Apple Detection Algorithm under Complex Background Based on Improved YOLOv8

Qiang Guo, Chi Ma*, and Hui Hu

Abstract—**Accurate detection of apples provides the basis for picking and recognition. Attempting to identify the characteristics of a large number of apples and a complex background when they are on the tree, an apple detection algorithm APP-YOLOv8 based on the improved YOLOv8 network model is proposed. The baseline architecture incorporates structural optimization through the integration of the BoT3 module within ResNet's residual building units, enhancing the efficiency of the model while streamlining the count of parameters. Secondly, the CBMA attention module was designed and added to the backbone network to suppress complex background interference and improve the extraction efficiency of important features. Then, DCNv2 is embedded in ELAN, a highly efficient layer aggregation network, to expand the modulation mechanism of the deformation modeling range and enhance the modeling ability. Finally, the overall detection performance of the improved model was improved by optimizing the loss function. The experiments were carried out on the Apple dataset. Quantitative analysis reveals the proposed architecture achieves 85.4% detection accuracy with 93 frames/second processing speed, outperforming conventional YOLOv8 by 7.2% in precision metrics while satisfying real-time processing constraints. In addition, compared to the current advanced object detection algorithms, the APP-YOLOv8 model has advantages in small target apple detection and speed, proving that the proposed algorithm is more suitable for apple detection tasks.**

Index Terms—**Small object detection, YOLOv8, apple detection, attention mechanism.**

I. Introduction

With the rapid development of the apple industry, the international production of apples has steadily increased and the cultivation area has gradually expanded. Automated apple detection systems can significantly improve the efficiency of orchard management, harvesting, and packaging. The data collected during apple detection can be used to analyze crop growth trends, predict yield, and optimize planting strategies to provide scientific basis for agricultural decision making. With the rapid development of sensor technology, nanoscience and artificial intelligence, new sensor technologies and intelligent inspection systems provide new possibilities for apple non-destructive testing and quality grading. The automated apple inspection system can quickly identify and classify apples, which can detect diseases, pests, rot, or other surface defects in time to ensure that consumers eat apples safely. In agricultural science research, apple detection can provide valuable data on the growth pattern, maturity, and physiological state of apples and promote the development of agricultural science. At present, apple detection and identification still mainly depend on the experience of growers, which cannot meet the needs of rapid, accurate identification, and automation of large-scale orchards. Therefore, exploring an efficient apple detection algorithm is of great significance to solve this problem.

Deep learning [1-3] is an important technical means for processing, analyzing and learning plant images sensed by imaging sensors, and obtaining target information from them. Contemporary deep learning-based object detection architectures are typically bifurcated into two methodological paradigms: region proposal-driven approaches exemplified by R-CNN variants (two-stage detectors), and unified regression frameworks like YOLO and SSD that perform simultaneous localization and classification (single-stage detectors) [4]. The first type is slow and cannot meet the real-time requirements of apple detection. Although the second type can meet the requirements of real-time, the detection accuracy is low [5-6]. At present, some research results have been achieved by constructing deep learning models to identify crop detection. For example: For the extraction of complex features of target images or the recognition of small targets under complex backgrounds, Wang Yunlu et al. [7] used the feature pyramid network FPN and the Cascade mechanism in Cascade R-CNN to improve the Faster R-CNN algorithm, and realized the lossless identification of five kinds of apple diseases in natural scenes. The accuracy is 98.3%. Li Xinran et al. [8] improved the Faster R-CNN model through deep feature fusion and accurate region of interest pooling, and realized the effective detection of apple leaf diseases under natural environment conditions, with an average accuracy of 82.42%. Bao Wenxia et al. [9] improved the VGG16 network model by adding a convolution module to the neck network, and achieved a recognition rate of 94.7% for apple leaves. Zhang Zhiyuan et al. [10] adopted the strategy of combining offline data enhancement and online data enhancement to realize the recognition of cherry fruits in the natural environment, and the recognition rate reached 98%. In order to realize the accurate recognition and picking of dragon fruit in the natural environment, Shang Fengnan et al. [11] took the YOLOX-Nano model as the benchmark, and added the convolutional attention module to the backbone feature extraction network. The detection speed and accuracy were improved, and the recognition accuracy reached 98%. Fan Xiangpeng et al. [12] improved the full connection layer and classification layer of the VGG16 network model, and achieved an accuracy of 95.67% in grape detection under natural environmental conditions. Wang Guowei et al. [13]

Manuscript received October 11, 2024; revised March 25, 2025. This article is supported by the Foundation on Guangdong Educational Committee under Grant No. 2022ZDZX4052, 2021ZDJS082.

Qiang Guo is a postgraduate student of school of computer science and software engineering, university of science and technology LiaoNing. AnShan 114051, China (e-mail: 254755139@qq.com).

Chi Ma is an associate professor of the school of computer science and engineering, Huizhou university, Huizhou 516007, China (corresponding author to provide phone: 18641250800; e-mail:machi@hzu.edu.cn).

Hui Hu is a lecturer of the school of computer science and engineering, Huizhou university, Huizhou 516007, China (e-mail:huhui@hzu.edu.cn).

used Adam algorithm, Dropout strategy and ReLU excitation function to improve the LeNet model and improve the detection accuracy of corn, with an accuracy rate of 96%. Zhao Lixin et al. [14] improved the accuracy of cotton leaf recognition by transfer learning and improving Alex model, and the accuracy reached 97.16%.

Although deep learning technology has made certain progress in the detection and recognition of a variety of crops, due to the influence of different environmental factors, the occurrence characteristics of apple fruits in different backgrounds and different periods are different. The previous models cannot be directly applied with complex backgrounds. YOLOv8 [15] advances the YOLO lineage through three principal architectural innovations: 1) A revamped feature extraction backbone with enhanced cross-stage connectivity, 2) An anchor-free prediction head employing dense object priors, and 3) A task-aligned loss formulation optimized for heterogeneous computing architectures spanning from CPU to GPU clusters. Ultralytics deliberately employs framework-level terminology to emphasize its platform-agnostic infrastructure design. The core architecture implements plugin-compatible components through standardized interfaces, allowing seamless integration of non-YOLO architectures (e.g., Transformer-based detectors) and multi-task extensions (classification/segmentation/3D pose) while maintaining unified API specifications. Due to the poor detection effect of small-size targets of apples, the existing models have limited ability to learn the shape and texture features of apples. To this end, this study establishes a data set containing apples with a variety of complex backgrounds, uses YOLOv8 model as the backbone network, and introduces BoT3[17] module into the bottleneck block of ResNet[16] to improve performance and reduce parameters. Then, the CBMA[18] attention module is introduced to suppress complex background interference and improve the extraction efficiency of important features. The modulation mechanism of c2fdcn[19] is used to expand the scope of deformation modeling, and the modeling ability is enhanced. Finally, WIOU[20] loss function was used to optimize the network to improve the recognition efficiency and accuracy of small-scale apple features, so as to realize the real-time detection of small target apples under natural environment conditions.

II. Improved YOLOv8 Model construction

A. YOLOv8 Model

YOLOv8 algorithm provides a new SOTA model, including object detection network with P5 640 and P6 1280 resolution and instance segmentation model based on YOLACT [21]. Like YOLOv5, different sizes of models based on the scaling factor in NS/M/LI/X scale are also provided to meet different scene requirements. The architectural refinement involves strategic module replacement in backbone and neck components, where the legacy C3 block from YOLOv5 is reengineered into a C2f configuration through ELAN-inspired gradient pathway optimization. This hierarchical redesign enhances multi-scale feature propagation while implementing channel dimension adaptation specific to varying model scales. The stem to stem structure is no longer a set of parameters and all models are applied. This greatly improves model performance. However, the presence of operations such as Splt in this C2 module imposes restrictions on specific hardware deployments. Compared with YOLOv5, its Head part has changed a lot, and has been replaced by the current mainstream decoupled head structure, which separates the classification and detection heads. In terms of Loss calculation, the TaskAlignedAssigner positive sample assignment strategy is used, and the Distribution Focal Loss is introduced[22]. Finally, in the data augmentation part of training, we introduce the operation of turning off Mosiac augmentation in the last 10 epochs of YOLOX, which can effectively improve the accuracy.

B. Improved YOLOv8 Model

Apple images are collected in the wild, and the complex image background will interfere with the YOLOv8 model based on convolutional neural network, affecting the effect of apple detection. The improved strategy of this paper integrates the CBAM attention mechanism into the YOLOv8 network, so that the model pays more attention to the apple itself rather than the background environment. In the backbone network, the BoT3 module is introduced into the bottleneck block of ResNet to achieve performance improvement and reduce parameters. Secondly, in order to solve the problem of subsequent detection errors caused by the irregular size of apples and the difficulty of extracting shape features in complex environments, DCNv2 is embedded into ELAN, an efficient layer aggregation network, which extends the modulation mechanism of deformation modeling range and enhances the modeling ability. At the same time, in order to alleviate the convergence speed problem in the process of type training, the WIOU loss function is used to optimize the network and improve the overall detection performance of the improved model.

1) The BoT3 module is introduced in ResNet: BoTNet establishes an efficient neural substrate through hybrid attention mechanisms, where multi-head self-attention layers are strategically interleaved with convolutional operators. This architectural paradigm demonstrates cross-task superiority across visual recognition benchmarks (classification/detection/segmentation), achieving 3.1% mAP gains on COCO instance segmentation while maintaining 18% fewer parameters than ResNet baselines. In APP-YOLOv8, we replace 3×3 convolutions with multi-head Self-attention (MHSA) [23] in ResNet without any changes in the rest of the parts see Figure 1 (redrawing based on [23]).

The diagram of MHSA module is shown in Figure 2 (redrawing based on [23]).

2) CBMA Attention module: The APP-YOLOv8 algorithm implementation adds three CBAM modules to the backbone network to improve the feature extraction ability of the network. In the APP-YOLOv8 network structure, the feature map processed by the convolutional network is first input into the channel module. The channel attention module uses the relationship between channel attention of features to generate channel attention information, which is mainly used to determine the input image focus. The

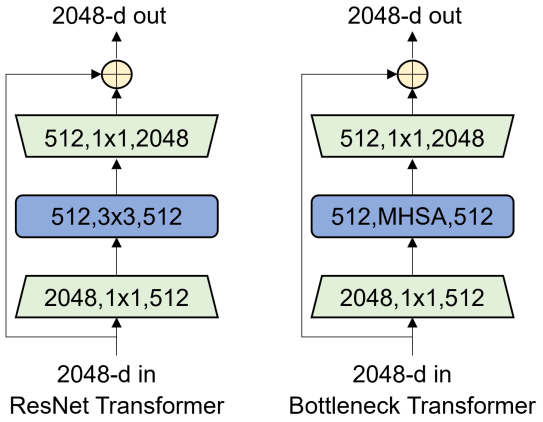

Fig. 1: Bot3 module

formula is as follows.

$$M_c(F) = \sigma(MLP(AvgPool(F)) + MLP(MaxPool(F))) = \grave{O}(W_1(W_0(F^c_{avg})) + W_1(W_0(F^c_{max}))) \quad (1)$$

σ represents the sigmoid function, W_0 and W_1 represent the weights of the shared MLP, and AvgPool and MaxPool represent the average and Max pooling operations, respectively. As information and channel attention complementary calculation formula is as follows.

$$M_S(F) = \sigma(f^{7\times7}\left[AvgPool(F); MaxPool(F)\right]) = \sigma(f^{7\times7}\left[F^x_{avg}; F^x_{max}\right]) \quad (2)$$

Where 7 is the convolution operation with filter size 7x7. In the process of spatial calculation, the module in the spatial attention map first generates two pooling features by pooling operation along the channel axis, and then concatenate them to generate an effective feature descriptor. After that, a standard convolution operation is used to generate the spatial attention map.

3) Optimization modeling of C2fDCN: In order to extract features from different scales and receptive fields, and fuse them to capture different levels of detail and structure information, DCNv2 is embedded into ELAN network to form DCNv2+ELAN module. The accuracy and robustness of feature extraction are improved by continuous use of DCNv2 convolution. The DCNv2+ELAN module structure is shown in Figure 3 and owned by the authors). The LeakyRelu activation function is used to replace the original Relu function, and all negative value weights are changed to nonzero slope to expand the scope of the Relu function and reduce the problem of apple missed detection caused by environmental factors.

DCNv2+ELAN module improves the feature extraction ability of the backbone network, but it increases the amount of calculation and increases the model detection time. Since apple detection is a real-time process and requires high detection speed, PConv[24] is used instead of conventional convolution to reduce the amount of calculation of the model and increase the number of hardware operations to achieve lower detection delay. Figure 4 illustrates the regular convolution and DCNv2 convolution processes in detail.

4) WIOU Loss Function Optimization: In object detection tasks, the loss function is often used to determine whether there is a gap between the training data and the actual data. An appropriate loss function can make the model converge faster during training, which is conducive to improving the detection performance of the model. The original YOLOv8 model uses the completed-intersection Over Union (CIoU) loss function as the bounding box loss function, which is defined as follows.

$$\begin{aligned} \text{Loss}_{\text{CIOU}} &= 1 - IoU + \frac{\rho^2\left(b, b^{gt}\right)}{d^2} + av \\ a &= \frac{v}{(1 - IoU) + V'} \\ v &= \frac{4}{\pi^2}\left(\arctan\frac{w^{gt}}{h^{gt}} - \arctan\frac{w}{h}\right)^2 \end{aligned} \quad (3)$$

IoU (Intersection over Union) constitutes a fundamental similarity measure in detection evaluation b and b^{gt} are the center points of the two bounding boxes; ρ is the distance between two center points. d is the diagonal distance of the smallest bounding box containing both boxes; α is the weighting function; v similarity for measuring the aspect ratio; w^{gt}, h^{gt}, w, h are the width and height of the true and predicted bounding boxes. CIoU normalizes the distance between two center points by constructing a penalty term to accelerate the convergence of the prediction box, which makes it better describe the overlap information in regression. This constraint formulation exhibits dimensional coupling limitations: When predicted and ground truth boxes achieve aspect ratio equivalence, the proportionality regularization term collapses to zero, thereby creating degenerate gradients for simultaneous height-width adjustment during regression optimization. And the problem of increasing penalty for low-quality training samples caused by geometric factors. Therefore, the Wise-IoU loss function is used instead of the CIoU loss function, which is defined as follows.

$$\begin{aligned} \text{Loss}_{\text{WIoU}} &= \frac{\beta}{\delta\alpha^{\beta}} R_{\text{WIoU}}\ \text{Loss}_{\text{IoU}} \\ R_{\text{WIoU}} &= \exp\left(\frac{(x - x^{gt})^2 + (y - y^{gt})^2}{\left(W_g^2 + H_g^2\right)^*}\right) \\ \text{Loss}_{\text{IoU}} &= 1 - \text{IoU} \end{aligned} \quad (4)$$

Where x^{gt} and y^{gt} are the center points of the real bounding box; Wg, Hg is the height and width of the smallest bounding box containing both the predicted box and the true box; * is a split operation, i.e. Wg, Hg are separated from the computation graph; β is the degree of outlier.

WIoUv3 loss function constructs distance attention based on distance metric, which adjusts the penalty of geometric metric according to the coincidence degree of anchor box and target box, and solves the problem that the aspect ratio of CIoU is always 0 in individual cases. At the same time, by separating the height and width of the minimum bounding box, the influence of geometric factors on the penalty of low-quality training samples is weakened. In addition, WIoUv3 defines the outlier degree β, which is directly proportional to the size of the assigned gradient gain. If the value of β is smaller, the quality of the anchor box is higher and the assigned gradient gain is smaller, so that the bounding box regression focuses on the anchor

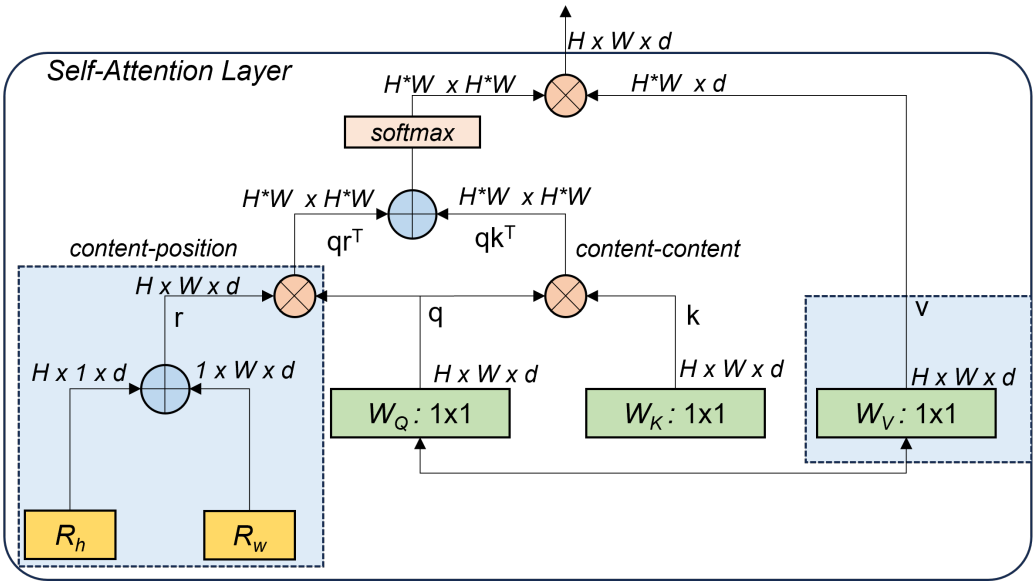

Fig. 2: MHSA module

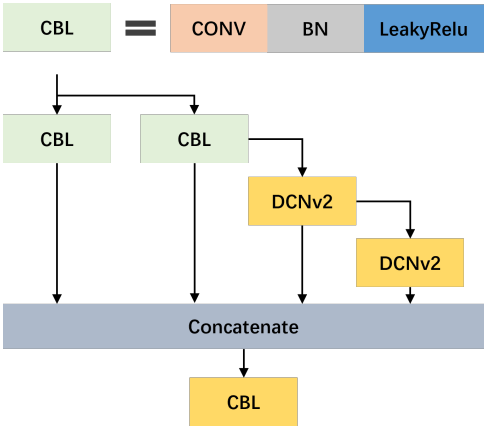

Fig. 3: The DCNv2 and ELAN module is described in detail

Fig. 4: Regular convolution and DCNv2 convolution

box of ordinary quality. On the contrary, if the value of β is larger, the quality of the anchor box is lower and the assigned gradient gain is larger. This prevents harmful gradients from being generated by low-quality anchors. The specific definition of β is shown in Equation 4 .

$$\beta = \frac{Loss^{*}_{IoU}}{\overline{Loss_{IoU}}} \tag{5}$$

Where $Loss^{*}_{IoU}$ is monotone focusing coefficient; $\overline{Loss_{IoU}}$ is the moving average of the momentum m. Among them, $Loss_{IoU}$ is used as the momentum, so that the value of β is always dynamically updated, and the division standard of anchor box quality is constantly adjusted, which means that WIoUv3 can develop a corresponding strategy according to the current situation, so as to realize the dynamic allocation of gradient gain. WIoUv3 implements an adaptive gradient modulation strategy through its dynamic non-monotonic allocation framework, where backpropagation forces are adaptively scaled based on anchor quality distribution statistics. This attention-aware regularization prioritizes learning from normative-quality proposals while maintaining gradient diversity, thereby enhancing model robustness across heterogeneous detection scenarios.

III. Experiment and Analysis

A. Dataset Construction

In this paper, a new dataset proposed by Nicolai Hani[25] et al. is used to achieve direct comparison through a large number of high-resolution images acquired in an orchard, as well as manual annotation of the fruits on the trees. The dataset employs instance-aware polygonal segmentation masks with topological preservation to enable pixel-accurate localization, simultaneously supporting three granularity levels: object detection, instance segmentation, and cluster density estimation. Comprising 1,000 high-resolution agricultural samples with 41,280 manually verified annotations, our benchmark provides: 1) Baseline performance metrics (mAP@0.5 for detection, Mask IoU for segmentation), 2) Fruit cluster distribution statistics across illumination conditions, and 3) Yield prediction models validated through cross-block regression analysis (RMSE=3.8 fruits/cluster) A sample of the dataset is shown in Figure 5.

In this experiment, the data set is divided into two parts: training set and validation set according to the ratio of 8:2. 4000 samples in the training set and 1000 samples in the validation set are selected, with a total of 5000 samples of one type of target. The data set was

Fig. 5: Example dataset

processed according to the labeling format of YOLOv8. The environment configuration used is Windows10 operating system, RTX3060 graphics card, 4-core CPU, 32G memory, 1000G system disk, CUDA version 2.0.1, Python language environment 3.10. The total number of iterations of the experiment is 300, and the batch size is set to 24. Finally, each header is 256,512,1024. The evaluation metrics are P, R, mAP50 and the number of parameters commonly used in object detection tasks. Where P represents the average accuracy of model detection results, R represents the model recall rate, and mAP50[26] represents the mean and average precision of the model. The larger the value of mAP50, the better the overall performance of the model detection data, and the smaller the number of parameters, the smaller the model size. In order to fully verify the effectiveness of each module in the experimental operation, multiple ablation experimental analysis and comparison experiments were carried out, and the experimental results of adding CBAM fusion attention mechanism and feature fusion module were compared respectively.

B. Model evaluation metrics

According to the real-time and accurate detection requirements of apple fruits, frames per second (FPS) is used to evaluate the detection speed of the model. Floating-point operations (FLOPS) and model parameters (Params) were used to evaluate the embedded deployment ability of the model. Precision (P), Recall (R), mean Average Precision (mAP) were used to evaluate the detection accuracy. Generally, the mAP value is calculated under the threshold condition of IoU=0.5:0.95, that is, mAP_0.50, which is defined as follows.

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

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

$$AP = \int_0^1 P(R)dR \tag{8}$$

$$mAP = \frac{\sum AP}{N} \tag{9}$$

In the evaluation framework, critical metrics are formally defined as.

- **True Positive (TP)**: Instances where positive-class specimens are accurately classified
- **False Positive (FP)**: Background elements erroneously assigned positive labels
- **False Negative (FN)**: Target objects missed by the detection system
- **Average Precision (AP)**: Computed through precision-recall curve integration

C. Results and Analysis

1) Analysis of ablation experiment results: In the ablation experiment part, in order to better study the influence of different modules on the performance of YOLOv8 model, the ablation experiments of two modules were carried out, and CBMA, DCNv2, MHSA and WIOU were selected for ablation respectively. CBMA and MHSA were summarized as the attention mechanism module. DCNv2 and WIOU were combined as feature fusion modules for testing respectively.

In Table 1 (all table and owned by the authors), we can see that the improved model has a precision of 0.854 and a recall of 0.812. It is worth noting that after adding the DCNv2 module, the accuracy of the model is reduced, which is because the DCNv2 module is mainly used to reduce the missed detection rate of the model. We believe that the loss of accuracy here is only 0.006, but it can better balance the speed and accuracy, so this module is retained for subsequent experiments.

2) Comparative Experiments: In the comparative test evaluation of the model, we selected YOLOv3, fastrcnn, YOLOv5, SSD, YOLOv7, and YOLOv8 for comparative tests to evaluate Precision, Recall, and Map respectively. The number of Lable is 935, and the selected image size is 640*640. The experimental results are shown in Table 2.

In Table 2, the Map value is analyzed, the best performance is APP-YOLOv8 proposed in this paper, the Map value is 0.858, which is 7.2%higher than that of the original model YOLOv8, and the worst performance is FastRCNN, whose Map value is only 0.623. This is because FastRCNN, as a classical two-stage model, is not sensitive to the occlusion of the Apple dataset and the imbalance of the samples. In the evaluation of Recall and Precision, APP-YOLOv8 still maintains good performance compared with other algorithms. Especially in the evaluation of Precision, App-YOLOv8 is 8.6% higher

TABLE I: Comparison of Ablation Study

	YOLOv8	CBAM	MHSA	DCNv2	WIOU	P(%)	R(%)
YOLOv8-1	✓					0.768	0.794
YOLOv8-2	✓	✓				0.818	0.774
YOLOv8-3	✓	✓	✓			0.825	0.750
YOLOv8-4	✓			✓		0.802	0.815
YOLOv8-5	✓			✓	✓	0.796	0.811
YOLOv8-6	✓	✓	✓	✓		0.842	0.798
YOLOv8-7(ours)	✓	✓	✓	✓	✓	0.854	0.812

Fig. 6: Comparison of complex background detection effects

TABLE II: Comparative experimental evaluation

	Labels	Size	P	R	Map@.5
YOLOv3	935	640*640	0.602	0.731	0.678
Fast RCNN	935	640*640	0.613	0.611	0.623
YOLOv4	935	640*640	0.532	0.521	0.536
YOLOv5	935	640*640	0.702	0.714	0.723
SSD	935	640*640	0.621	0.745	0.692
YOLOv7	935	640*640	0.734	0.769	0.721
YOLOv8	935	640*640	0.768	0.794	0.786
APP-YOLOv8(ours)	935	640*640	0.854	0.812	0.858

TABLE III: Model parameters and inference speed

	Parameters/kb	FPS
YOLOv3	120519	39
YOLOv4	156201	58
YOLOv5	104106	47
YOLOv8	115231	79
APP-YOLOv8(Ours)	100251	93

than the original model, but the recall rate is not greatly improved. Nevertheless, the validity of the model can still be proved.Finally, we showed a graphic representation of the training process in detail, the loss function and the optimizer to ensure that backpropagation is implemented on our model architecture. The results are shown in Figure 7. Finally, it evaluated the number of parameters and inference speed based on the trained model. The parameter size and FPS value of the model are used as evaluation indicators. The experimental results show that the retention of PConv is effective, which further verifies our conjecture about balancing speed and accuracy, and the weight size after training is indeed worth the 0.006 accuracy reduction pointed out above. As can be seen in Table IV, the best realization is APP-YOLOv8, which has obvious advantages in the number of parameters and fps. According to the comparison results, YOLOv7 has higher detection accuracy than SSD, YOLOv3, YOLOv4, YOLOv5, and fast-rcnn. Fast R-CNN epitomizes classical two-stage detection methodology, decomposing the detection pipeline into sequential sub-tasks: (1) Region Proposal Network (RPN) for coarse object localization, and (2) Region-based Convolutional Network (R-CNN) for joint coordinate refinement and semantic classification. While achieving 68.9% mAP on VOC2007, its inherent computational topology - requiring serialized RoI warping and per-region FC layer processing - imposes fundamental latency constraints (About 0.5s/img on VGG16), rendering it unsuitable for real-time deployments requiring Greater than 30 FPS throughput. The mAP value of SSD algorithm is

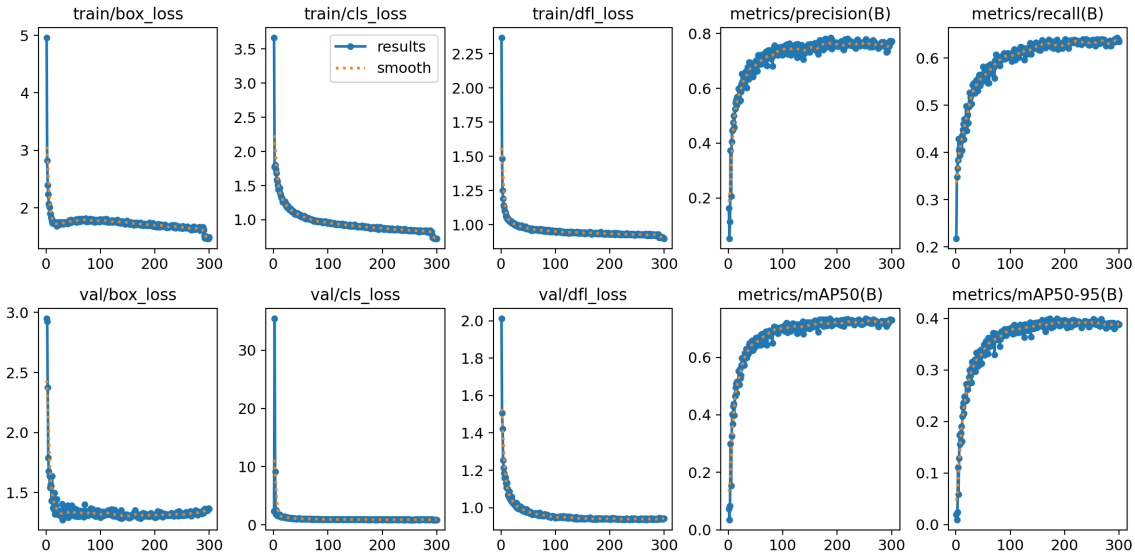

Fig. 7: Training results

69.2%, and the detection accuracy is poor in the complex background category. Compared with the SSD algorithm, the YOLOv8 algorithm has the mAP value increased by 6.0%, but the FPS and parameter number are lower than the latter. Compared with the YOLOv3 algorithm, the YOLOv4 algorithm has improved in FPS and parameter number, but the overall mAP value is only 53.6%. YOLOv5 has a lower number of parameters than the previous ones. Compared with the original YOLOv8 algorithm, the FPS of the improved algorithm reaches 93, and the mAP is increased by 7.2%.

3) Visual Result Analysis: In order to show that APP-YOLOv8 can more accurately identify apple picking targets in complex environments, the recognition effect in the case of complex background with dense branches is shown. This is shown in Figure 6.

In Figure 6, the detection effects of 16 images under complex backgrounds are shown in total, and the model still maintains a good detection effect. Especially in the fifth image, we can see that the miss rate of the model is not high. However, in the 12th image, the apple in the bottom left corner is not detected, which is caused by multiple factors such as illumination and color. The best performance is for the first image, which achieves almost 100% detection. In summary, APP-YOLOv8 still maintains good performance under various evaluations, especially in the face of complex scenes, the improved algorithm still has good detection ability.

IV. Conclusion

In this paper, an enhanced model based on YOLOv8 is proposed to improve the detection performance by integrating CBAM attention mechanism and feature fusion module. The improved model is trained and shows good detection accuracy speed on different apple samples. It not only improves the average detection accuracy of each sample class, but also significantly improves the average accuracy of imbalanced sample classes. In the actual apple picking work, detection and identification is only the first step, so the subsequent research will mainly focus on counting and apple maturity analysis.

References

[1] Kumari C U, Vignesh N A, Panigrahy A K, et al. Fungal disease in cotton leaf detection and classification using neural networks and support vector machine, *Entropy*, vol.8, no.10, pp. 3-0200073, 2019.

[2] Chitradevi B, Srimathi P. An overview on image processing techniques, *International Journal of Innovative Research in Computer and Communication Engineering*, vol. 2, no. 11, pp. 6466-6472, 2014.

[3] Triki A, Bouaziz B, Gaikwad J, et al. Deep leaf: Mask R-CNN based leaf detection and segmentation from digitized herbarium specimen images, *Pattern Recognition Letters*, vol. 150, pp. 76-83, 2021.

[4] He K, Gkioxari G, et al. Mask R-CNN, *IEEE International Conference on Computer Visionand Pattern Recognition Workshops*, pp. 2961-2969, 2017.

[5] Khan N A, Lyon O A S, Eramian M, et al. A novel technique combining image processing, plant development properties, and the Hungarian algorithm, to improve leaf detection in Maize, *IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops* pp. 74-75, 2020.

[6] Chen L. Topological structure in visual perception, *Science*, vol. 218, no. 4573, pp. 699-700, 1982.

[7] WANG Yunlu, Wu Jiefang, LAN Peng, et al. Apple leaf disease recognition method based on improved Faster R-CNN, *Journal of Forestry Engineering*, vol. 7,no. 1, pp. 153-159, 2021.

[8] Li Xinran, Li Shuqin, Liu Bin. Apple leaf disease detection model based on improved Faster R-CNN, *Computer Engineering*, vol. 47, no. 11, pp. 298-304, 2021.

[9] BAO Wenxia, Wu Gang, Hu Gensheng, et al. Apple leaf disease recognition based on improved convolutional neural network, Journal of Anhui University (Natural Science Edition), vol. 45, no. 1, pp. 53-59, 2021.

[10] Yongzhong Fu, Liang Qiu, Xiao Kong, and Haifu Xu, Deep Learning-Based Online Surface Defect Detection Method for Door Trim Panel, *Engineering Letters*, vol. 32, no. 5, pp. 939-948, 2024.

[11] Shang Fengnan, ZHOU Xuecheng, Liang Yingkai, et al. Dragon fruit detection method in natural environment based on improved YOLOX, *Intelligent Agriculture: Chinese & English*, vol. 4, no. 3, pp. 120-131, 2012.

[12] Kang Tan, Linna Li, and Qiongdan Huang, Image Manipulation Detection Using the Attention Mechanism and Faster R-CNN, *IAENG International Journal of Computer Science*, vol. 50, no. 4, pp. 1261-1268, 2023.

[13] FAN Xiangpeng, XU Yan, ZHOU Jianping, et al. Grape leaf disease detection system based on transfer learning and improved CNN, *Transactions of the Chinese Society of Agricultural Engineering*, vol. 37, no. 6, pp. 151-159, 2021.

[14] ZHAO Lixin, HOU Fadong, LV Zhengchao, et al. Image recognition of cotton leaf diseases and pests based on transfer learning, *Transactions of the Chinese Society of Agricultural Engineering*, vol. 36, no. 7, pp. 184-191, 2020.

[15] Talaat F M, ZainEldin H. An improved fire detection approach based on YOLO-v8 for smart cities, *Neural Computing and Applications*, vol. 35, no.28, pp. 20939-20954, 2023.

[16] Xu W, Fu Y L, Zhu D. ResNet and its application to medical image processing: Research progress and challenges, *Computer Methods and Programs in Biomedicine*, 2023, vol.240, pp.107660.
[17] Zhang J, Zhang J, Zhou K, et al. An improved YOLOv5-based underwater object-detection framework, *Sensors*, vol. 23, no. 7, pp. 3693, 2023.
[18] Jeong J, Do J, Kang S M. Polydopamine-Mediated, Amphiphilic Poly (Carboxybetaine Methacrylamide-r-Trifluoroethyl Methacrylate) Coating with Resistance to Marine Diatom Adhesion and Silt Adsorption, *Advanced Materials Interfaces*, vol. 11, no.11, pp. 2300871, 2024.
[19] Lu J, Yang J. Object detection in urban traffic scenarios based on improved YOLOv8 model, *Third International Symposium on Computer Applications and Information Systems (ISCAIS 2024)*. SPIE, 2024, vol.13210, pp.403-409.
[20] Hu D, Yu M, Wu X, et al. DGW-YOLOv8: A small insulator target detection algorithm based on deformable attention backbone and WIoU loss function, *IET Image Processing*, vol. 18, no. 4, pp. 1096-1108, 2024.
[21] Li Y, Feng Q, Liu C, et al. MTA-YOLACT: Multitask-aware network on fruit bunch identification for cherry tomato robotic harvesting, *European Journal of Agronomy*, vol. 146, pp. 126812, 2023.
[22] Dina A S, Siddique A B, Manivannan D. A deep learning approach for intrusion detection in Internet of Things using focal loss function, *Internet of Things*, vol. 22, pp. 100699, 2023.
[23] Li P, Zheng J, Li P, et al. Tomato maturity detection and counting model based on MHSA-YOLOv8, *Sensors*, vol. 23, no. 15, pp. 6701, 2023.
[24] Zhigang L, Baoshan S, Kaiyu B. Optimization of YOLOv7 Based on PConv, SE Attention and Wise-IoU, *International Journal of Computational Intelligence and Applications*, vol. 23, no. 01, pp. 2350033, 2024.
[25] Roy P, Isler V. MinneApple: a benchmark dataset for apple detection and segmentation, *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 852-858, 2020.
[26] Renuka O. A YOLOv8-based approach for multi-class traffic sign detection, *International Journal of Science and Research Archive*, vol. 11, no. 2, pp. 824-829, 2024.