: Smaller models and faster training', *International Conference on Machine Learning*, PMLR, pp.10096–10106.

Tang, Y., Han, K., Guo, J. et al. (2022) 'GhostNetv2: enhance cheap operation with long-range attention', *Advances in Neural Information Processing Systems*, Vol. 35, pp.9969–9982.

Tu, S., Huang, J., Lin, Y. et al. (2021) 'Automatic detection of passion fruit based on improved faster R-CNN', *Res. Explor. Lab.*, Vol. 40, No. 11, pp.32–37.

Wang, Z., Wang, J. and Wang, X.X. (2022) 'Lightweight real-time apple detection method based on improved YOLO v4', *Transactions of the Chinese Society for Agricultural Machinery*, Vol. 53, No. 8, pp.294–302.

Wu, L., Ma, J., Zhao, Y. et al. (2021) 'Apple detection in complex scene using the improved YOLOv4 model', *Agronomy*, Vol. 11, No. 3, p.476.

Xing, W., Zeyu, Q., Longjun, W. et al. (2020) 'Apple detection method based on light-YOLOV3 convolutional neural network', *Nongye Jixie Xuebao/Transactions of the Chinese Society of Agricultural Machinery*, Vol. 51, No. 8, pp.17–25.

Zhang, W., Hu, J., Zhou, G. et al. (2020) 'Detection of apple defects based on the FCM-NPGA and a multivariate image analysis', *IEEE Access*, Vol. 8, pp.38833–38845.