AC-YOLO: citrus detection in the natural environment of orchards

Xu Xiao,[1] Yaonan Wang,[2] Yiming Jiang,[2] Haotian Wu,[2] Zhe Zhang,[2] Rujing Wang[3]

[1]College of Electrical and Information Engineering, Hunan University, Changsha; [2]National Engineering Laboratory for Robot Visual Perception and Control Technology, Hunan University, Changsha; [3]Anhui key Laboratory of Bionic Sensing and Advanced Robot Technology, Hefei, China

Abstract

In the natural environment, the shape and color of fruits can vary greatly due to various factors, and the growth of fruits is irregular, shaded by leaves and branches, and there are phenomena such as overlapping fruits. The complex background causes the difficulty of fruit recognition by the picking robot to increase, which affects the positioning of subsequent picking points, greatly increasing the difficulty of picking, and even causing damage to the end effector. To address such issues, this study adopts panoramic photography to capture images of citrus fruit trees, and proposes an AC-YOLO based citrus recognition method in the natural environment of orchards. Firstly, in the Resblock module of the YOLOv4 backbone feature extraction network, the AC network structure is integrated with different levels of feature mapping to fuse context information as small targets. At the same time, a self-attention mechanism is introduced to suppress the impact of complex backgrounds and underlying noise, improving the detection ability of small target citrus; Finally, the Mish activation function is used to replace Leaky Re LU, improving the generalization ability of the model and improving the accuracy of citrus detection. The panorama image is divided into sub images, and an improved YOLOV4 model is used for recognition. By comparing the orange recognition effects of different network models such as Fast R-CNN, Center Net, YOLOV4 series algorithms, and YOLOV5 series algorithms on the panorama image, the improved YOLOV4 network model has an accuracy rate of 96.19%, a recall rate of 95.47%, and an average accuracy of 97.27%, Compared with the original YOLOv4 model, it has increased by 1.07, 2.59, and 2.02 percentage points respectively. This method has a good recognition effect for citrus in the natural environment of orchards.

Correspondence: Yaonan Wang, College of Electrical and Information Engineering and the National Engineering Research Center for Robot Visual Perception and Control Technology, Hunan University, Changsha 410082, China.
E-mail: yaonanwang2022@163.com

Key words: citrus testing; YOLO; AC network; non maximum fusion.

Contributions: all the authors made a substantive intellectual contribution, read and approved the final version of the manuscript and agreed to be accountable for all aspects of the work.

Availability of data and materials: all data and related metadata underlying the findings reported in a submitted manuscript should be deposited in an appropriate public repository, unless already provided as part of the submitted article.

Acknowledgments: this work is partially supported by the National Key Research and Development Program of China (Grant Nos. 2022YFB3303801), the National Natural Science Foundation of China (Grant Nos. 62293511, 62293515, 62303172), the National Natural Science Foundation of Hunan Province (Grant Nos.2023JJ40179), and China Postdoctoral Science Foundation (Grant Nos. 2022M721097). It is also supported by China Scholarship Council (CSC NO. 201806130027).

Received: 12 December 2023.
Accepted: 13 July 2024.

Journal of Agricultural Engineering 2024; LV:1654
doi:10.4081/jae.2024.1654

Publisher's note: all claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article or claim that may be made by its manufacturer is not guaranteed or endorsed by the publisher.

Introduction

Citrus is the largest type of fruit in the world, with a long history of cultivation (Guo *et al.*, 2018; Vasconez *et al.*, 2019; Lv *et al.*, 2022). The production of citrus requires a large amount of manual work, and with the development of urbanization, issues such as increased labor costs and low operational efficiency have brought a certain degree of impact on the development of the citrus industry (Bai *et al.*, 2023). Automated picking robots can improve picking efficiency, reduce damage to fruits, and reduce labor costs (Ju *et al.*, 2022). However, the operating environment of the harvesting robot is complex and there are many uncertain factors, making it difficult to harvest. Efficient and efficient fruit and vegetable picking requires accurate target recognition and three-dimensional positioning support (Barbashov *et al.*, 2022). The vision system of the harvesting robot runs through four stages, namely, target detection, target recognition, three-dimensional reconstruction, and three-dimensional positioning (Chen *et al.*, 2022; Hannan *et al.*, 2009; Nehme *et al.*, 2021; Xu *et al.*, 2022). The accuracy of target recognition and positioning will directly determine the harvesting efficiency of the harvesting robot, whether crops will be damaged, and whether the harvesting robot body will be damaged due to collision (Zhuang *et al.*, 2018).

Currently, researchers at home and abroad have conducted a large amount of research on the recognition of fruit trees (Jana *et al.*, 2017). The main method for fruit recognition is visual recognition, or image recognition, which uses image acquisition equipment to collect images, and then classifies and recognizes the images (Mai *et al.*, 2020). Traditional image recognition mainly involves manually extracting features from images, and then performing fruit recognition (Wan *et al.*, 2020). With the development of computer technology, image recognition based on machine learning and deep convolution neural networks has also been widely used in fruit detection.

Traditional visual fruit recognition: RGB images collected by image acquisition devices include main features such as color, shape, and texture. Early research generally focused on extracting single features for fruit recognition. Liu *et al.* (2024a) extracted an area where a detection target may exist based on the color characteristics of the identified target, and then fitted a multiple elliptical boundary model within the area. The fitted result was used as the recognition result to achieve the recognition of overlapping oranges, with an accuracy rate of 90.8% and an error detection rate of 11.2%, respectively.

Fruit recognition based on machine learning: the above manual feature extraction methods are time-consuming and laborious and require a large number of feature combination experiments to obtain the best results. Since the 1980s, machine learning has gained rapid development. Many scholars have integrated machine learning theory with fruit recognition technology, mainly using SVM, Canny, HOG, and other methods for feature extraction to achieve fruit recognition. Wang *et al.* (2023a) used the Canny edge detection operator to extract all contour information from the image, and then constructed a three-dimensional convex object contour function based on the three-dimensional characteristics of the apple. The contour function filtered all contour information to identify the apple, achieving an accuracy of 94%. Xiao *et al.* (2022) established an adaptive red-blue color map (ARB) and used absolute transform difference sum (SATD) block matching methods to identify potential fruit pixels. Then, based on texture features, a support vector machine classifier was established to identify immature green fruits, achieving a recognition rate of 83.4% (Bac *et al.*, 2016). Wang *et al.* (2022) performed clustering analysis on the image to extract potential target regions, then used the normalized cut algorithm to extract target contour information, and then used an interpolation algorithm to reconstruct the contour. This solved the problem of difficulty in distinguishing overlapping fruits, achieving a coincidence ratio of 93.81% between the identified region and the original target region under occlusion.

Fruit recognition based on convolutional neural networks: in 2012, Krizhevsky and colleagues proposed the AlexNet convolutional neural network for the first time, and its excellent performance quickly attracted scholars' interest (Krizhevsky *et al.*, 2012). With the development of convolutional neural networks and the popularity of computer hardware, convolutional neural networks have evolved from being able to only classify simple images at first to being able to identify various objects (Fang and Liang, 2022; Alaaudeen *et al.*, 2024). More and more scholars have also used convolutional neural networks to identify orchard crops. Wan and Goudos (2020) have optimized the structures of convolution and pooling layers in the Faster R-CNN model to make recognition more accurate and faster. Compared with traditional recognition methods, this algorithm has higher detection accuracy and lower processing time, and has achieved an average accuracy rate of 91% for detecting multiple types of fruit on a self-made fruit dataset (Alaaudeen *et al.*, 2024).

Currently, fruit tree fruit recognition is mainly based on image recognition of local scenes and simple scenes (Liu *et al.*, 2024a). However, for most domestic orchards, local scene recognition and simple scene recognition cannot meet the needs of precision gardening. This paper proposes an AC-YOLO detection model for citrus fruits in the natural environment of orchards. Aiming at the problem that medium and small target citrus fruits are prone to miss detection, an AC network structure is proposed, which fuses different levels of feature maps as context information for small targets. At the same time, a self-attention mechanism is introduced to suppress the impact of complex backgrounds and underlying noise, improving the detection ability of small target citrus fruits; Finally, the Mish activation function is used to replace Leaky Re LU, improving the generalization ability of the model and improving the accuracy of citrus detection. Experimental results show that the improved algorithm has better detection performance than the original algorithm on citrus dataset.

Methods

Citrus data set

The selection of training dataset and image calibration are two crucial steps in the target detection process. Due to the lack of publicly available general citrus datasets, this paper constructs and expands the g-citrus dataset to train and validate the citrus fruit detection algorithm. Considering the influencing factors of citrus fruit detection, 2300 citrus images were taken and collected from different time, different environment, citrus shading and citrus shape perspectives in this paper. Some examples of the citrus dataset are given in the following figure. When training convolutional neural networks, too small data sets can cause overfitting of the model, resulting in weakened generalization ability of the model, which is excellent in the training set but poor in the test set. Appropriate data enhancement methods can enrich the images of the data set and rationalize the data distribution, which can well avoid the model overfitting problem and enhance the generalization performance of the model.

At the same time, considering the influence of complex outdoor environment and illumination, this paper uses image transformation methods such as random brightness adjustment, random rotation, random contrast adjustment and random cropping to expand the data set, and divides the 2300 original images into 2000 training sets, 2300 original images into 2000 training sets and 300 verification sets. After data enhancement and expansion, the training set is 6000 images and the test set is 900 images (Figure 1).

AC-YOLO algorithm

This paper mainly studies the detection of citrus fruits in natural environments, which is the basic work for realizing automatic citrus fruit harvesting. A citrus picking robot collects images of citrus fruits through an airborne binocular camera, preprocesses the images to obtain 640 * 640 images, and extracts features using a Darknet convolutional neural network. In order to solve the complex problems in citrus detection, AC network is proposed to fuse context information, suppress the impact of complex environments, and improve the accuracy and recall rate of citrus detection (Liu *et al.*, 2024b).

After the analysis of the citrus fruit data set, most of the citrus fruit labels are less than 0.1 of the original figure, that is, the actual citrus fruits are mostly small targets. The deeper network layers of Darknet-53 improve the feature extraction ability of the model (Xiao *et al.*, 2024). However, with the deepening of the network layers, high-level features will disappear, especially the feature information of small targets. Although the feature pyramid network FPN in YOLOv4 algorithm improves the detection accuracy of small targets, its detection accuracy for small targets still fails to meet the requirements compared with the detection of large and medium-sized targets. Therefore, in order to obtain a deep learning network more suitable for small target citrus fruit detection and reduce the missed detection rate of small and medium targets in citrus fruit, this paper first reduced the number of feature extraction network layers of the YOLOv4 model, rejected the 32 times down sampling layer of the original network, and added 4 times sampling layer to obtain more abundant texture information of small targets, as shown in Figure 2.

AC network structure

In order to provide sufficient contextual information for small citrus targets and mitigate the impact of complex backgrounds, this paper proposes an AC network structure that integrates self-attention mechanism and contextual information, as shown in Figure 3. The AC network structure fuses the higher level of the target feature layer with the low-level feature mapping enhanced by the attention module to generate a new feature combination that contains target context information (Guo *et al.*, 2024). For example, when using an 8-fold down sampling layer P3 to detect a target, its contextual features come from the 4-fold down sampling layer P2 and the 16-fold down sampling layer P4 (Du *et al.*, 2024). Before feature stitching, convolution down sampling is performed on the P2 feature layer to make it have the same spatial size as the target feature layer P3, and the number of channels is set to 1/2 of the target feature layer. Deconvolution the P4 feature layer to obtain the same scale as the P3 feature layer, with the channel number set to 1/2 of the P3 feature layer. Finally, the target features and context features are superimposed to obtain enhanced feature information P5. In the YOLOv4 network, there will be 5 consecutive 3 × 3 and 1 × 1 convolution layer. Generally speaking, the operation of repeated convolution in high-level convolution can handle situations with multiple categories (Zhang and Su, 2023). However, there is only one category for citrus detection on the road surface, which means that the recognition effect of the model can be improved by reducing the number of high-level convolution layers.

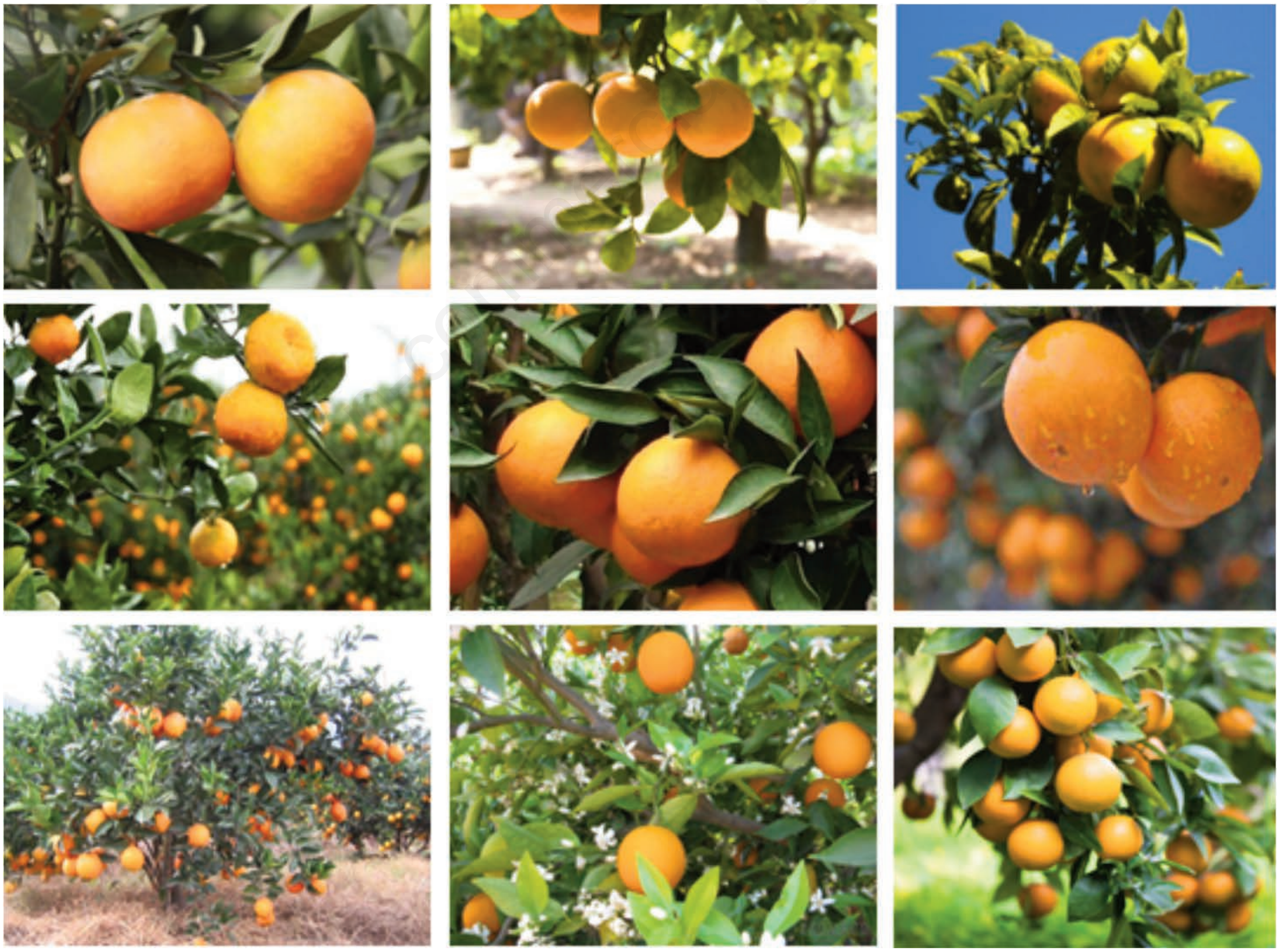

Figure 1. Example of citrus dataset.

Therefore, in this paper, the continuous 5-layer convolution layer is reduced to 3 layers to improve the detection effect of citrus. Among them, the addition of 4x down sampling features not only enriches the texture information of small targets, but also brings in more background texture information and noise, which has a negative impact on the detection effect. At the same time, citrus detection is mostly conducted in outdoor environments, with complex and variable backgrounds, resulting in a high false detection rate of the model. Therefore, a self-attention mechanism module (CBAM) is added after sampling low-level features to reduce the negative impact of the background and make the model pay more attention to the target itself. The CBAM network structure is shown in Figure 4. CBAM consists of a spatial attention module and a channel attention module, which weights useful information and suppresses noise and background feature information. Given an intermediate feature graph F? $R^{C\times H\times W}$ as input, CBAM sequentially calculates a one-dimensional channel attention graph M_C? $R^{C\times 1\times 1}$ and a two-dimensional spatial attention graph M_S? $R^{1\times H\times W}$.

The complete attention process can be summarized in Eq. (1):

$$F'=M_C\,(F)?\ F$$
$$F''=M_s\,(F')?\ F'$$

where ? represents bitwise multiplication; when multiplied by bits, the attention value is broadcast accordingly: the channel attention value is broadcast along the spatial dimension, and *vice versa;* Fis the final refined output.

Each channel of a feature represents a classifier, and the channel attention mechanism is to select important channels and increase their weight. The channel attention module in the CBAM module uses global pooling and average pooling to separately utilize different information and summarize spatial characteristics. As shown in Figure 5, input the feature F? $R^{C\times H\times W}$ for global pooling and average pooling. Then switch to shared MLP and get two 1's × one × channel description for C. The channel weight coefficient M_c is obtained by adding the two feature layers and activating the sigmoid. The spatial attention module focuses on the meaningful areas in each channel. As shown in Figure 6, the feature layer F' of channel attention undergoes maximum pooling and average pooling in turn, and then passes through the convolution layer and is activated by sigmoid to obtain the spatial weight coefficient M_s.

The channel attention module uses the parallel pooling method of Max and Avg to increase the weight of important channels through the relationship between channels, reduce the weight of channels such as background, and obtain more citrus feature information to achieve better recognition and classification effects. The spatial attention module also includes maximum pooling and average pooling, which undergo a single core convolution to enable the

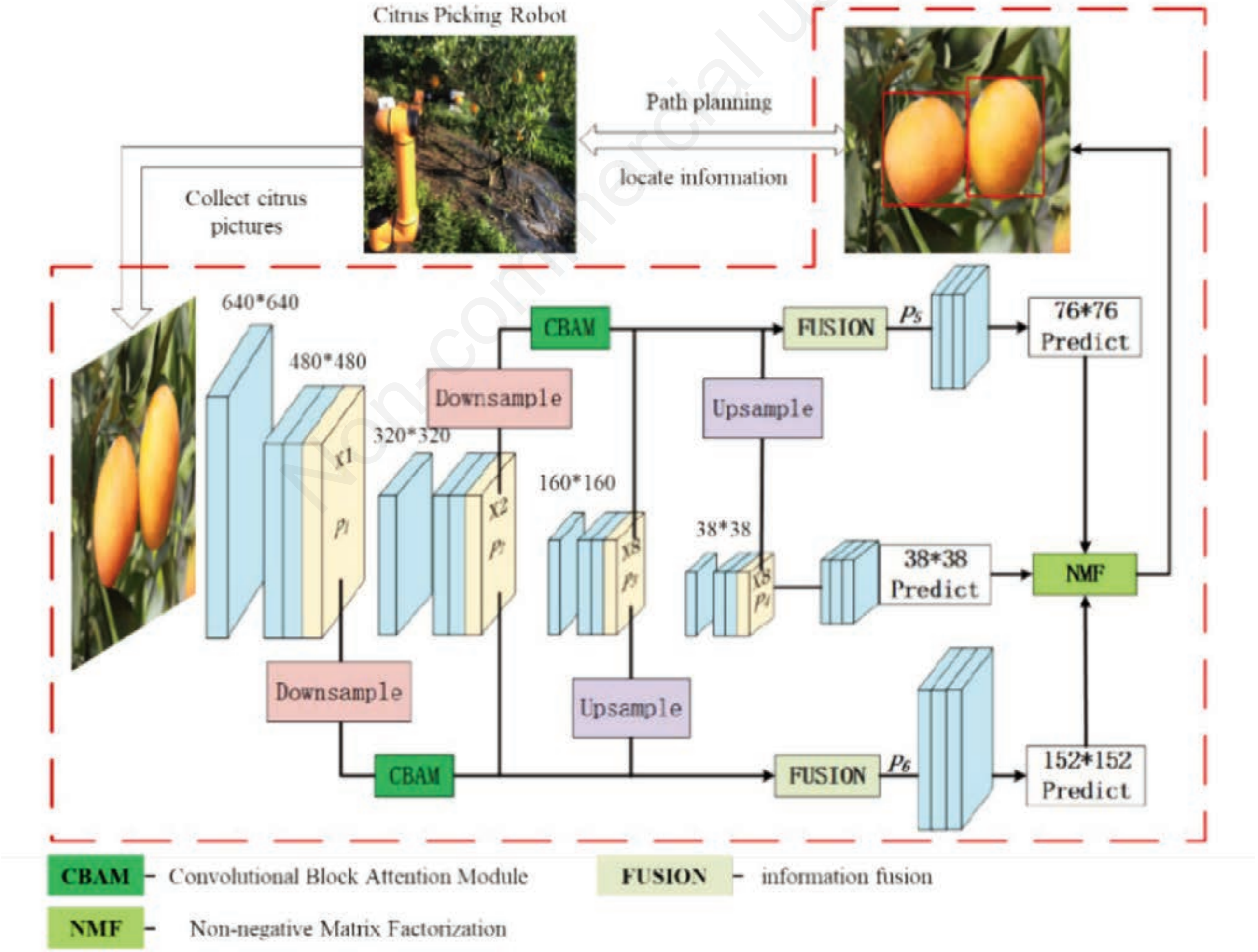

Figure 2. AC-YOLO network structure.

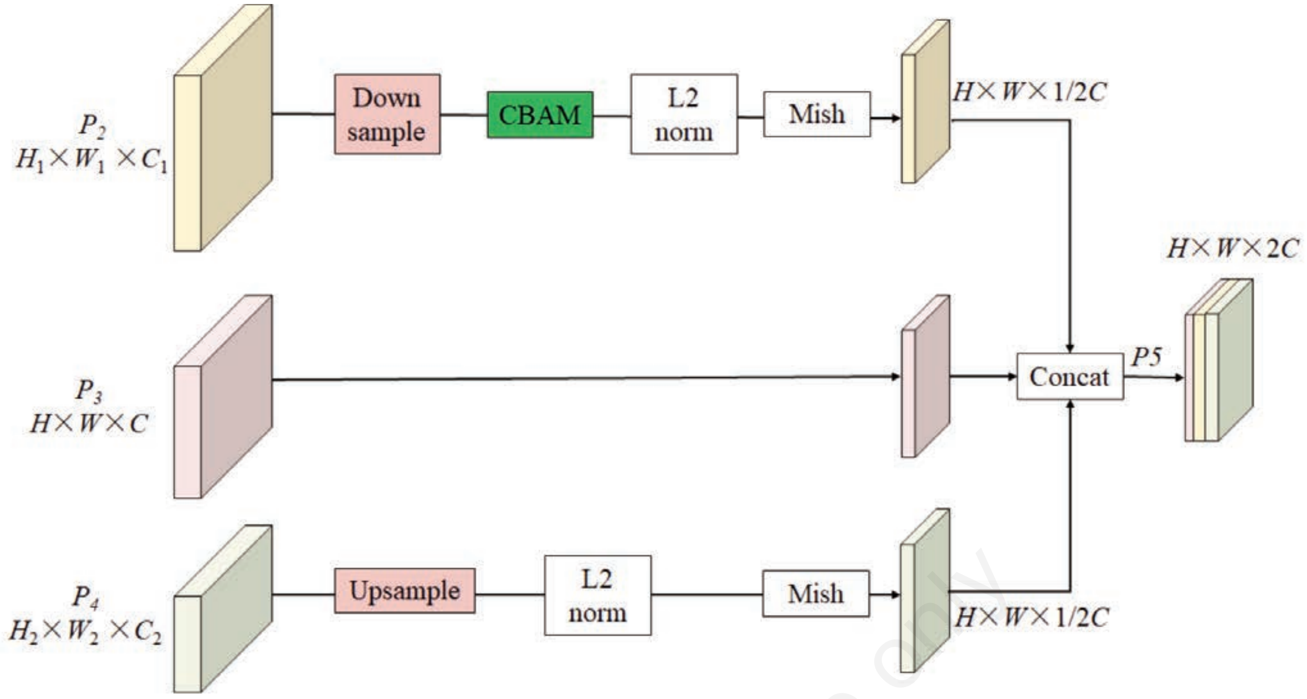

Figure 3. AC network

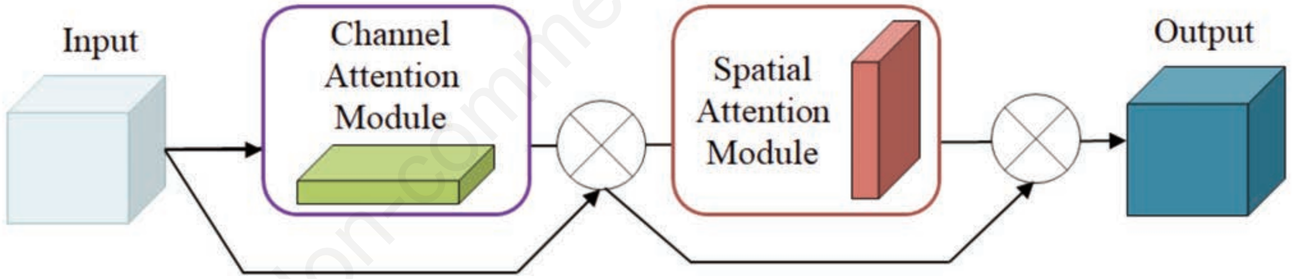

Figure 4. Self-attention mechanism module structure.

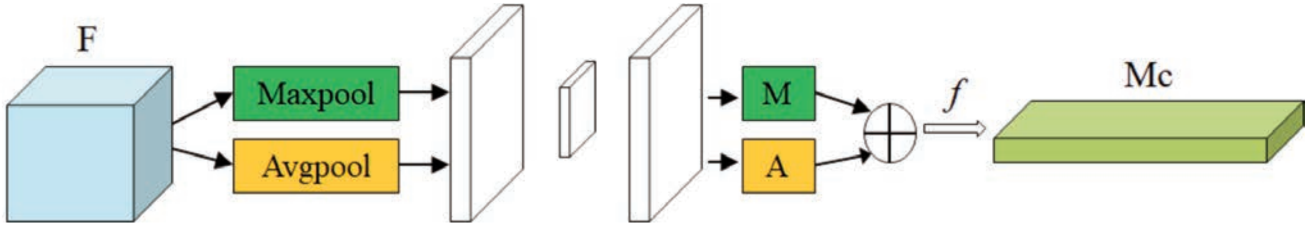

Figure 5. Channel attention module.

model to more accurately learn the location of the citrus. The single core convolution in the spatial attention module defaults to a 7 * 7 convolution core, which can be adjusted according to different detection targets when used.

Mish activation function

The role of the activation function is to map the input data of neurons to a nonlinear domain, enabling neural networks to have the ability to characterize nonlinear characteristics. As shown in Figure 7, if the neuron structure in the above figure does not have the nonlinear mapping ability given by the activation function, the resulting output is always linear, and the neural network always obtains a linear model regardless of how many layers are added (Zheng *et al.*, 2024). However, the data distribution in nature is usually not a simple linear distribution, but a complex nonlinear distribution. Simple linear models cannot fit the data distribution (Wang *et al.*, 2023b). Therefore, an excellent activation function can enable neural networks to have better nonlinear mapping capabilities, while enabling neural networks to have stronger feature learning capabilities (Liu and Liu, 2024). Sigmoid is one of the classical neural network activation functions. Eq. (2) is a sigmoid expression, and Figure 8 is a functional coordinate diagram. From the graph, it can be concluded that sigmoid is a monotonically continuous function. When the input is from negative infinity to positive infinity, the output is (0,1), and the optimization process is relatively stable during network training. However, the soft saturation of sigmoid can easily lead to the phenomenon of gradient disappearance during training, resulting in the loss of the ability of neural networks to continue optimizing network weights during training (Luo *et al.*, 2023).

$$sigmoid(x) = \frac{1}{1+e^{-x}} \tag{2}$$

The Tanh neural network activation function is similar to the sigmoid activation function, and is also a monotonically continuous function with soft saturation. It can map any value from negative infinity to positive infinity onto the [-1,1] interval. Eq. (3) is the Tanh expression, and Figure 9 shows the coordinate graph of the Tanh function. From the figure, it can be seen that the Tanh activation function has a larger slope than the sigmoid activation function, so it can make the training network converge faster. However, the Tanh activation function is also prone to gradient disappearance during training.

$$tanh(x) = \frac{e^{x}-e^{-x}}{e^{x}+e^{-x}} \tag{3}$$

The ReLU activation function is essentially a piecewise function. Eq. (4) is the expression of the *ReLU* function, and Figure 10 is the coordinate diagram of the *ReLU* function. When x ≥0, the function output is x, and the function derivative is equal to 1, without saturation. Therefore, *ReLU* alleviates the problem of gradient disappearance during the training of the network. Moreover, the calculation process of the Re LU activation function is very simple and fast. When x<0, the Re LU activation function is in a saturated state and the value of the function is always 0, which causes the network to generate many inactive neurons, also known as dead neurons. The Death Sutra element does not play any role in the process of network training, which results in a seemingly large network with a large number of parameters, but only a small number of neurons actually work, greatly reducing the efficiency of the neural network.

$$ReLU(x) = \begin{cases} 0, x < 0 \\ x, x \geq 0 \end{cases} \tag{4}$$

The Mish activation function can be expressed as follows:

$$f(x) = xtanh(softplus(x)) = xtan(\ln(1+e^{x})) \tag{5}$$

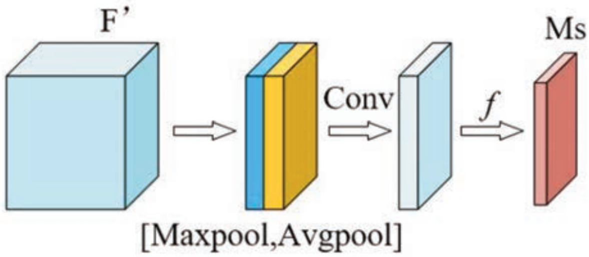

Figure 6. Spatial attention module.

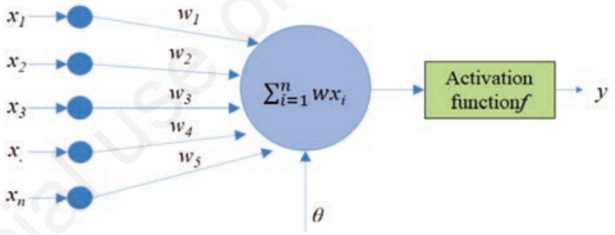

Figure 7. Neuron activation function.

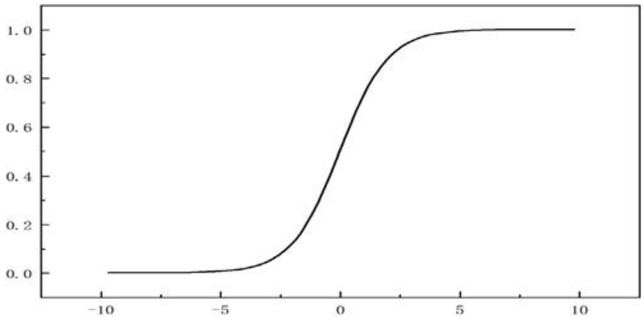

Figure 8. Sigmoid activation function.

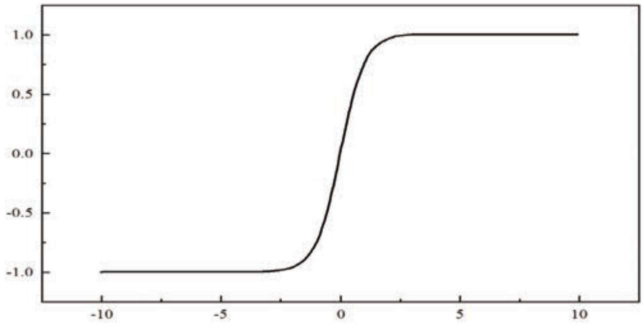

Figure 9. Tanh activation function..

As shown in Figure 11, both Mish and Leaky Re LU activation functions have the characteristics of upper unbounded and lower bounded, and positive convergence rates are comparable. The unbounded nature of the activation function ensures that gradient saturation is avoided during model training and speeds up model training; Having a lower bound and a smaller negative value can ensure the realization of regularization effects and the stability of the gradient flow of the network. However, compared to Leaky Re LU activation function, Mish activation function has better nonlinear characteristics, making the model have better generalization ability, and can improve the accuracy of the model in predicting citrus.

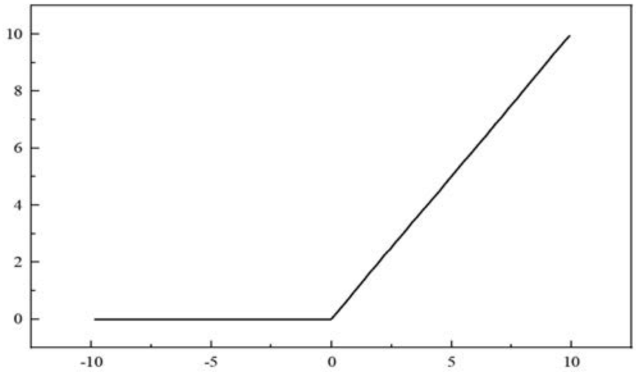

Figure 10. ReLU activation function.

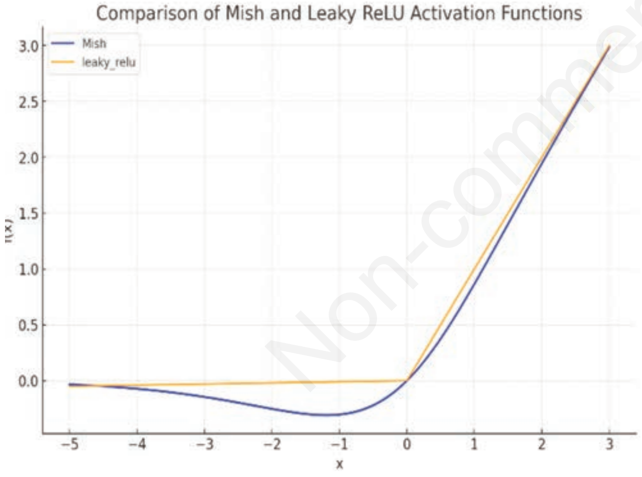

Figure 11. Mish and Leaky Relu activation functions.

Results and Analysis

Evaluating indicator

Considering the requirements for the speed and accuracy of citrus detection in practical applications, the experimental analysis uses the average of accuracy, recall, and average accuracy as evaluation indicators.

i) Accuracy and recall:

$$P_{Precision} = \frac{TP}{TP + FP}$$

$$P_{Recall} = \frac{TP}{TP + FN}$$

where TP represents the number of samples where the detected target category is consistent with the real target category, FP represents the number of samples where the detected target category is inconsistent with the real category, and FN represents the number of samples where the real target exists but has not been detected. The citrus dataset used in this article has many small targets, and improving the recall rate is even more important.

Table 1. Prior frame combination.

Target feature layer	Prior frame combination
×4	(20, 10) (26, 15) (31, 22)
×8	(37, 30) (42, 22) (52, 56)
×16	(55, 19) (70, 32) (94, 98)

Table 2. Experimental results of citrus data set.

Algorithm	mAP@0.5	APS	ARS
YOLOv5	80.5	11.5	29.8
AC-YOLO	78.8	12.8	32.1

Table 3. Results of ablation experiment.

Algorithm	P	R	mAP@0.5/%
YOLOv4	0.875	0.862	88.9
AC-YOLO	0.898	0.948	91.2
	0.902	0.933	92.8
	0.912	0.921	93.6

Table 4. Comparison of different models for sub image-recognition results.

Network model	Precision (%)	Recall (%)	AP value (%)	F1 score	File size (MB)	Detection time
Fast R-CNN	93.55	82.73	89.87	0.88	124	0.276
Faster R-CNN	94.57	91.45	87.53	0.71	521	0.053
YOLOv4	95.12	92.88	95.25	0.94	244	0.052
Improved YOLOv4	95.08	88.73	93.25	0.92	54	0.039
YOLOv5-l	95.48	88.90	93.66	0.92	225	0.054
YOLOv5-x	95.81	88.43	93.26	0.92	347	0.064
AC YOLO	96.19	95.47	97.27	0.96	171	0.052

ii) Average accuracy:

$$AP = \int_{0}^{1} P(R)dR$$

$$mAP = \frac{\sum_{i=1}^{N} AP_i}{N}$$

Anchor parameter optimization

YOLOv4 draws on prior frame Anchors to predict the frame coordinates, which are sets of prior frames with fixed width and height. In the process of target detection, the size of prior frame and the matching degree of target directly affect the speed and accuracy of model detection. Therefore, it is particularly important to set the matching Anchor parameter combination according to the characteristics of citrus fruit labels in the citrus fruit data set. The citrus fruit targets in the data set used in this paper are mostly small targets, and the original prior frame size of YOLOv4 no longer meets the detection requirements, so it is necessary to re-cluster and optimize Anchor parameters. Considering that the K-means algorithm has great randomness in the selection of the initial cluster center, which will have a negative impact on the clustering results, this paper selects the k-means++ algorithm with less randomness to perform the clustering calculation. The k-means++ algorithm was used to recluster the homemade citrus fruit data set. Through multiple clustering comparison, when the combination of prior frames is greater than 9 groups, there are redundant prior frame combinations, so 9 groups of prior frames are the optimal combination of citrus fruit data sets. The distribution of detection scales is shown in Table 1.

Image recognition results

Training loss value and training process

The AC-YOLO model proposed in this article is used to train the training set, and the trained model is used to detect the verification set. The loss curve and verification set training process of each training generation's training set and verification set are shown in Figure 12.

As can be seen from the training process in Figure 12, when the number of iterations reaches 300, the loss curve of the verification set tends to flatten out, and the evaluation indicators for the

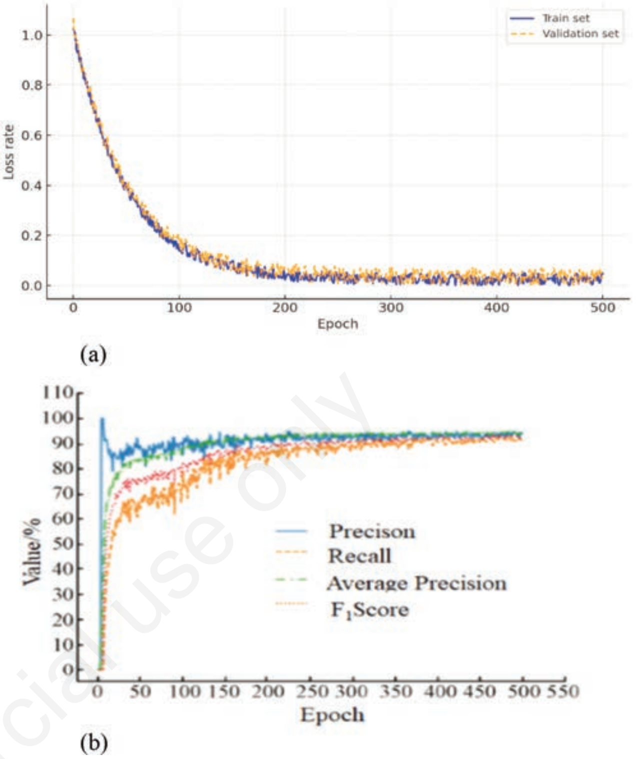

Figure 12. Training process of AC-YOLO model. **a**) Loss curve of training set and validation set; **b**) each evaluation curve of validation set.

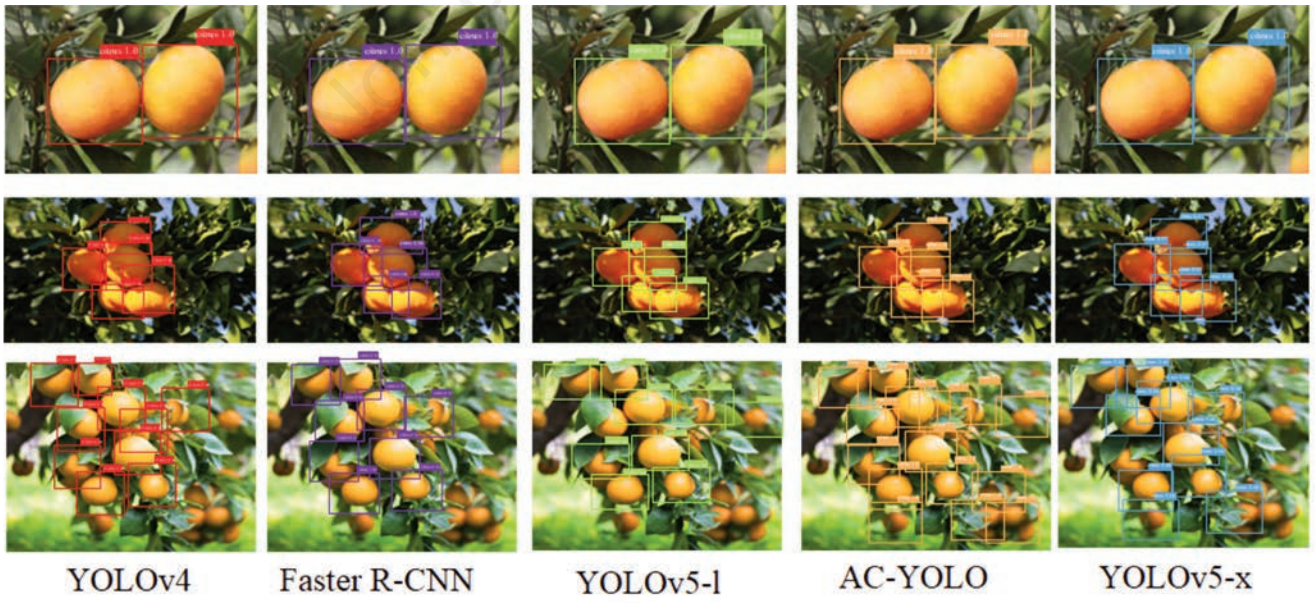

Figure 13. Visualization of our AC-YOLO detection results and some comparison methods in the natural environment of the orchard.

verification set gradually stabilize. Finally, the loss rate of the validation set fluctuates around 0.04, and the AP value of the validation set reaches over 93%, resulting in model convergence.

Quantitative comparison experiment

In order to verify that the proposed method can effectively improve the efficiency of small target detection, the average accuracy of the evaluation indicators mAP@0.5, APs and ARs were selected to conduct a comparative experiment on the citrus fruit dataset, and the input image size was 640×640. As can be seen from Table 2, on the basis of ensuring the detection accuracy, AC-YOLO has achieved better results in the accuracy and recall rate of small targets than YOLOv5 and super resolution methods, which indicates that AC-YOLO can effectively improve the detection effect on small target citrus fruits. In order to further verify that the proposed method effectively improves the efficiency of citrus fruit detection, target detection algorithms were selected to evaluate index accuracy (P), recall rate (R) and average accuracy (m AP@0.5) to conduct ablation comparison experiments on citrus fruit data sets. As can be seen from Table 3, the addition of AC network structure significantly improves the recall rate and accuracy of the model, indicating that the AC structure effectively improves the detection ability of the network for small target citrus fruits. At the same time, the use of Mish activation function enhanced the generalization ability of the model, further improving the detection performance of the citrus fruit detection model.

Performance evaluation of different models

In the experiment, the training sets with the same partitioning strategy were used to train the currently widely used target detection network models Fast R-CNN, Faster R-CNN, YOLOv4, Improved YOLOv4, YOLOv5-l, YOLOv5-x, and the AC YOLO model proposed in this article. The same test set data were tested separately, and the results obtained are shown in Table 4.

As can be seen from Table 4, the Faster R-CNN model occupies a large space and has an accuracy rate of only 57.94%, resulting in poor recognition performance. The accuracy and recall rate of the Center Net model are relatively low, 93.55% and 82.73%, respectively, which are lower than the YOLOv4 and YOLOv5 series algorithms. The YOLOv4-Lite model takes up less space and has a faster detection speed, but the recall rate is only 88.73%. The accuracy of YOLOv5-l and YOLOv5-x models is relatively high, reaching 95.48% and 95.81% respectively, but the recall rate is nearly 4 percentage points lower than YOLOv4. The improved YOLOV4 model proposed in this article has a detection speed comparable to that of the YOLOV4 model, with an accuracy rate of 96.19%, a recall rate of 95.47%, and an AP value of 97.27%. Compared with the previous YOLOV4 model, it has increased by 1.07, 2.59, and 2.02 percentage points, respectively, with an F1 score of 0.96. The recognition effect of the YOLOv4 model before and after the improvement is shown in Figure 13.

There is a phenomenon of missing recognition in the recognition of citrus fruits using the model trained using the AC-YOLO model, which is mainly reflected in the fact that in citrus images taken on cloudy and sunny days, the missing recognition phenomenon is relatively serious for fruit targets with severe occlusion or deviation in fruit color caused by overexposure. The improved YOLOv4 model in this article can effectively detect the fruits that are not identified above by detecting the same image.

Conclusions

In order to solve the problem of high miss detection rate and false detection rate in citrus detection under complex background, this paper proposes an AC-YOLO based citrus detection algorithm in orchard natural environment. By collecting citrus images and labeling images in the natural environment of the orchard, and analyzing the characteristics of citrus on the road, this paper established a citrus dataset in the natural environment of the orchard. Then, K-means++algorithm is used to cluster and analyze the citrus dataset to obtain the optimal Anchor parameter. Aiming at the difficulties of citrus detection in orchard natural environment, this paper proposes an AC network based on self-attention and context feature information, and calls on Mish activation function to enhance the generalization ability of the model, improving the detection accuracy of citrus in orchard natural environment.

In this paper, the attention mechanism of AC network structure and deep separable convolution are introduced to improve the YOLOv4 network model for citrus recognition in the natural environment of orchards. Through testing different target detection network models and the AC-YOLO model proposed in this paper, the results show that the model proposed in this paper performs better than other models, with accuracy, recall, and average accuracy reaching 96.19%, 95.47%, and 97.27%, respectively, with an F1 score of 0.96.

This paper proposes a boundary box matching and merging algorithm based on threshold value, which combines the recognition results of citrus fruits in the orchard natural environment using the AC-YOLO model. The accuracy rate of the merged citrus recognition results reaches 96.17%, the recall rate reaches 95.63%, and the average accuracy reaches 95.06%. Compared with the direct recognition of citrus, the algorithm has significantly improved the effect, and can better identify citrus fruits in the orchard natural environment with larger resolution.

References

Alaaudeen, K.M., Selvarajan, S., Manoharan, H., Jhaveri, R.H. 2024. Intelligent robotics harvesting system process for fruits grasping prediction. Sci. Rep. 14:2820.

Bac, C.W., Roorda, T., Reshef, R., Berman, S., Hemming, J., van Henten E.J. 2016. Analysis of a motion planning problem for sweet-pepper harvesting in a dense obstacle environment. Biosyst. Eng. 146:85-97.

Bai, Y., Zhang, B., Xu, N., Zhou, J., Shi, J., Diao, Z. 2023. Vision-based navigation and guidance for agricultural autonomous vehicles and robots: A review. Comput. Electron. Agr. 205:107-124.

Barbashov, N.N., Shanygin, S.V., Barkova, A.A. 2022. Agricultural robots for fruit harvesting in horticulture application. IOP C. Ser. Earth Environ. 981:9-32.

Chen, C., Lu, J., Zhou, M., Yi, J., Liao, M. Gao, Z. 2022. A YOLOv3-based computer vision system for identification of tea buds and the picking point. Comput. Electron. Agr. 198:107-116.

Du, X., Han, X., Shen, T., Meng, Z., Chen, K., Yao, X., et al. 2024. Natural frequency identification model based on BP neural network for Camellia oleifera fruit harvesting. Biosyst. Eng. 237:8-49.

Fang, Z., Liang, X. 2022. Intelligent obstacle avoidance path planning method for picking manipulator combined with artificial potential field method. Ind. Robot 49:835-850.

Guo, Y., Dong, H., Wang, G., Ke, Y. 2018. A robotic boring system for intersection holes in aircraft assembly. Ind. Robot 4:28-336.

Guo, Z., Shi, Y., Ahmad, I. 2024. Design of smart citrus picking model based on Mask RCNN and adaptive threshold segmentation. PeerJ Comput. Sci. 10:e1865.

Hannan, M.W., Burks, T.F., Bulanon, D.M. 2009. A machine vision algorithm combining adaptive segmentation and shape analysis for orange fruit detection. CIGR J. 6:1-17.

Jana, S., Basak, S., Parekh, R. 2017. Automatic fruit recognition from natural images using color and texture features. IEEE Conf. Devices for Integrated Circuit, Kalyani. pp. 620-624.

Ju, C., Kim, J., Seol, J., Son, H.I., 2022. A review on multirobot systems in agriculture. Comput. Electron. Agr. 202:107-136.

Krizhevsky, A., Sutskever, I., Hinton, G.E. 2012. ImageNet classification with deep convolutional neural networks. Proc. Advances in Neural Information Processing Systems 25:1097-1105.

Liu, J., Liu, Z. 2024. The vision-based target recognition, localization, and control for harvesting robots: a review. Int. J. Precis. Eng. Manuf. 25:409-428.

Liu, M., Bian, Y., Liu, Q., Wang, X., Wang, Y. 2024a. Weakly supervised tracklet association learning with video labels for person re-identification. IEEE Trans. Pattern. Anal. Mach. Intell. 46:3595-3607.

Liu, M., Wang, F., Wang, X., Wang, Y., Roy-Chowdhury, A.K. 2024b. A two-stage noise-tolerant paradigm for label corrupted person re-identification. IEEE Trans. Pattern. Anal. Mach. Intell. 46:4944-4956.

Luo, K., Zhang, X., Cao, C., Wu, Z., Qin, K., et al. 2023. Continuous identification of the tea shoot tip and accurate positioning of picking points for a harvesting from standard plantations. Front. Plant Sci. 14:1211279.

Lv, J., Xu, H., Xu, L., Zou, L., Rong, H., Yang, B., et al. 2022. Recognition of fruits and vegetables with similar-color background in natural environment: A survey. J. Field Robot. 39:888-904.

Mai, X., Zhang, H., Jia, X., Meng, M.Q-H., 2020. Faster R-CNN with classifier fusion for automatic detection of small fruits. IEEE T. Autom. Sci. Eng. 17:1555-1569.

Nehme, H., Aubry, C., Solatges, T., Savatier, X., Rossi, R., Boutteau, R. 2021. Lidar-based structure tracking for agricultural robots: Application to autonomous navigation in vineyards. J. Intell. Robot. Syst. 103:61.

Vasconez, J.P, Kantor, G.A, Cheein, F.A.A. 2019. Human-robot interaction in agriculture: A survey and current challenges. Biosyst. Eng. 179:35-48.

Wan, S., Goudos, S. 2020. Faster R-CNN for multi-class fruit detection using a robotic vision system. Comput. Netw. 168:107-126.

Wang, L., Wang, Z., Liu, M., Ying, Z., Xu, N., Mdeng, Q. 2022. Full coverage path planning methods of harvesting robot with multi-objective constraints. J. Intell. Robot. Syst. 106:17.

Wang, Y., He, Z., Cao, D., Ma, L., Li, K., Jia, L., Cui, Y. 2023a. Coverage path planning for kiwifruit picking robots based on deep reinforcement learning. Comput. Electron. Agr. 205:107593.

Wang, Y., Wu, H., Zhu, Z., Ye, Y., Qian, M. 2023b. Continuous picking of yellow peaches with recognition and collision-free path. Comput. Electron. Agr. 214:108273.

Xiao, X., Huang, J., Li, M., Xu, Y., Zhang, H., Wen, C., Dai, S. 2022. Fast recognition method for citrus under complex environments based on improved YOLOv3. J. Eng. 2022:148-159.

Xiao, X., Jiang, Y., Wang, Y. 2024. A method of robot picking citrus based on 3D detection. IEEE Instru. Meas. Mag. 27:50-58.

Xu, R., Li, C. 2022. A modular agricultural robotic system (MARS) for precision farming: concept and implementation. J. Field Robot. 39:387-409.

Zhang, Q, Su, H-W. 2023. Real-time recognition and localization of apples for robotic picking based on structural light and deep learning. Smart Cities (Basel) 6:3393-3410.

Zheng, X., Rong, J., Zhang, Z., Yang, Y., Li, W., Yuan, T. 2024. Fruit growing direction recognition and nesting grasping strategies for tomato harvesting robots. J. Field Robot. 41:300-313.

Zhuang, J.J, Luo, S.M., Hou, C.J., Tang, Y., He, Y., Zue, X.Y. 2018. Detection of orchard citrus fruits using a monocular machine vision-based method for automatic fruit picking applications. Comput. Electron. Agr. 152:64-73.