APPLE FRUIT RECOGNITION METHOD BASED ON IMPROVED YOLOv5 / 基于改进 YOLOV5 的苹果果实识别方法

Yang BAI[1)], Shengqiao XIE[2)], Jian SONG*[1)], Cunyu ZHAO[1)], Fuxiang XIE[1)]
[1)]School of Machinery and Automation, Weifang University, Weifang/China;
[2)] China National Heavy Duty Truck, Jinan/China
Tel: 13256361611; E-mail: 20210025@wfu.edu.cn
Correspondent author: Jian Song
DOI: https://doi.org/10.35633/inmateh-75-81

Keywords: *YOLOv5; Apple; Attention mechanism; Apple fruit recognition*

ABSTRACT

This study addressed the practical problems of complex picking environments, difficult image recognition, and low picking efficiency in apple harvesting, combined with China's agricultural requirements and picking systems. An improved apple fruit recognition method based on attention mechanisms and YOLOv5 was proposed. A dataset was created by collecting 3,600 apple images under front-light, side-light, and backlight conditions at different coloring stages in natural environments. The SENet and CBAM attention mechanisms were used to enhance YOLOv5's feature extraction network, and the model was trained to improve detection accuracy. Experimental verification showed that the YOLOv5x model embedded with the CBAM module achieved the highest mean average precision (mAP) of 98.3%. The CBAM module outperformed the SENet module. Actual tests of the apple-picking robot's vision system prototype showed that when the IOU threshold was set at 0.5 and 0.3, the average detection accuracy was over 85% in both cases. The results demonstrated that the improved YOLOv5 model exhibited robustness to light intensity variations. This approach provides a technical reference for developing apple picking robot vision systems.

摘要

针对苹果采摘存在采摘环境复杂、图像准确识别困难、采摘效率低下等实际问题，结合我国苹果采摘农艺要求及采摘体系。本文提出了一种基于注意机制和改进的 YOLOv5 的苹果果实识别方法。该方法通过收集 3600 张自然环境中顺光、侧光和背光的不同着色天数的苹果图像，创建了一个数据集，注意机制 SENet 和 CBAM 用于改进 YOLOv5 的特征提取网络，并对模型进行训练以提高模型的检测精度。经过实验验证，嵌入 CBAM 模块的 YOLOv5x 的平均检测精度最高，mAP 为：98.3%。CBAM 模块的性能优于 SENet 模块。结果表明：改进的 YOLOv5 模型对光强变化具有良好的鲁棒性。通过采摘机器人视觉识别系统样机的实际试验验证，当 IOU 阈值设为 0.5 和 0.3 时，该系统样机平均检测精度均在 85%以上。改进后的 YOLOv5 模型可为苹果采摘机器人视觉系统的开发提供参考。

INTRODUCTION

China is the largest producer and consumer of apples, with a large cultivation area and high yield. However, the mechanization level of apple harvesting is relatively low (*Lu et al., 2020*). Apple-picking robots can improve harvesting efficiency and save costs, but the complex working environment and various uncertain factors make apple picking challenging. Rapid detection of fruits in complex natural environments is the primary task in research on apple-picking robots (*Wang et al., 2021*). Only when the visual system recognizes the target can the robotic arm and end effector be driven to complete fruit picking (*Tian et al., 2020; Liu et al., 2019; Zhao et al., 2021*). The accuracy of fruit recognition directly affects the efficiency and quality of robot harvesting.

Traditional target recognition methods for apple-picking robots mainly rely on the color and grayscale threshold of the fruits (*Felzenszwalb et al., 2010*). They require manual extraction of target features and are greatly influenced by natural light intensity, as well as branch and leaf occlusion. The robustness of the algorithm is poor. Target detection based on machine learning requires predetermined parameters, and the parameter size significantly affects the classification results.

Currently, deep learning-based detection algorithms are divided into single-stage algorithms and two-stage algorithms based on the different ways of predicting target categories and positions (*Luo et al., 2020*).

Since 2014, *Girshick et al., (2016)*, have successively proposed a series of two-stage algorithms, such as R-CNN, Fast R-CNN, Faster R-CNN, and Mask R-CNN.

The calculation area generation in the two-stage detection algorithm improves the detection accuracy, but increases the calculation amount, training difficulty and parameter quantity of the algorithm, reduces the prediction speed of the algorithm, and is difficult to meet the real-time detection requirements. To solve the problem of algorithm efficiency, *Redmon J et al., (2016),* have successively proposed single-stage object detection algorithms represented by the YOLO series since 2016. Single-stage algorithms regress the position and class of bounding boxes in a single output layer, resulting in a significant improvement in detection speed compared to two-stage algorithms. Among them, the YOLOv5 algorithm is currently a more advanced object detection algorithm, surpassing YOLOv4 in both detection accuracy and speed. Therefore, this model has a significant advantage in hardware deployment and is suitable for rapid detection of apples in natural environments.

In this study, an improved YOLOv5 object detection method based on attention mechanisms is proposed to optimize the apple fruit detection process. The goal is to overcome errors and influences caused by environmental factors and enhance the detection performance of the visual system for fruits.

MATERIALS AND METHODS

Data acquisition and pre-processing

This study focuses on the Red Fuji apple in Yantai. The data collection was conducted at the Le Feng apple plantation in Linqu, Weifang. Figure 1 shows images of apples at different stages of maturity. Collecting images of apples at different coloring stages facilitates testing the impact of different maturity stages on the detection performance of the network model.

1 2 3

Fig. 1 – Apple images collected at different coloring periods

1. Coloring for 5 days; 2. Coloring for 10 days; 3. Coloring for 15 days

1 2 3

Fig. 2 – Images of an apple under different lighting angles

1. Front-lighting; 2. Side-lighting; 3. Backlighting

To ensure the diversity of collected image samples, apple images were captured under different weather conditions (sunny and cloudy) at the time range of 8:00-17:00. The captured images were taken in three lighting modes: front-light, side-light, and backlight. Figure 2 shows apple images captured under different lighting angles.

Data pre-processing

The labelImg software was used to annotate the apple fruits in the collected 3600 images. Figure 3 shows the distribution of sample attributes in the dataset.

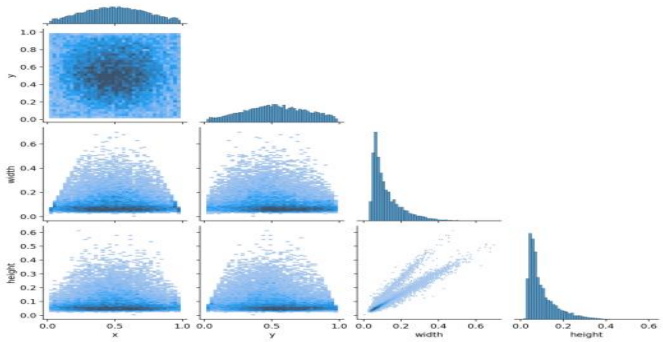


Fig. 3 – Distribution of sample attributes in the dataset

To ensure the randomness and rationality of dataset partition, manual partitioning was conducted on the collected apple images in this study. A total of 300 images were randomly selected from the original dataset to create Test Set 1, which consists of 100 images taken in front-lighting conditions, side-lighting conditions, and backlighting conditions, respectively. Additionally, Test Set 2 was created by randomly selecting 100 images of apples that were 5 days, 10 days, and 15 days old, respectively, while the remaining images were used as training data for the model. The number of samples in each dataset and the corresponding number of target boxes are shown in Table 1.

Table 1

Factors and level of orthogonal test

Datasheet	Training sets		Test set 1		Test set 2	
	Number of pictures	Number of target boxes	Number of pictures	Number of target boxes	Number of pictures	Number of target boxes
Coloring for 5 days	1000	12038	100	1314	100	1328
Coloring for 10 days	1000	13117	100	1248	100	1196
Coloring for 15 days	1000	12864	100	1293	100	1159
Grand total	3000	38019	300	3855	300	3683

The 3600 original apple images collected in this study are not sufficient to cover all the factors such as lighting intensity, weather, noise, and clarity in natural environments. Therefore, data augmentation is performed on the original images to enhance the generalization ability of the object detection model and prevent overfitting. This study mainly adds random brightness, random contrast, random Gaussian noise, random saturation, and random flipping to the collected apple images to simulate various states of fruit trees in natural environments as much as possible. The augmented apple images are five times more than the original data. Figure 4 shows some examples of the apple images after augmentation.

1

2

3

4 5 6

Fig. 4 – Examples of image augmentation

1.Original Image; 2. Random Brightness; 3. Random Contrast; 4. Random Gaussian noise; 5. Random saturation; 6. Random rotation

Construct an apple recognition model based on improved YOLOv5

The YOLOv5 algorithm added modules such as image compression and Mosaic data augmentation at its input end (*Yonghui et al., 2022*). Mosaic utilized random scaling, cropping, and arrangement of four images to generate a new image, achieving data augmentation (*Chen et al., 2022*).

There were four different versions of the YOLOv5 algorithm: YOLOv5x, YOLOv5l, YOLOv5m, and YOLOv5s. The structural principles of the YOLOv5 versions were similar, with differences only in network width and depth (*Liu et al., 2020, Yongpeng et al., 2024*). The network structure of the YOLOv5 algorithm mainly consisted of modules such as Focus, CBL, CSP_1, CSP_2, and SPP, as shown in Figure 5.

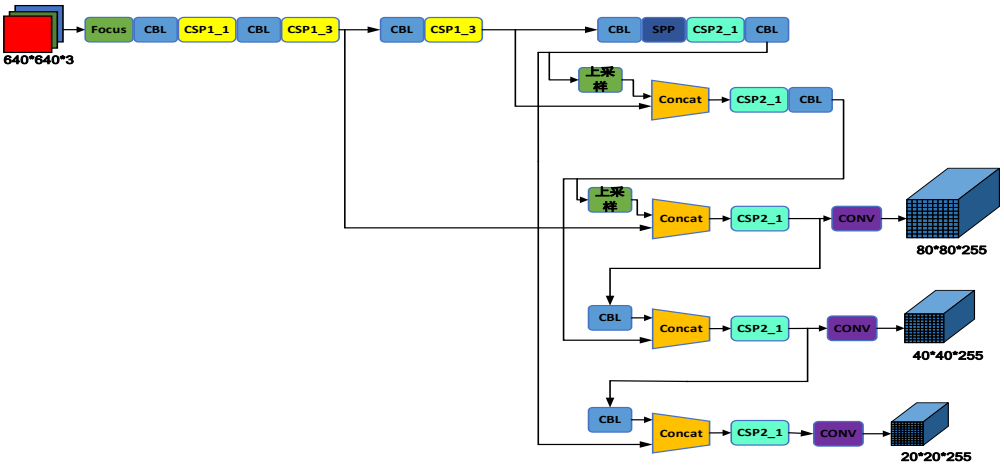


Fig. 5 – The network structure of YOLOv5

The CBL module consists of Convolutional layers, Batch Normalization (BN) layers, and Leaky ReLU (LR) layers. The CBL module plays a role in downsampling and helps reduce information loss in the downsampling process. The computational cost of the Focus module in YOLOv5 is approximately three times higher than that of downsampling with convolution, but it helps reduce information loss in the downsampling process.

The feature extraction network in the YOLOv5 algorithm consists of various functional modules such as Focus, CBL, CSP, and SPP. The deep network layers ensure the network's ability to extract features. The YOLOv5 algorithm utilizes convolutional operations to perform feature re-extraction on the output feature maps. The algorithm takes in input image data of size 640x640x3 and outputs three feature maps of sizes 80x80x18, 40x40x18, and 20x20x18, respectively. The feature map with a higher resolution is used for predicting smaller objects, while the ones with lower resolutions are used for predicting larger objects. Using feature maps of different resolutions helps improve the accuracy of object recognition for objects of different sizes.

Attention mechanism

The introduction of attention mechanisms in computer vision can effectively enhance the feature extraction capability of networks, thereby improving the accuracy of object recognition. Currently, widely used attention mechanisms in deep learning include SENet and CBAM.

SENet attention mechanism

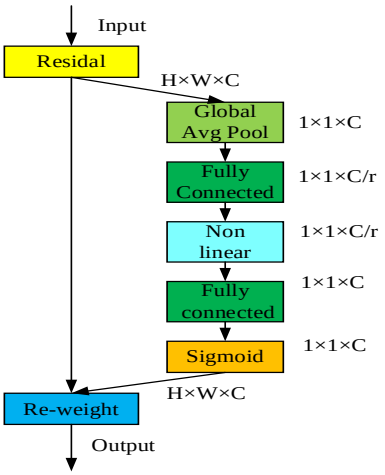


Fig. 6 – Schematic diagram of the seNet attention mechanism structure

Figure 6 shows the structure of the SENet network. The SENet takes an H×W×C feature map as input, with C channels (*Bai et al., 2022; Lin et al., 2021; Peng et al., 2022*). First, the input feature map is globally average pooled, reducing the height and width of the feature map to 1×1, as shown in Eq.(1). Then, two fully connected layers are applied, followed by the sigmoid activation function to normalize the values. This process obtains the weights for each channel in the input feature map. By multiplying these weights with the input feature map, a new calibration of the input feature map using channel attention is achieved, as shown in Eq.(2).

$$z_c = F_{sq}(u_c) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} u_c(i, j) \tag{1}$$

$$\overset{\Lambda}{X} = X \cdot \sigma(\overset{\Lambda}{z}) \tag{2}$$

z_c: the c-th element of z;

H: the height of the feature map;

W: the width of the feature map;

$u_c(i, j)$: the (i, j)-th element of the c-th channel of u;

$\sigma(\hat{z})$: channel weight.

After passing through the first fully connected layer, the dimension of the feature map decreases, significantly reducing the model's parameters and computational complexity. After passing through the second fully connected layer, the dimension of the feature map is restored to the same as the input, establishing correlations between channels. This increases the weights of effective feature map channels and decreases the weights of other feature map channels, allowing the model to achieve better training performance.

CBAM attention mechanism

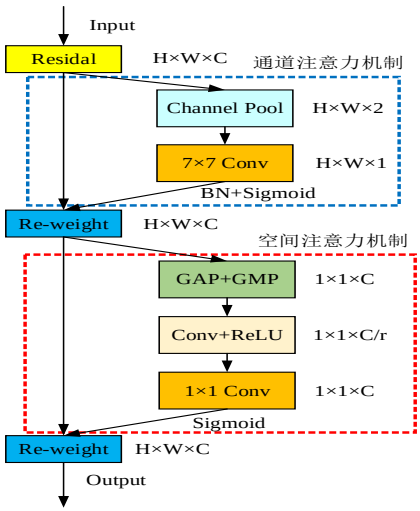


Fig. 7 – Schematic diagram of the structure of the CBAM attention mechanism

The structure shown in Figure 7 is the Convolutional Block Attention Module (CBAM), which consists of a Channel Attention Module and a Spatial Attention Module (*Liu et al., 2021; Xia et al., 2023; Huang et al., 2021)*. The input feature map first goes through the Channel Attention Module, where the feature map's height and width are globally averaged and globally max-pooled. The resulting values are then passed through a Multi-Layer Perceptron (MLP) to obtain the channel attention weights. These weights are then added to the original input feature map through a multiplication and addition operation. This completes the calibration of the original input feature map using channel attention, as shown in Equation·(3).

$$M_c(F) = (MLP(AvgPool(F)) + MLP(MaxPool(F))) = \sigma(W_1(W_0(F_{avg}^c)) + W_1(W_0(F_{\max}^c))) \quad (3)$$

σ: Sigmoid function;

C: Number of channels;

W_1, W_0: Input shared weights;

F_{avg}^c: Feature mapping generated by average pooling in space;

F_{max}^c: Feature mapping generated by maximum pooling in space.

After passing through the CBAM module, the feature map generates new feature maps with channel and spatial attention weights, enhancing the relationship between image features in both spatial and channel dimensions. This is beneficial for extracting typical features of the target (*Su et al., 2021, Lingqing et al., 2024*).

Loss function

In the target detection task, the loss function can better reflect the gap between the predicted value and the real value of the data, and then reflect the detection effect. The loss function used in this paper consists of three parts, they are Loss of confidence (L_{con}),Positioning loss (L_{GIOU}) and Classification loss (L_{class}). The total loss function (L_{total}) can be obtained by accumulating the three, as shown in Equation (4).

$$L_{total} = L_{con} + L_{GIOU} + L_{class} \quad (4)$$

In the formula, the confidence loss measures the confidence level of the prediction box, with the calculation shown in Equation (5). The calculation method of the function is cross entropy error, which determines whether the predicted bounding boxes contain the predicted target. If there is a predicted target in the current bounding box, the value of I_{ij}^{obj} is 1. If there is no predicted target in the current bounding box, the value of I_{ij}^{obj} is 0. The confidence values are weighted and summed to obtain the value of L_{con}.

$$\mathrm{L_{con}} = \lambda_{\mathrm{obj}} \sum_{i=0}^{S^2} \sum_{j=0}^{B} I_{ij}^{obj} \left[-\hat{C}_i \ln C_i - \left(1 - \hat{C}_i\right) \ln\left(1 - \hat{C}_i\right) \right] + \lambda_{\mathrm{no}} \sum_{i=0}^{S^2} \sum_{j=0}^{B} I_{ij}^{no} \left[-\hat{C}_i \ln C_i - \left(1 - \hat{C}_i\right) \ln\left(1 - \hat{C}_i\right) \right] \quad (5)$$

λ_{obj}: there is a target weight coefficient in the grid;

λ_{no}: no target weight coefficient in the grid;

S^2: number of grids;

I_{ij}^{obj}: to determine whether the j-th bounding box in the i-th grid needs to be predicted,

I_{ij}^{no}: to determine whether the j-th bounding box in the i th grid has a target that does not need to be predicted,

C_i: predict the target confidence value;

$\hat{C}_i$: the actual target confidence value.

L_{class} is the classification loss, which is used to calculate the difference between the predicted value and the actual value of the category. The calculation is shown in Equation (6).

The calculation of L_{class} is similar to that of L_{con}.The values of I_{ij}^{obj} and I_{ij}^{obj} are the same as shown in Equation (7). The calculated probability value is weighted and summed to obtain the value of L_{class} .

$$\mathrm{L_{class}} = \sum_{i=0}^{S^2}\sum_{j=0}^{B}\sum_{c\subset class} I_{ij}^{obj}\left[-\hat{p}_i(c) - \ln\left(\hat{p}_i(c)\right) - \left(1-\hat{p}_i(c)\right)\ln(1-p_i(c)\right] \quad (6)$$

c: Boundary box prediction category;

$p_i(c)$: The probability that the target is predicted to be c in the i-th grid;

$\hat{p}_i(c)$: The actual probability that the target is c in the i-th grid.

In object detection tasks, IoU is often used to calculate the coordinate differences between predicted bounding boxes and ground-truth bounding boxes, and can more directly reflect the detection performance of the algorithm. The calculation formula is shown in (7) ~ (9).

$$IoU = \frac{A\bigcap B}{A\bigcup B} \quad (7)$$

$$GIoU = IoU - \frac{\left|C-\left(A\bigcup B\right)\right|}{\left|C\right|} \quad (8)$$

$$L_{GIoU} = \sum_{i=0}^{S^2}\sum_{j=0}^{B}\left(1-GIoU\right) \quad (9)$$

A: Real box area;

B: Predicting box area;

C: A and B minimum circumscribed rectangle area.

RESULTS

The effect of adding attention mechanism on the test results

To verify the detection performance of the YOLOv5 algorithm after adding the attention mechanism module, a comparative experiment was conducted between the YOLOv5 algorithm with the added attention mechanism module and the original YOLOv5 algorithm. The performance was tested using a dataset composed of samples mixed from different maturity stages. Table 2 shows the experimental results of apple detection before and after adding the attention mechanism module to the YOLOv5 algorithm.

Table 2

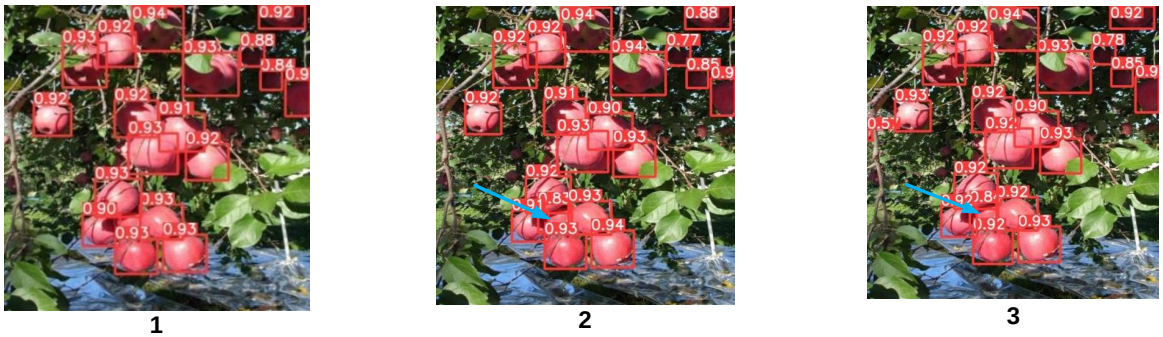
Test results of Apple detection before and after the improvement of the YOLOv5 algorithm

Algorithm	Number of network parameters (MB)	Examination speed (ms)	mAP (%)
YOLOv5s	13.7 (0.00)	31.5 (0.00)	92.3 (0.00)
YOLOv5s-SENet	14.8 (+8.03%)	31.9 (+1.26%)	93.5 (+1.30%)
YOLOv5s-CBAM	15.1 (+10.22%)	32.1 (+1.9%)	94.2 (+2.06%)
YOLOv5m	40.2 (0.00)	80.2 (0.00)	93.1 (0.00)
YOLOv5m-SENet	42.8 (+6.47%)	81.4 (+1.50%)	94.4 (+1.40)
YOLOv5m-CBAM	43.6 (+8.46%)	82.3 (+2.62%)	95.6 (+2.69%)
YOLOv5l	88.5 (0.00)	148.3 (0.00)	95.6 (0.00)
YOLOv5l-SENet	92.8 (+4.86%)	149.8 (+1.01%)	96.9 (+1.36%)
YOLOv5l-CBAM	94.2 (+6.44%)	151.8 (+2.36%)	97.8 (+2.30%)
YOLOv5x	171.6 (0.00)	304.2 (0.00)	96.4 (0.00)
YOLOv5x-SENet	179.1 (+4.37%)	308.4 (+1.38%)	97.7 (1.35%)
YOLOv5x-CBAM	181.2 (+5.59)	311.7 (+2.47%)	98.3 (+1.97%)

According to Table 2, it could be seen that in the original YOLOv5 algorithm, YOLOv5x had an average detection accuracy of 96.4%, higher than the other three networks, but it had a longer detection time.

After separately embedding the SENet attention mechanism module, the average detection accuracy of YOLOv5s increased by 1.3% to 93.5%. YOLOv5m, YOLOv5l, and YOLOv5x also showed some improvements in average detection accuracy. The parameter count of YOLOv5s-SENet increased by 8.03% compared to the original version, resulting in a 1.26% increase in detection time. Compared to the original version, the parameter count of YOLOv5m-SENet, YOLOv5l-SENet, and YOLOv5x-SENet increased by 4.37% to 6.47%, with the highest increase being only 7.5MB. However, all versions achieved more than 1% improvement in accuracy, demonstrating good detection performance.

When the CBAM module was embedded in all four different network structures of the YOLOv5 algorithm, the parameter count of YOLOv5s, YOLOv5m, YOLOv5l, and YOLOv5x increased by 10.22%, 8.46%, 6.44%, and 5.59% respectively. The lighter the network, the greater the increase in parameter count. The actual detection results of YOLOv5 model before and after adding the attention mechanism modules were shown in Figure 7.



1 2 3

Fig. 8 – YOLOv5 test effect before and after improvement

1.Unimproved YOLOv5, 2. YOLOv5+SENet, 3. YOLOv5+CBAM.

From Figure 8, it could be seen that compared with Figure (a), the occluded fruits in Figure (b) and Figure (c) were successfully recognized, indicating that the YOLOv5 network model with the added attention mechanism module had improved accuracy in identifying small and occluded targets. Figure (c) showed that the improved YOLOv5-CBAM model successfully detected the small target fruit on the left, while the other two algorithms failed to detect this target. Taking into account various factors, embedding the attention mechanism module helped improve the model's accuracy in recognizing difficult samples, and the performance of the SENet module was slightly lower than that of the CBAM module.

Impact of different lighting intensity on experimental results

In natural environments, different lighting intensities have some influence on the brightness of captured images. Images captured in direct light are clearer, images captured in side-light have some variations in lighting and shadows, while images captured in backlight are darker, all of which can affect the accuracy of fruit detection. To further validate the detection performance of the proposed attention mechanism-based improved YOLOv5 algorithm, this section conducted relevant experiments using YOLOv5s, YOLOv5s-SENet, and YOLOv5s-CBAM as examples.

The test sets were then evaluated using different weight files before and after the addition of the attention mechanism, and the experimental results were shown in Table 3.

Table 3

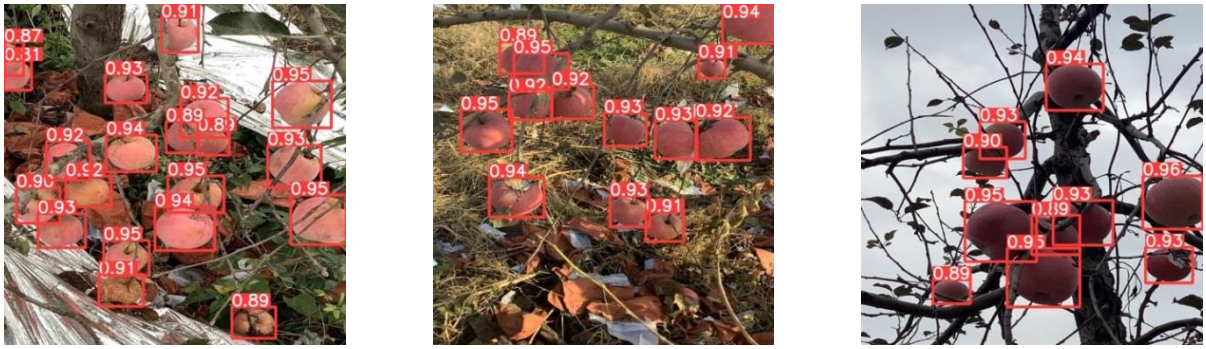
Average detection accuracy at different light intensity datasets

Data set	YOLOv5s	YOLOv5s-SENet	YOLOv5s-CBAM
Front-lighting	94.3%	95.1%	95.6%
Side-lighting	92.6%	93.7%	94.3%
Backlighting	91.4%	92.5%	92.8%

According to Table 3, the average accuracy of the original YOLOv5s algorithm on the front-light dataset was 94.3%. The YOLOv5s - SENet and YOLOv5s - CBAM algorithms achieved average accuracies of 95.1% and 95.6%, respectively.

This indicated that embedding attention mechanism modules in the YOLOv5 algorithm improved the average detection accuracy across different lighting intensities, with the YOLOv5s - CBAM algorithm achieving the best detection performance. The YOLOv5s - CBAM algorithm achieved the highest average accuracy of 95.6% on the front-light dataset, which was 1.38% higher than that of YOLOv5s. Both the original and improved YOLOv5 algorithms achieved average accuracies above 94% on the front-light dataset. On the backlight dataset, YOLOv5s - CBAM achieved an average detection accuracy of 92.8%, while YOLOv5s achieved only 91.4%, a difference of 1.4%. Furthermore, the average accuracy of YOLOv5s - CBAM on the backlight dataset was 2.93% lower than that on the front-light dataset. This also explained why the average detection accuracy of both the original and improved YOLOv5 algorithms was lowest on the backlight dataset. On the side-lighting dataset, the average accuracy of YOLOv5s, YOLOv5s - SENet, and YOLOv5s - CBAM was 92.6%, 93.7%, and 94.3%, respectively, all of which fell between the average detection accuracy of the front-lighting and backlighting datasets.

The average accuracy of YOLOv5s on datasets with different lighting intensities differed by only 2.9%, indicating that both the original and improved YOLOv5 algorithms had good robustness to changes in lighting intensity. The detection results were shown in Figure 9.



1 2 3

Fig. 9 – Detection effect at different light angles

1. Front-light; 2. Side-light; 3. Backlight

Prototype testing of a robotic target recognition system based on the proposed algorithm

As shown in Figure 9, the prototype of the vision recognition system for the picking robot was designed in this paper. The system was composed of a JETSON NANO embedded development board, a 5V4A power supply, a USB camera, and an HDMI touch display. Figure 9 presents the scene when the system prototype was functioning properly, with the recognition object being a simulated apple tree in a real - world scenario. As can be seen in Figure 9, the system could normally detect the fruit targets in the field of view, and the detection speed was stable at 30 frames per second.

Fig. 10 – Prototype of target recognition system for harvesting robot

Experimental verification showed that when the IOU threshold was set at 0.5 and 0.3, the average detection accuracy of the system prototype was over 85% in both cases, with a detection speed of 30 frames per second.

Fig. 11 – Actual testing results of Apple

CONCLUSIONS

This article proposed an apple detection method using attention mechanisms and improved YOLOv5 to facilitate picking robots' detection. Firstly, SENet and CBAM mechanisms enhanced YOLOv5's feature extraction network, improving model accuracy.

Experimental results showed that after embedding the attention mechanism module, the mAP of the YOLOv5m-CBAM model was 95.6%, which was improved by 2.69%, with an increase in detection time of 2.62%. Among the YOLOv5x models with embedded CBAM module, the average detection accuracy was the highest, with an mAP of 98.3%, and the performance of the CBAM module was superior to that of the SENet module. The improved YOLOv5s model achieved apple recognition accuracy ranging from 91.4% to 95.6% under different lighting conditions, with a difference in fruit recognition accuracy within 3% for the same model. These results demonstrated that the improved YOLOv5 model had good robustness to changes in light intensity. The improved YOLOv5 model could provide reference for the development of the visual system of apple picking robots. Actual tests of the apple-picking robot's vision system prototype showed that when the IOU threshold was set at 0.5 and 0.3, the average detection accuracy was over 85% in both cases.

REFERENCES

[1] Bai Qiang, Gao Ronghua, Zhao Chunjiang, Li Qifeng, Wang Rong, Li Shuqin (2022). Multi-scale behavior recognition method for dairy cows based on improved YOLOV5s network (基于改进 YOLOV5s 网络的奶牛多尺度行为识别方法) [J]. *Transactions of the Chinese Society of Agricultural Engineering (Transactions of the CSAE)*, Vol. 38, no.12, pp. 163-172, doi: 10.11975/j.issn.1002-6819.2022.12.019.

[2] Chen Yixiao, Alifu·Kurban, Lin Wenlong, Yuan Xu (2022).CA-YOLOv5 for Crowded Pedestrian Detection (面向拥挤行人检测的 CA-YOLOv5) [J/OL]. *Computer Engineering and Applications*: Vol. 52, no.2, pp. 1-10, doi:10.3778/j.issn.1002-8331.2201-0058

[3] Du Yonghui, Gao Ang, Song Yuepeng, Guo Jing, Ma Wei, Ren Longlong (2022). Young Apple Fruits D etection Method Based on Improved YOLOv5 [J]. *INMATEH - Agricultural Engineering*, Vol. 73 no.2, pp. 84-93, doi: https://doi.org/10.35633/inmateh-72-17.

[4] Girshick R, Donahue J, Darrell T, Malik J. (2016). "Region based convolutional networks for accurate object detection and segmentation", *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 1, pp. 142– 158;

[5] Huang Linsheng, Luo Yaowu, Yang Xiaodong, Yang Guijun, Wang Daoyong (2021). Crop Disease Recognition Based on Attention Mechanism and Multi-scale Residual Network (基于注意力机制和多尺度残差网络的农作物病害识别) [J]. *Agricultural Machinery Journal*, Vol. 52, no.10, pp. 264-271, doi:10.6041/j.issn.1000-1298.2021.10.027

[6] Lin Sen, Liu Meiyi, Tao Zhiyong (2021). Detection of underwater treasures using attention mechanism and improved YOLOv5 (采用注意力机制与改进 YOLOv5 的水下珍品检测) [J]. *Transactions of the Chinese Society of Agricultural Engineering (Transactions of the CSAE)*, Vol. 37, no.18, pp. 307-314, doi:10.11975/j.issn.1002-6819.2021.18.035.

[7] Feng L., Liu Y., Jia Z., Yang H., Guan J., Zhu H., Ho Y. (2024). Detection of Apple Leaf Diseases Target Based on Improved YOLOv7[J]. *INMATEH - Agricultural Engineering*, Vol. 72. no.1, pp. 280-290, doi: https://doi.org/10.35633/inmateh-72-26.

[8] Felzenszwalb P., Girshick R., McAllester D. (2010) Cascade object detection with deformable part models, *Computer vision and pattern recognition (CVPR)*, *IEEE conference on. IEEE*, 2010, pp. 2241–2248.

[9] Liu Fang, Liu Yukun, Lin Sen, Guo Wenzhong, Xu Fan, Zhang Bai (2020). Fast Recognition Method for Tomatos under Complex Environments Based on Improved YOLO (基于改进型 YOLO 的复杂环境下番茄果实快速识别方法) [J]. *Journal of Agricultural Machinery*, Vol. 51, no.6, pp. 229-237, doi: 10.6041/j.issn.1000-1298.2020.06.024

[10] Liu Xiaoyang, Zhao Dean, Jia Weikuan, Ruan Chengzhi, Ji Wei (2019). Fruit Segmentation Method Based on Superpixel Features for Apple Harvesting Robot (基于超像素特征的苹果采摘机器人果实分割方法) [J]. *Agricultural Machinery Journal*, Vol. 50, no.11, p.9, doi:10.6041/j.issn.1000-1298.2019.11.002.

[11] Liu Mochen, Gao Tiantian, Ma Zongxu, Song Zhanhua, Li Fade,Yan Yinfa (2021). Target detection model of corn weeds in field environment based on MSRCR algorithm and YOLOv4 tiny (基于 MSRCR YOLOv4 tiny 的田间环境玉米杂草检测模型). Transactions of the Chinese Society for Agricultural Machinery, Vol. 53, No. 2, 246-255, 335, doi:10.6041/j.issn.1000-1298.2022.02.026

[12] Lu Yiqiu, Yang Ye (2020). Study on the Development of Apple Harvesting Mechanization (苹果采摘机械化发展研究) [J]. *Southern Agricultural Machinery*, Vol. 51, no.14, pp. 2.

[13] Luo Yuan, Wang Boyu, Chen Xu (2020). Research Progresses of Target Detection Technology Based on Deep Learning (基于深度学习的目标检测技术的研究综述) [J]. *Semiconductor Optoelectronics*, Vol. 41, no.1, pp. 1-10, doi:10.16818/j.issn1001-5868.2020.01.001

[14] Peng Hongxing, Xu Huiming, Liu Huanai (2022). Model for identifying grape pests and diseases based on two-branch feature fusion and attention mechanism (融合双分支特征和注意力机制的葡萄病虫害识别模型) [J]. *Transactions of the Chinese Society of Agricultural Engineering*, Vol. 38, no.10, pp. 156-165, doi:10.11975/j.issn.1002-6819.2022.10.019

[15] Redmon J., Divvala S., Girshick R., Farhadi A. (2016). You only look once: Unified, real-time object detection, *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 779–788.

[16] Su Baofeng, Shen Lei, Chen Shan, Mi Zhiwen, Song Yuyang, Lu Nan (2021). Multi-features Identification of Grape Cultivars Based on Attention Mechanism (基于注意力机制的葡萄品种多特征分类方法) [J]. *Agricultural Machinery Journal*, Vol.52, no.11, pp.226-233, 252, doi:10.6041/j.issn.1000-1298.2021.11.024.

[17] Tian Bokai (2020). *Research on Apple Detection Classification and Location Technology in Complex Environment Based on Deep Learning* [D]. Tianjin University of Technology.

[18] Wang Fang, Cui Dandan, Li Lin (2021). Target recognition and positioning algorithm of picking robot based on deep learning (基于深度学习的采摘机器人目标识别定位算法) [J]. *Electronic measurement technology*, Vol. 44, no.20, pp.6, doi:10.19651/j.cnki.emt.2107366

[19] Xia Ye, Xiaohui Lei, Andreas Herbst, Xiaolan Lyu (2023). Research on Pear Inflorescence Recognition Based on Fusion Attention Mechanism with YOLOv5. *INMATEH-Agricultural Engineering*, Vol. 69, No.1, pp.11-20, doi: https://doi.org/10.35633/inmateh-69-01

[20] Yongpeng Chen, Yi Niu, Weidong Cheng, Laining Zheng, Dongchao SUN (2024). Apple Detection Method in the Natural Environment Based on Improved YOLOv5[J]. *INMATEH - Agricultural Engineering*, Vol.72, no.1, pp. 183-192, doi: https://doi.org/10.35633/inmateh-72-17

[21] Zhao Jun'ai (2021). Design and Experiment of Apple Picking Robot Based on Machine Vision (基于机器视觉的苹果采摘机器人的设计与试验) [J]. *Henan Science and Technology*, Vol. 0, no.20, pp.3, doi: 10.3969/j.issn.1003-5168.2021.20.017