Engenharia Agrícola

ISSN: 1809-4430 (on-line)

www.engenhariaagricola.org.br

Scientific Paper
Doi: http://dx.doi.org/10.1590/1809-4430-Eng.Agric.v44e20230175/2024

LIGHTWEIGHT YOLOV5S-SUPER ALGORITHM FOR MULTI-DEFECT DETECTION IN APPLES

Jinan Yu[1], Rongchang Fu[1*]

[1*]Corresponding author. XinJiang University/Urumuqi, China.
E-mail: 2781642414@qq.com | ORCID ID: https://orcid.org/0000-0002-7045-7597

KEYWORDS

apple defect, Yolov5s-Super, lightweighting, EAM attention, Wise-IoU.

ABSTRACT

As the application scenarios of embedded devices become increasingly extensive, the use of high-performance convolutional neural networks can solve the problem of low accuracy of multiple defects detection in apples. However, owing to the overly large parameters and network structure of the convolutional neural network, perfectly integrating it with the embedded devices is difficult. Therefore, this study proposes a lightweight and improved algorithm based on Yolov5s. First, the structure of the optimized MobileNetV3 is introduced in the backbone layer to reduce the computational and parametric quantities of the model. Wise-IoU is used as the loss function of the localization regression of the bounding box to reduce the harm of low-quality samples on anchor box regression. The efficient multiscale attention mechanism is embedded in each downsampling layer of the backbone, and small target detection is added to the neck layer to improve the attention of the convolutional layer on important features. The experimental results showed that the Yolov5s-Super model parametric count decreased by 78%, and accuracy P, mAP@50, and mAP@50:95 improved by 10.3%, 3.2%, and 4.2%, respectively, compared to the original model. Theoretical support is provided for the migration of this network model to embedded devices.

INTRODUCTION

The yield of crops significantly affects the economic and social development (Wang et al., 2023). As the country with the largest apple production and the largest planting area in the world (Yu et al., 2022), China has a pivotal position in the agricultural field of apples in China. However, owing to the backwardness of the existing apple grading technology in China that leads to the uneven quality of apples, enterprises and farmers are unable to obtain higher sales profits (Wang et al., 2022). Real-time detection of the health status of apples can effectively control the large-scale proliferation of pathogens (Samajpati & Degadwala, 2015) and avoid incalculable losses. Most farm managers select apples manually, a monotonous and tedious process that is inefficient and expensive. In recent years, deep learning has also been widely used in apple defect detection (Kamilaris & Prenafeta, 2018). The use of convolutional neural networks to achieve intelligent, high-precision identification of apple surface defects is of great significance (Dong et al., 2024) to ensure the food safety of people and improve the income of enterprises and farmers.

With the development of machine vision technology, particularly the application of image processing and pattern recognition technology (Dong & Wang, 2023), the current vision algorithms based on deep learning are divided into two types: one-stage, where typical algorithms are the Yolo and SSD (Liu et al., 2016); and two-stage, where typical algorithms are R-CNN (Ross et al., 2014) and Faster R-CNN (Ren et al., 2015). Among these algorithms, the Yolo algorithm has the advantages of high detection accuracy and fast detection speed and is currently widely used in various fields. Among the Yolo series, Yolov5 is widely used in the field of target detection owing to its advantages of small model size, low deployment cost, high flexibility, fast detection speed, and high detection accuracy. However, the disadvantages of its oversized downsampling rate and the lack of feature fusion in the first-order algorithm of the SSD lead to poor detection of small targets (Hu et al., 2023).

[1] XinJiang University/Urumuqi, China.

Area Editor: Gizele Ingrid Gadotti
Received in: 5-30-2023
Accepted in: 11-15-2023

Scholars worldwide have conducted relevant research to address these problems. Mei et al. (2023) proposed a lightweight apple defect detection method based on Mo-M2Det that reduces the number of parameters; however, the value of its model loss function is high that is prone to leakage and misdetection. Hu et al. (2023) proposed an architecture based on the Yolov5l model for the detection of defects in apples, that increases the channel attention mechanism, and a small target detection layer to improve the accuracy; however, owing to the use of a large network structure, the number of parameters is large, the requirements on the device are high, and perfectly integrating it with embedded devices is difficult. Tian et al. (2019) proposed a Yolov3-based surface defect detection method for apples; based on the use of image data enhancement, the use of densely connected neural networks (DenseNet) for the lower resolution of the Yolov3 model feature layer was optimized. Valdez (2020) classified the detection of defects in apples as an object detection problem rather than a simple image classification problem and used the Yolov3 network for the detection that had a detection accuracy of only 69%; this must still be improved.

Because general detection and sorting devices are embedded devices, their storage capacity is small. However, convolutional neural networks have a large number of parameters and high computational complexity; hence, perfectly integrating them with embedded devices is difficult (Han et al., 2020). To improve the accuracy of the apple defect detection model and reduce the number of parameters, this study proposes a high-precision, lightweight apple surface multi-defect detection algorithm based on Yolov5s. First, the network of MobileNetv3 is modified to eliminate its redundant SELayer, and the modified model is used as the backbone of Yolov5s to realize an entire network model that is lightweight. Second, to reduce the damage caused by low-quality images to the anchor frame regression, W-IoU is introduced. Finally, because the target detected in this case is a small target in the entire image, the efficient multiscale attention (EMA) mechanism and small-target detection layer are embedded in each downsampling layer of the backbone, thereby improving the attention to the features and enhancing the sensory field of the deep convolution. The improved model temporarily uses less space and also improves the detection accuracy, thereby effectively improving the efficiency of apple defect detection.

MATERIAL AND METHODS

To make the algorithm lightweight while improving detection accuracy, the network structure is first slimmed down. Subsequently, an attention mechanism is added to improve the detection accuracy. To reduce the damage caused by the box regression, the loss function is modified. Finally, for the characteristics of apple defects that are small targets in the entire apple, a feature extractor is added to ensure the detection accuracy of the small targets.

Improvement Of The Backbone Network

Certain embedded and edge devices are difficult to integrate perfectly with convolutional reach-in networks owing to their smaller storage capacities. To realize the slimming of the network model, the more lightweight MobileNetv3 (Howard et al., 2019) is used to replace the backbone network, CSPDarknet-53, of Yolov5s. In this study, MobileNetv3 is further lightened by removing the redundant SELayer. The core task is to develop the model using the separable convolution and linear activation functions, thus reducing the number of parameters and computation of the model to achieve a faster operation speed, as shown in Figure 1.

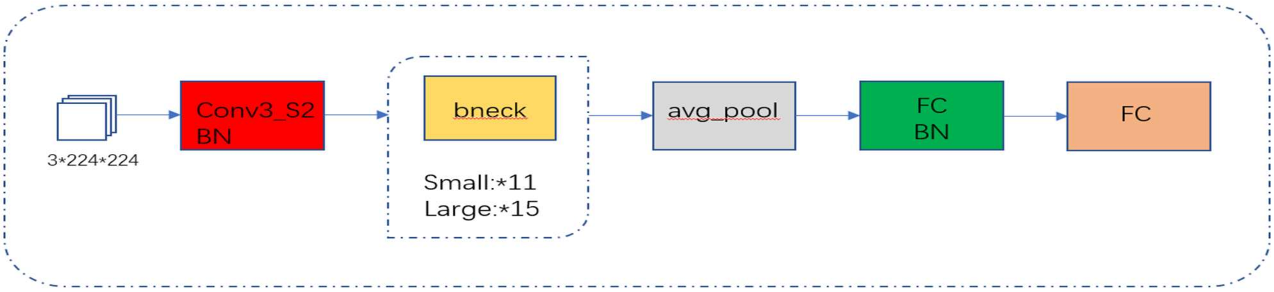


FIGURE 1. MobileNetV3 structure.

MobileNet is a lightweight neural network model proposed by the Google team that uses the idea of the depthwise separable convolution and inverse residual structure to build the model (Yi & Jia, 2023). As a lightweight network structure, it has considerably fewer operations and parameters than the traditional convolution modules. MobileNetv3 not only has the depth-separable convolutional structure of MobileNetv1 but also inherits the linear bottleneck-to-residual structure of MobileNetV2, in addition to the introduction of the Hard-swish function instead of the previous swish function that reduces the number of operations and improves performance. Table 1 presents the structural hierarchy.

TABLE 1. MobileNetV3 structure.

Input	Operator	Exp size	# Out	SE	NL	S
224^2*3	Conv2d	-	16	-	HS	2
112^2*16	Bneck,3*3	16	16	√	RE	2
56^2*16	Bneck,3*3	72	24	-	RE	2
28^2*24	Bneck,3*3	88	24	-	RE	1
28^2*24	Bneck,5*5	96	40	√	HS	2
14^2*40	Bneck,5*5	240	40	√	HS	1
14^2*40	Bneck,5*5	240	40	√	HS	1
14^2*40	Bneck,5*5	120	48	√	HS	1
14^2*48	Bneck,5*5	144	48	√	HS	1
14^2*48	Bneck,5*5	288	96	√	HS	2
7^2*96	Bneck,5*5	576	96	√	HS	1
7^2*96	Bneck,5*5	576	96	√	HS	1
7^2*96	Conv2d,1*1	-	576	√	HS	1
7^2*576	Pool,7*7	-	-	-	-	1
1^2*576	Conv2d 1*1,NBN	-	1024	-	HS	1
1^2*1024	Conv2d 1*1,NBN	-	K	-	-	1

MobileNetv3 optimizes the backbone network through two aspects.

(1) The main feature of MobileNetV3 is channel separable convolution (Howard et al., 2017). Compared with the traditional convolution process, channel separable convolution reduces the parameters and computational complexity while retaining the model feature expression capability. This is mainly divided into two processes. The first step is deep convolution. Each channel has its own individual convolution kernel for the convolution operation as a way of extracting features, compared with the traditional convolution process. This operation greatly reduces the computational complexity. The second step is point-by-point convolution, the results obtained in the first step, point-by-point convolution with a 1*1 convolution kernel for convolution operation. This operation makes the features of each channel get excellent integration effect. The specific process is shown in Figure 2.

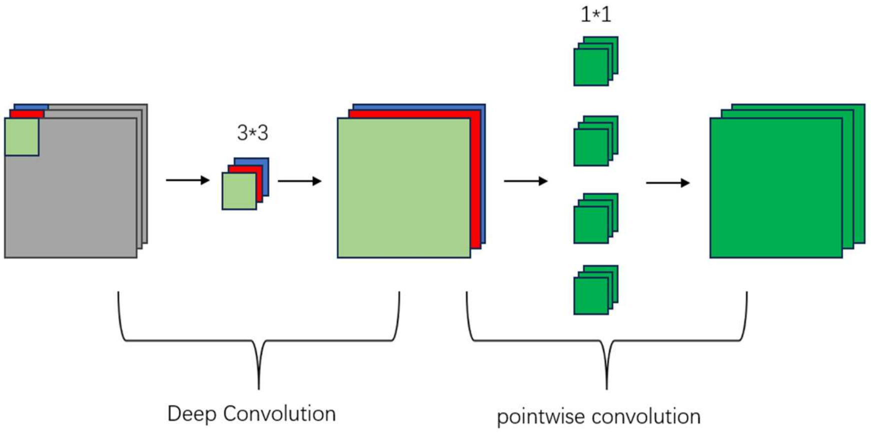


FIGURE 2. Depth-separable convolution process.

(2) MobileNetV3 introduces the H-swish function that not only solves the problem of one-sided suppression of the ReLU function but also has less computational overhead than the Swish function that is more suitable for mobile devices. The expression for the H-swish function is as follows.

$$h - swis\hbar[x] = x\frac{Re\,LU6(x+3)}{6} \qquad (1)$$

Attention Mechanism

The EMA mechanism (Ouyang et al., 2023) can fuse the feature information under different scale pictures, divide the channel dimension into multiple sub-features, encode the global information, and aggregate the output features of two parallel branches using the cross-dimension interaction method. First, the attention weight descriptors of the grouped feature maps are extracted using three parallel routes, where two parallel routes are located in the 1*1 branch and the third route is located in the 3*3 branch. Two one-dimensional average pooling operations are used to encode the channels along two different directions in the 1*1 branch. In the 3*3 branch, a 3*3 convolutional kernel is placed to capture multiscale features. Two-dimensional global average pooling is used to encode the global information in the 1*1 branch, where the encoded one-dimensional global average pooling operations at H and W along the horizontal direction in C can be expressed as in eqs (2)-(4).

$$Z_C^H(H) = \frac{1}{W}\sum\nolimits_0^W X_C(H,i) \qquad (2)$$

$$Z_C^H(W) = \frac{1}{H}\sum\nolimits_0^H X_C(j,W) \qquad (3)$$

$$Z_c = \frac{1}{H*w}\sum\nolimits_j^H \sum\nolimits_i^W X_C(i,j) \qquad (4)$$

EMA is added after each downsampling layer in the backbone layer. This attention mechanism can obtain the feature information of the image in different sizes, perform the weighting operation, and place it in each downsampling layer such that the EMA mechanism can link the context and can better handle the relationship between the feature maps of each channel. The specific network structure is illustrated in Figure 3.

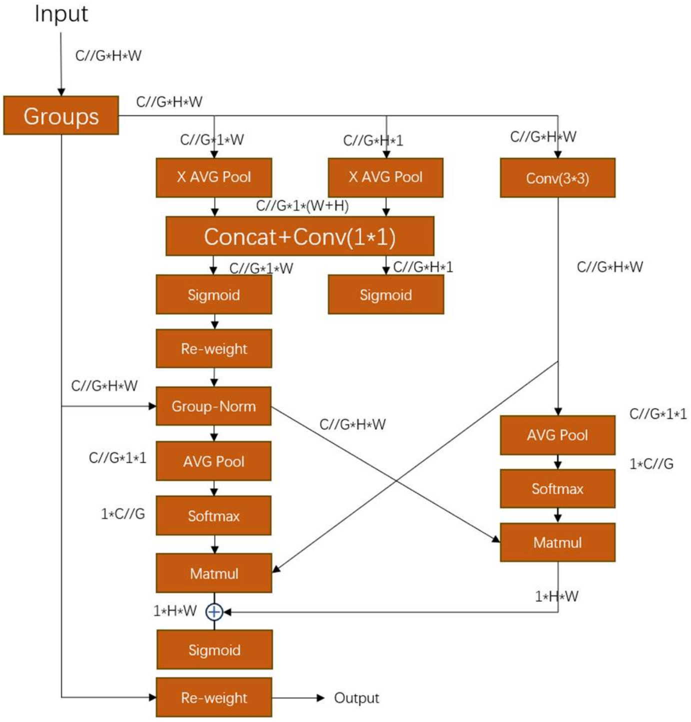


FIGURE 3. EMA network architecture diagram.

Improvement of the loss function

The loss function is used to measure the difference between the predicted and real values of a model. During model training, the loss function calculates a value based on the predicted output value of the model and the actual labeled value; this value is the loss value during training. The smaller the loss value, the more accurate the model prediction ability, the closer the model parameter update will be to the optimal solution, and the better the training results. IoU refers to the intersection and merger of the target and prediction frames, as shown in Figure 4.

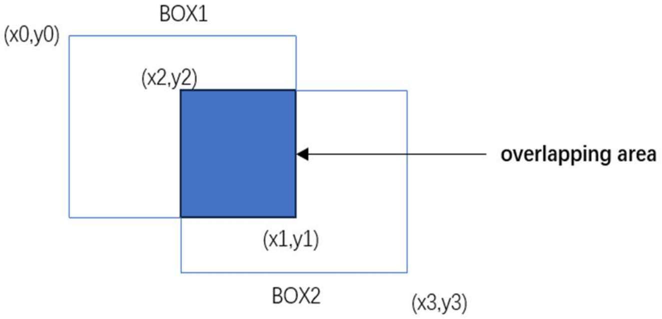


FIGURE 4. IoU loss function.

IoU loss is the edge loss regression function used by the YOLO algorithm that is mainly used to calculate the intersection of the predicted and real regions and calculate the de-intersection ratio; the smaller the value, the lower the overlap between the predicted region and the real region, and the higher the loss, that is written as LioU. However, when BOX1 and BOX2 do not intersect, the value of IoU is zero, and a specific intersection between the two frames cannot be derived. Zheng et al. (2020) and other researchers proposed D-IoU using IoU as the basis and adding a penalty term to converge the distance between the predicted box and the standard box more quickly. Thus, D-IoU fills a part of the IoU gap but does not consider the aspect ratio of the bounding box, and in Yolov5, C-IoU (Bodla et al., 2017) is used to solve the problem of bounding box accuracy. The specific expressions for IoU and C-IoU are presented in eqs (5)-(8).

$$L_{IoU} = 1 - IoU = 1 - \left|\frac{B \cap B^{gt}}{B \cup B^{gt}}\right| \tag{5}$$

$$L_{cIou} = 1 - IoU + \frac{\rho^2(b, b^{gt})}{c^2} + \beta v \tag{6}$$

$$\beta = \frac{v}{(1 - iou) + v} \tag{7}$$

$$v = \frac{4}{\pi^2}\left(\arctan\frac{w_{gt}}{h_{gt}} - \arctan\frac{w}{h}\right)^2 \tag{8}$$

Here:

B represents the range of the target frame;

B_{gt} represents the range of the prediction frame;

ρ() represents the Euclidean distance;

c denotes the diagonal length of the minimum envelope of the two frames;

β denotes the influence factor; the larger the IoU, the larger it influences;

v characterizes the consistency of the aspect ratio;

w and h represent the width and height of the target frame, respectively;

w_{gt} and h_{gt} represent the width and height of the prediction frame, respectively.

Because low-quality examples inevitably appear in the training dataset and C-IoU adopts a monotonic focusing mechanism, if we continue emphasizing the regression of the bounding box on the low-quality examples, it will inevitably affect the enhancement of the detection performance of the model. In this study, to enhance the generalization performance of the model, the loss function of Yolov5s is improved to that of the V3 version of Wise-IoU (Tong et al., 2023). The Wise- IoU is a dynamic non-monotonic FM loss that reduces the competitiveness of high-quality anchor frames and reduces the deleterious gradient generated by low-quality samples; this allows the Wise-IoU loss function to better improve the detection performance of the network. The anchor and target frames of Wise-IoU are shown in Figure 5.

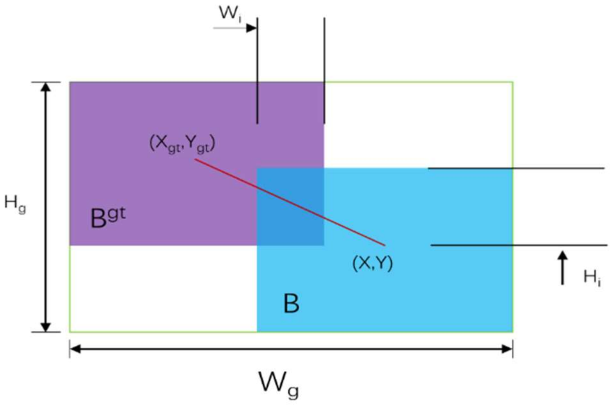


FIGURE. 5. Schematic of W-IoU.

Here, B represents the anchor frame and Bgt represents the target frame that are expressed as in eqs (9)-(11).

$$R_{WIoU} = \exp(\frac{(x - x_{gt})^2 + (y - y_{gt})^2}{(W_g^2 + H_g^2)^*}) \tag{9}$$

$$L_{WIoUv1} = R_{WIoU} L_{IoU} \tag{10}$$

$$\beta = \frac{L^* IoU}{\overline{LIoU}} \qquad r = \frac{\beta}{\delta \alpha^\beta} \tag{11}$$

$L_{IoU} \in [0,1]$ significantly reduces R_{WIoU} of the high-quality anchor frame and reduces the effect of geometric factors when the center distance between the anchor frame and the target frame is small. $R_{WIoU} \in [1,e)$ significantly amplifies L_{IoU} of the general quality anchor frame. W_g and H_g represent the width and height of the minimum enclosing frame, respectively; (x,y) and (x_{gt},y_{gt}) represent the centroids of the anchor and target frames, respectively; r denotes the nonmonotonic focusing coefficient; β denotes the degree of outlier that represents the quality of the regression frames; $L^{*}IoU$ denotes the monotonic focusing coefficient; $\overline{LIoU}$ represents the sliding mean of momentum m; δ and α represent the hyperparameters that can be tuned according to the different environments.

Feature Extractor

The feature pyramid network structure used by Yolov5s uses 8-fold, 16-fold, and 32-fold downsampling to obtain three additional sizes of feature images. However, according to the idea of a feature pyramid network (Lin et al., 2017a), after many convolution and feature extraction operations, a part of the feature information is lost in the deep and large feature maps. Because deep feature maps have a large sensory field, the loss of a small portion of the information will result in the incomplete extraction of the features of the entire image.

By combining the feature pyramid with the path aggregation network, the feature pyramid network conveys deep semantic features from top to bottom, and the path aggregation network conveys the location information of the target from bottom to top. By fusing top-down and bottom-up feature information, the model can learn features better and improve its accuracy for small target detection (Li & Wu, 2022). The target in this study occupies only a small part of the image; therefore, adding a detection layer with four times downsampling in the head layer can provide more attention to the small target, increase the accuracy of small target detection, and retain more feature information. The working principle of the small target detection layer is illustrated in Figure 6.

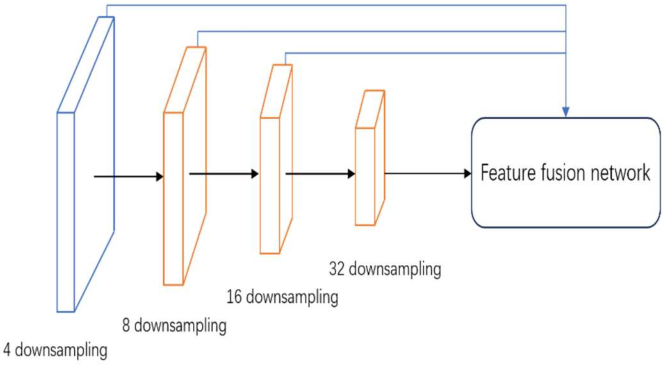


FIGURE 6. Small target detection layer.

Yolov5s-Super Network Structure

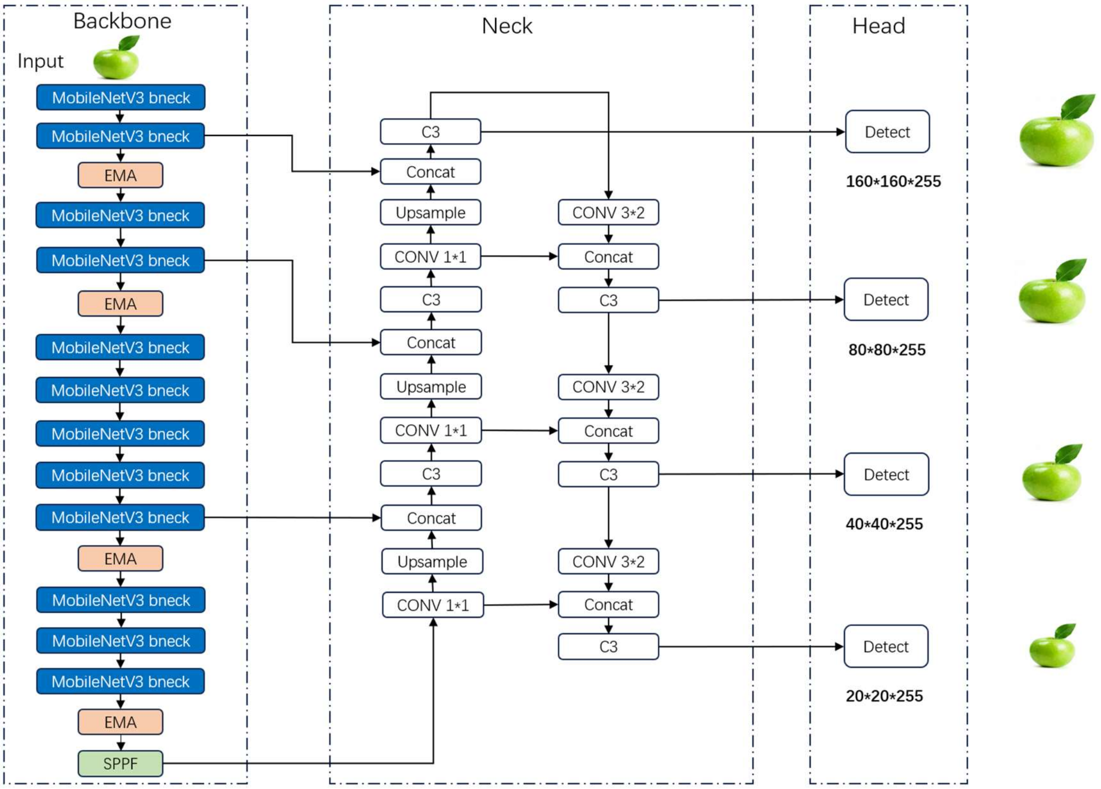


FIGURE 7. Yolov5s-Super network structure.

Experimental Environment and Datasets

The GPU used for the experiment is Nvidia GeForce RTX 3080, CPU is Xeon(R) Platinum 8255C, graphics driver is CUDA 11.7, programming language is Python 3.8, programming platform is PyCharm 2023, batch size is 24, and the maximum number of epochs is 700.

In this study, we use web public and homemade datasets. A total of 374 images are included, and 1196 training sets, 60 image validation sets, and 80 test sets are obtained by adding noise, rotating, cropping, and other operations. Labeling is performed using LabelImg with four categories: Blotch, Healthy, Rot, and Scab.

Evaluation Indicators

This study uses precision, recall, mAP@50 (Girshick, 2015), map@50:95 (Lin et al., 2017b), and F1 (Dempster et al., 1977) scores. Precision P is the ratio of positive samples correctly identified by the algorithm to the total number of samples identified as positive. Recall R, also known as the check-all rate, is the ratio of positive samples correctly identified by the algorithm to the total number of positive samples in the original sample. See eqs (12) and (13).

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

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

Here:

TP indicates a positive sample and positive test result;

FP indicates a negative sample and positive test result;

TN indicates a negative sample and negative test result;

FN indicates a positive sample and negative test result.

mAP@50 is the mean accuracy value, a widely used metric to measure the performance of target detection and object recognition algorithms. The mAP combines precision and recall to comprehensively evaluate the detection accuracy and detection rate of the model in different categories (Lowe 1999). Denotes the average AP of each category over all images calculated for IoU=0.5. mAP@0.5:0.95 means stacked starting with step i=0.05 until the average IoU=0.95. The calculation formula is expressed as in [eq. (14)].

$$mAP = \frac{\sum_{i=1}^{K} AP_i}{K} \tag{14}$$

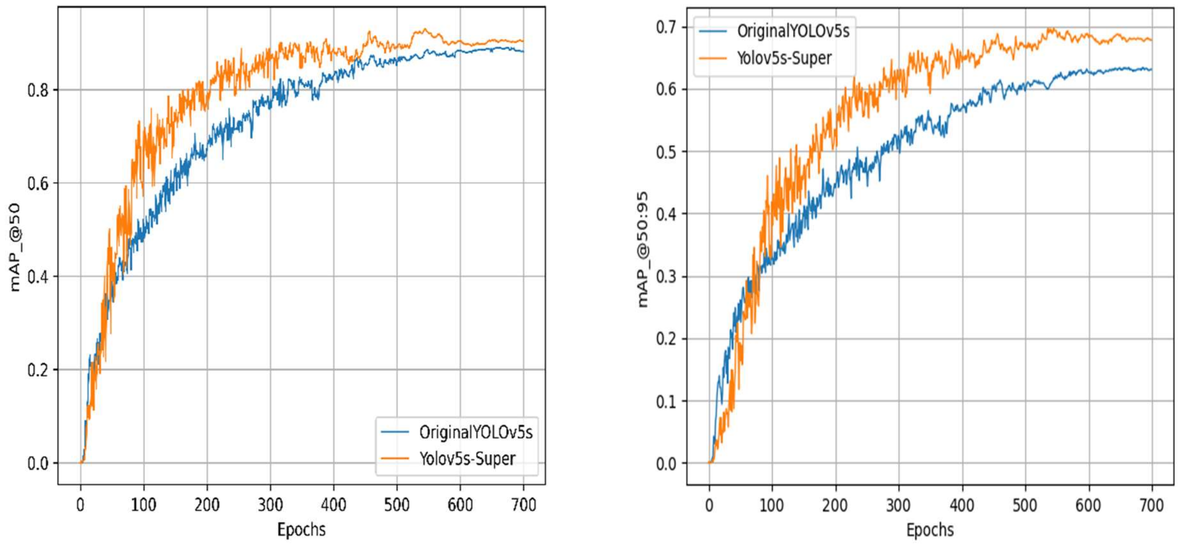
F1 is another performance metric that combines precision and recall and represents the reconciled average of precision and recall. Its expression is as in [eq. (15)].

$$F_1 = 2 * \frac{(precision*recall)}{precision+recall} \tag{15}$$

RESULTS AND DISCUSSION

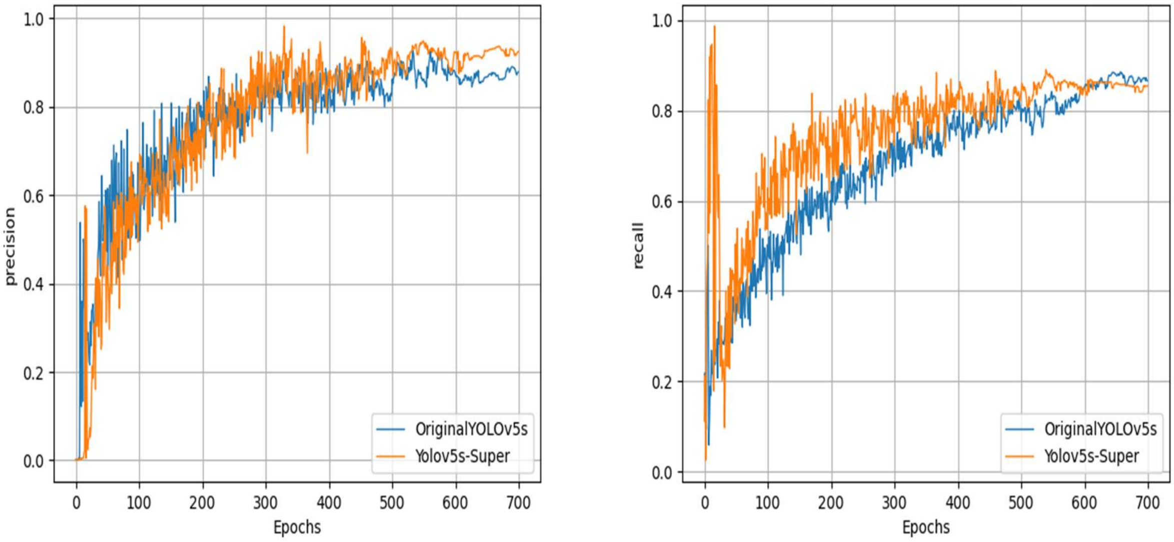
Comparison Experiment

Comparative experiments were conducted to compare the training performances of Yolov5s-Super and Yolov5s. To more intuitively visualize the gap between the original Yolov5s algorithm and the Yolov5s-Super algorithm, we plotted line graphs and tables. The results are presented in Figure 8 and Table 2.



(a)

(b)



(c)

(d)

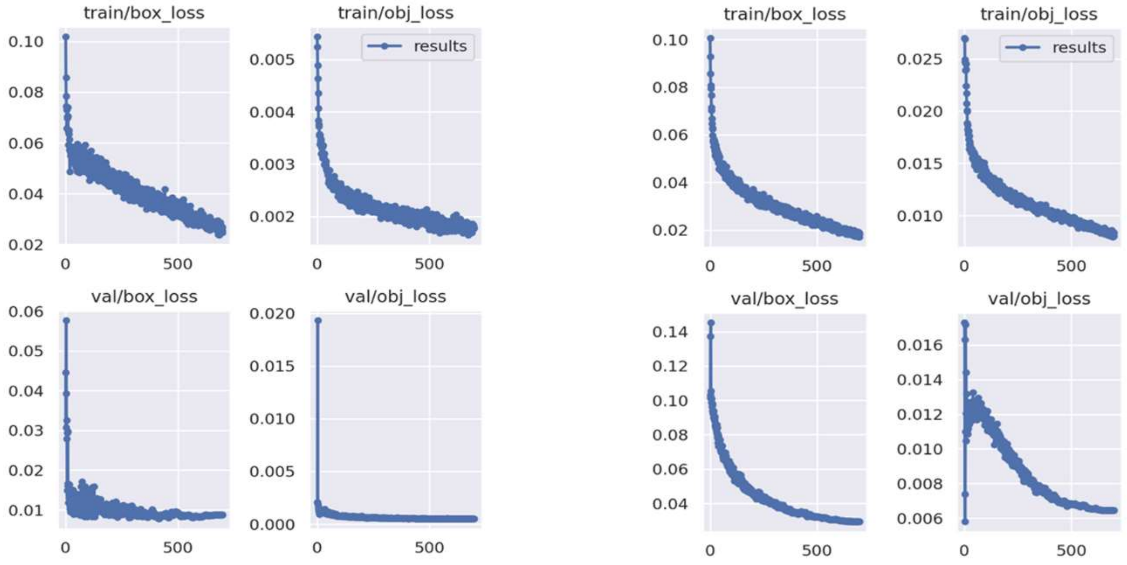
FIGURE 8. Comparison results between Yolov5s and Yolov5s-Super in 700 epochs. (a) mAP@50, (b) mAP@50:95, (c) precision, (d) recall.

TABLE 2. Comparison results between Yolov5s and Yolov5s-Super.

Model	mAP@50(%)	mAP@50:95(%)	P(%)	R(%)	Parameter(M)	F1(%)
Yolov5s	90	65.4	84.3	87.4	7.23	87
Yolov5s-Super	93.2	69.6	94.6	88.2	1.6	91

As shown in Figure 8 and Table 2, the improved algorithm achieved significant improvements in all indicators, including key indicators, while reducing the overall network structure parameter count by 77.8%. mAP@0.5, mAP@50:95, and F1 improved by 5%, 6%, and 3%, respectively, with an accuracy (P) improvement of approximately 10%. This indicated that the model structure proposed in this study achieved light weight and significantly improved the key indicators of each visual algorithm, with all indicators performing better than the original Yolov5s algorithm.

The loss of the detection frame indicates how well the algorithm can localize the centroid of the object and whether the detection target is covered by the predicted bounding box. The smaller the value of the loss function, the more accurate the predicted frame. The target loss function is essentially a measure of the probability that the detection target exists in the anchor frame region; the smaller the value of the loss function, the higher the accuracy (Chen et al., 2022). Box_loss denotes the box regression loss that is used to measure the difference between the positions of the predicted and actual frames in the target detection; a lower value indicates that it is closer to the position of the real frame. Obj_loss denotes the target confidence loss that indicates the accuracy of the judgment of the model regarding the presence of a target to be detected in the image; a lower value indicates that the model can accurately identify the target in the image.



(a)

(b)

FIGURE 9. Bounding box loss functions. (a) Yolov5s-super, (b) Yolov5s.

As shown in Figure 9, in the entire training process, the function image exhibited a downward trend, with the stochastic gradient algorithm used to optimize the network, and the model in the continuous learning process of weights and parameters was constantly updated and finally converged. In the original Yolov5s algorithm, the convergence speed is slower and its function image has a large fluctuation, indicating that the model for target detection is prone to omission and incorrect detection. In summary, the final effect and convergence results of the proposed method were better than those of the original Yolov5 algorithm.

Results Of Ablation Experiments

To analyze the effect of introducing each module in the model for apple defect detection, ablation experiments were designed to characterize different defects in apples, improve target, and calculate the contribution of each module independently, as listed in Table 3.

TABLE 3. Results of ablation experiments.

Model	P（%）	R（%）	mAP@50(%)	mAP@50:95(%)	Parameter(M)
Yolov5s	84.3	87.4	90	65.4	7.2
Yolov5-MobileNetV3	78.5	84.8	86.3	60.8	0.93
Yolov5s-MobileNetV3+W-IoU	83.1	76.9	89.2	63	0.93
Yolov5s-MobileNetV3+W-IoU+EMA	88.1	84.5	88.9	66.7	0.94
Improved-Yolov5	94.6	88.2	93.2	69.6	1.6

As presented in Table 3, replacing the backbone network of Yolov5s with the improved MobileNetV3 achieved overall network lightweighting, with a decrease of 87% in parameter count, but resulted in improved accuracy, recall, mAP@50 and mAP@50:95 decreased by 9.1%, 3.4%, 3.7%, and 4.6%, respectively. To reduce the decrease in detection accuracy caused by box regression, the loss function was replaced with W-IoU that significantly improved the other indicators but decreased the recall rate. At this time, the model could not correctly and completely identify the target object. Therefore, by introducing the EMA mechanism and embedding it into each downsampling layer of the backbone, the ability of the model to extract targets and their features was improved. Finally, by combining several improved methods with a small object detection layer, Yolov5s-Super achieved not only a lightweight structure but also improved accuracy, recall, mAP@50, and mAP@50:95.

The ablation experiment proved that the improved MobileNetV3 backbone network in this study effectively achieved lightweighting of the model and did not conflict with the neck layer of the original network. Targeted improvements were then made based on the different characteristics of apple defects. The results presented in Table 3 proved the effectiveness of each improvement method, providing theoretical support for the migration of the network model to apple defect detection embedded devices.

Comparison Of Experimental Results With Other Visual Networks

To verify the performance of the proposed model, the model was compared with four other commonly used depth detection based target detection models in the same experimental environment; the results are presented in Table 4.

TABLE 4. Results of comparative experiments with other models.

Object Detection Networks	P(%)	Number of Parameters(M)	Size of Model(MB)	mAP@50(%)
Yolov5s	84.3	7.2	14.1	90
Yolov5m	86.2	21.2	42.3	91.8
Yolov5l	79	109.1	92.9	85.3
Yolov8s	95	11.01	21.5	95.2
Improved Yolov5s	94.6	1.6	4.3	93.2

As presented in Table 4, the proposed algorithm was the lightest among the five models and outperformed the other algorithms in the Yolov5 series in terms of accuracy and mAP values. Yolov5s-Super differed in detection accuracy and mAP values by only 0.4% and 2% when the number of parameters was 85.4% less than that of Yolov8s.

Test Results

To better demonstrate the detection effect of Yolov5s-Super in this study, the detection results of Yolov5s-Super and the original Yolov5s were compared, as shown in Figure 10.

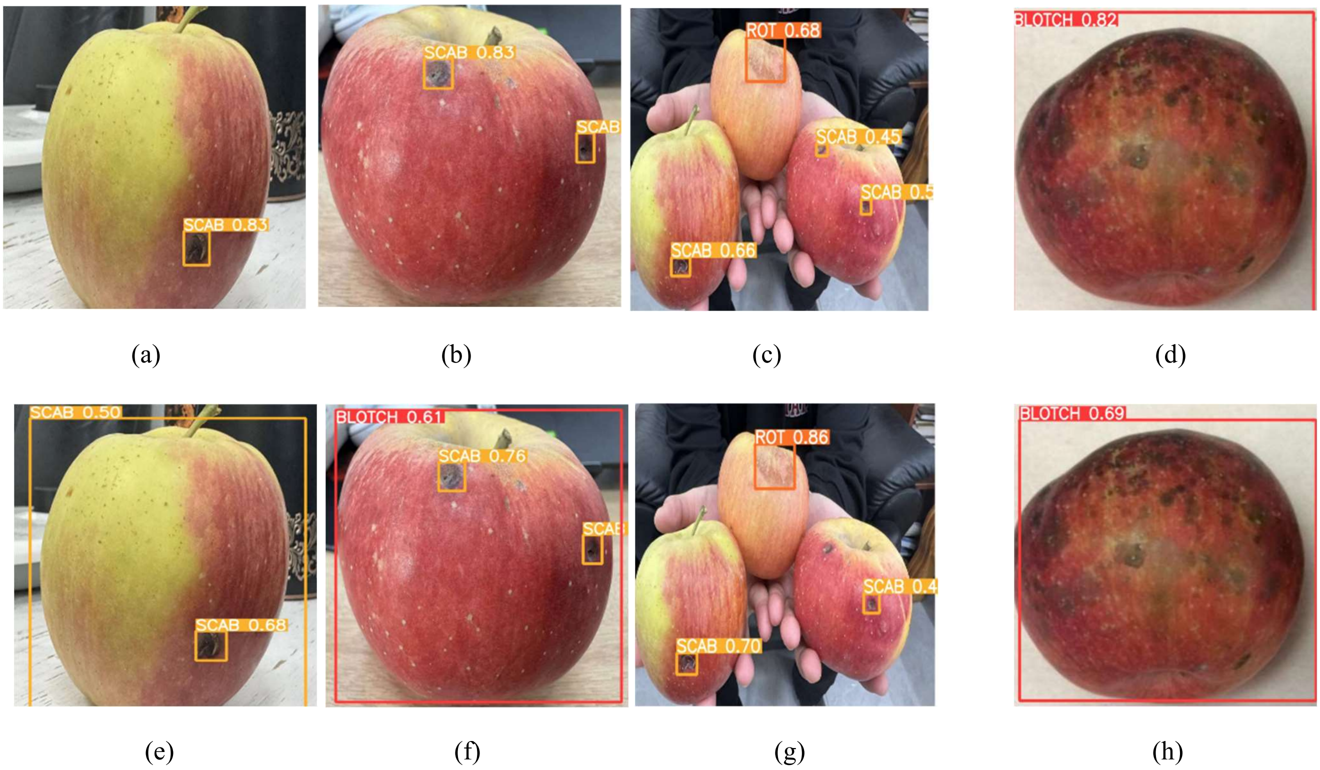


FIGURE 10. Comparison of the detection effect. (a),(b),(c),(d) are the test results of Yolov5s-Super, and (e),(f),(g),(h) are the test results of Yolov5s.

As shown in Figure 10, Yolov5s incorrectly detected the entire apple as a certain defect in the first two detections; in the third detection, Scab was undetected. In contrast, Yolov5s-Super effectively prevented this situation, and the detection accuracy of the target was higher than that of the original Yolov5s model, making Yolov5s-Super more suitable for the detection of apple defects.

CONCLUSIONS

Effective detection of defects in apples plays an important role in the economy of the farmer and company. In this study, we proposed a lightweight multi-defect vision algorithm for apples based on Yolov5s. To adapt to embedded devices, the entire network structure was lightened, and the loss function was optimized for the problem of difficult detection of low-quality pictures. Because the defects of apples are smaller targets in the entire detection range, an attention mechanism and a small target detection layer were introduced to improve detection precision and localization accuracy. The experimental results demonstrated the effectiveness of Yolov5s-Super in detecting defects in apples. This provides theoretical support for the migration to embedded devices.

REFERENCES

Bodla N, Singh B, Chellappa R, Davis LS (2017) Soft-NMS--improving object detection with one line of code. In: IEEE International Conference on Computer Vision. Venice, Proceedings… p. 5561-5569.

Chen Z, Wu R, Lin Y, Li C, Chen S, Yuan Z, Zou X (2022) Plant disease recognition model based on improved YOLOv5. Agronomy 12(02):365.

Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society series B (methodological) 39(01): 1-22.

Dong Z, Wang Y (2023) Crop disease and pest identification technology based on ACPSO-SVM algorithm optimization. Engenharia Agrícola 43(5): e20230104.

Dong Z, Jiang J, Wang Y (2024) Parameter optimization design of precision seeding device based on the bp neural network for panax notoginseng. Engenharia Agrícola 44: e20230161.

Girshick R (2015) Fast r-cnn. In: IEEE International Conference on Computer Vision. Santiago, Proceedings… p. 1440-1448.

Han K, Wang Y, Tian Q, Guo J, Xu C (2020) Ghostnet: More features from cheap operations. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. Proceedings… p. 1580-1589.

Howard A, Sandler M, Chu G, Chen LC, Chen B, Tan M, Adam H (2019) Searching for mobilenetv3. In: IEEE/CVF International Conference on Computer Vision. Proceedings… p. 1314-1324.

Howard AG, Zhu M, Chen B, Kalenichenko D, Wang W, Weyand T (2017) Mobilenets: efficient convolutional neural networks for mobile vision applications. ArXiv abs/1704.04861.

Hu TH, Gao XM, Hua YS, Cai LJ (2023) Deep lerning-based adaptive muti defet detection in aple images. Journal of Shandong University of Technology (Natural Science Edition) 38(01):42-47.

Kamilaris A, Prenafeta B (2018) Deep learning in agriculture: a survey. Computers and electronics in agriculture 147(01): 70-90.

Li R, Wu Y (2022) Improved YOLO v5 wheat ear detection algorithm based on attention mechanism. Electronics 11(11): 1673.

Lin TY, Dollár P, Girshick R, He K, Hariharan B, Belongie S (2017a) Feature pyramid networks for object detection. In: IEEE Conference on Computer Vision and Pattern Recognition. Proceedings… p. 2117-2125.

Lin TY, Goyal P, Girshick R, He K, Dollár P (2017b) Focal loss for dense object detection. In: IEEE International Conference on Computer Vision. Proceedings… p. 2980-2988.

Liu W, Dragomir A, Dumitru E, Christian S, Scott R, Fu CY (2016) Ssd: Single shot multibox detector. In: The Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, Netherlands, Proceedings… Part I 14: 21-37.

Lowe DG (1999) Object recognition from local scale-invariant features. In: IEEE International Conference on Computer Vision. Proceedings… p. 1150-1157

Mei JB, Li T, Qin YC (2023) Research on Apple Picking Robot Monitoring System and Surface Defect Detetion Method. Computer Measurement & Control 31(06):19-26.

Ouyang D, Su H, Zhang GZ, Luo MZ, Guo HY, Zhan J, Huang Z (2023) Efficient Multi-Scale Attention Module with Cross-Spatial Learning. In: ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Proceedings… p. 1-5

Ren SQ, He KM, Girshick (2015) Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in Neural Information Processing Systems. Proceedings… p. 28

Ross G, Jeff D, Trevor D, Jitendra M (2014) Rich feature hierarchies for accurate object detection and semantic segmentation. In: IEEE Conference on Computer Vision and Pattern Recognition. Proceedings… p. 580-587.

Samajpati BJ, Degadwala SD (2015) A survey on apple fruit diseases detection and classification. International Journal of Computer Applications 130(13): 0975-8887.

Tian, Y, Yang G, Wang Z, Li E, Liang Z (2019) Detection of apple lesions in orchards based on deep learning methods of cyclegan and yolov3-dense. Journal of Sensors 2019(01): 1-13.

Tong Z, Chen Y, Xu Z, Yu R (2023) Wise-IoU: Bounding box regression loss with dynamic focusing mechanism. arXiv preprint arXiv:2301.10051.

Valdez P (2020) Apple defect detection using deep learning based object detection for better post harvest handling. ArXiv: abs/2005.06089.

Wang QS, Lü L, Huang DF, Fu SQ, Yu HY (2022) Research of apple leaf disease defect detection based on improved YOLOv4 algorithm. Journal of Chinese Agricultural Mechanization 43(11):182-187.

Wang Z, Zhang T, Han J, Zhang L, Wang B (2023) Research on identification of crop leaf pests and diseases based on few-shot learning. Engenharia Agrícola 43(6): e20230140.

Yi ZJ, Jia Y (2023) Recognition of Color-ring Resitors Based on improved MobileNeV3. Computer Systems & Applications 32(04): 361-367.

Yu M, WEN Y, Mao M (2022) The pattern and development trend of China's apple foreign trade. China Fruit 225(07):100-104.

Zheng Z, Wang P, Liu W, Li J, Ye R, Ren D (2020) Distance-IoU loss: faster and better learning for bounding box regression. Proceedings of the AAAI Conference on Artificial Intelligence 34(07):12993-13000.