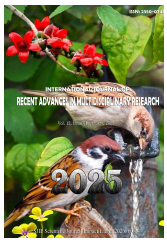
ISSN : 2350-0743

International Journal of Recent Advances in Multidisciplinary Research

Vol. 12, Issue 01, pp. 10619-10625, January, 2025



RESEARCH ARTICLE

FRUIT CLASSIFICATION ALGORITHM DESIGN BASED ON IMPROVED YOLOV8

***Changyou Wang, Qing Zhang and Jie Huang**

College of Applied Mathematics, Chengdu University of Information Technology, Chengdu, Sichuan, 610225, P. R. China

ARTICLE INFO

Article History

Received 20th October, 2024
Received in revised form
16th November, 2024
Accepted 27th December, 2024
Published online 24th January, 2025

Keywords:

YOLOv8, Multi-target detection, Receptive field attention, Fruit classification.

****Corresponding author:***
Muarrah Nisar

ABSTRACT

Fruit classification plays a pivotal role in the fruit industry. Traditional fruit classification methods rely heavily on manual inspection, which significantly hampers the accuracy and efficiency of harvesting. The advent of deep learning algorithms has empowered robots to classify fruits more swiftly, thereby enhancing picking efficiency. In this paper, we propose a deep optimization target detection algorithm based on the YOLOv8s framework to improve the accuracy and efficiency of fruit classification. This algorithm leverages the YOLOv8s model and introduces optimizations to enhance fruit classification effectiveness. Firstly, we focused on making crucial enhancements to the head and neck network's structure. Specifically, we replaced the original C2f module with the Enhanced Local Aggregation Network (ELAN) module, which retains more comprehensive feature information and learns directly from the original feature map. This modification effectively minimizes information loss, bolsters feature representation capabilities, and subsequently elevates detection accuracy. Secondly, we integrated receptive field attention into both the convolutional layer and feature extraction layers of the backbone and neck networks. By augmenting the network's capacity to capture local region and global context information, this design significantly improves the model's efficiency in learning detailed features, thereby accelerating target detection speed and refining detection accuracy. Furthermore, in fruit classification detection tasks, incorporating multiple attention mechanisms before the backbone network and the large object detection head enables dynamic adjustment of feature weights. This approach strengthens key target features, suppresses background noise, and markedly reduces the false detection rate. This effect is particularly pronounced in scenarios with numerous fruit picking targets and complex backgrounds. The proposed model exhibits excellent detection accuracy and is suitable for practical fruit detection applications.

Citation: Changyou Wang, Qing Zhang, and Jie Huang. 2024. *"*Fruit classification algorithm design based on improved YOLOv8*", International Journal of Recent Advances in Multidisciplinary Research,* 12, (01), 10619-10625.

INTRODUCTION

With the improvement of economic levels, the fruit industry occupies an important position in the agricultural planting industry in our country. The utilization of deep learning algorithms to replace traditional manual technology for fruit classification holds great significance for the development of the entire fruit industry. Deep learning algorithms can not only address the issue of low efficiency in robotic fruit classification but also aid in resolving the problem of inconsistent classification standards in robotic visual inspection (1). In fruit image classification, deep learning-based detection algorithms encounter challenges such as prolonged detection times and low accuracy. Object detection algorithms in deep learning are categorized into two-stage (e.g., R-CNN (2), etc.) and single-stage (SSD (3) and YOLO (4)) algorithms. The two-stage algorithm requires additional networks to form candidate frames, resulting in slower speeds.

While the single-stage algorithm offers faster detection speeds, its accuracy is not high. Therefore, various improvements have been made based on existing algorithms to better adapt them for future life scenarios. Ge *et al*. (5) proposed a novel method for detecting fruits using deep convolutional neural networks. They applied the Faster R-CNN model to multi-modal information through transfer learning, generating a new model that improved the detection accuracy and recall performance for sweet peppers. Yamamoto *et al*. (6) initially conducted tomato detection through color-based segmentation, subsequently trained CART classifiers to produce segmentation maps and grouped pixels using color and shape features. Additionally, they trained non-fruit classifiers with random forests in a controlled greenhouse environment to minimize false alarms. Zhang *et al*. (7) introduced a video sequence-based orange counting algorithm leveraging deep

learning, which incorporated fruit detection and tracking algorithms. They enhanced the YOLOv3 backbone network and incorporated a multi-scale fusion module to address the issue of repeated citrus fruit counting. Koirala *et al.* (8) compared the performance of six deep learning architectures in detecting mango fruits and developed a new architecture, "MangoYOLO," which demonstrated excellent and robust performance. They verified that MangoYOLO significantly improved detection on a mango image dataset during daytime. Apolo *et al.* (9), utilizing short-term memory, trained a Faster R-CNN deep learning dataset for citrus yield prediction and size estimation using a custom dataset. Wan *et al.* (10) improved the convolution and pooling layers and the model framework within the Faster R-CNN model, and appropriately expanded the dataset to enhance the detection accuracy of multiple types of fruits while reducing processing time. Akiva *et al.* (11) constructed a Triple-S network, employing low-cost central point annotation for supervision. This network, combined with three-part losses incorporating shape priors, better suited common, known objects in agricultural scenes.

From the aforementioned research, it is evident that the accuracy of current deep learning models for fruit detection needs improvement. The complexity of backgrounds and varying lighting conditions pose significant challenges for detection algorithms in fruit harvesting. Based on YOLOv8s, an advanced model, this paper proposes a series of innovative enhancements. Specifically, the C2f modules located before the pooling layer in the backbone network and before the small object detection head in the neck network are replaced with the Elan module. This modification aids the network in integrating more gradient information prior to feature fusion and small object detection, thereby enhancing detection accuracy. Furthermore, receptive field attention is integrated into the convolutional layers of both the backbone and neck networks to improve the recognition capability for multi-scale targets. Additionally, a non-parametric attention module is introduced before the pooling layer of the backbone network and the large target detection head, which enhances the comprehensiveness of feature extraction and boosts the model's generalization ability. Through these measures, this paper successfully constructs a target detection model that is better suited for fruit classification and detection scenarios, exhibiting superior performance. The experimental results demonstrate that the improved algorithm has achieved remarkable improvements in various evaluation indicators.

- The algorithmic scheme of this paper
- YOLOv8 Base Model

In January 2023, Ultralytics team YOLOv5 proposed the YOLOv8 (12) model, which consists of Backbone, Neck and Head. Its structural model is shown in Figure 1. The structure of YOLOv8 is similar to that of YOLOv5. The obvious changes are as follows: Firstly, the C3 structure in the backbone network of YOLOv5 is replaced with C2f structure; SPPF structure is used to further increase the feature semantic expression ability, retain the original features, and improve the information richness of the pooling layer. Remove redundant convolution in the upper sample of YOLOv5; The coupling head of YOLOv5 was changed to the decoupling head, and the frame-based detector was changed to the frameless detector. The Distribution Focal Loss and fusion CIoU Loss were introduced to calculate the rectangular frame loss.

Improved YOLOv8 Model: In the task of fruit detection, the classification target detection of fruit is complicated due to the variety of fruits, different shapes and the problems of illumination change and occlusion in the image. At the same time, due to the complex and changeable background of fruit detection, it may contain various environmental elements, such as branches and leaves, soil, etc. In this case, the YOLOv8s algorithm model may have the following problems: insufficient adaptability to complex background, easy to be disturbed, resulting in false detection and missing detection; The recognition accuracy of fruits of different shapes and sizes is not high enough; Poor performance in dealing with light changes and other conditions, affecting the detection effect.

In this paper, a series of improvements were made based on the YOLOv8s model. Firstly, Elan(13) module replaced the C2f feature extraction module in the backbone network and neck network, and more level and scale feature information was obtained by increasing the channel dimension and the number of branches, so as to achieve more comprehensive and fine feature fusion. Secondly, the C2f feature extraction module of the backbone network and the neck network and the common convolutional layer are integrated into the receptive field attention (14), and the processing of the receptive field is adjusted according to the different feature maps of each region, so as to improve the precision of the network's understanding of the input features. Finally, multiple attention mechanisms are introduced in front of the backbone network and small object detection head (15), and by adjusting the attention weights of channel dimension and spatial dimension, attention is focused on important areas, so that the model can focus attention on the target area. The final algorithm model structure after modification is shown in Figure 2.

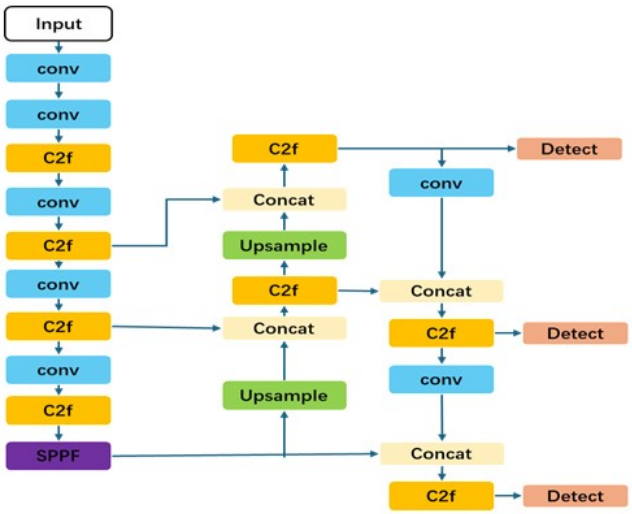


Figure 1. YOLOv8 schematic diagram

Improved feature extraction module: YOLOv8's neck network uses CSPDarknet to extract features from input images. C2f module in YOLOv8 is composed of CSP and FFM, which is the core part of feature extraction of YOLOv8 neck network. The specific structure is shown in Figure 3.

Although C2f in YOLOv8 uses a simpler convolution method to simplify the model, more gradient information is lost and the model tracking effect is still poor.

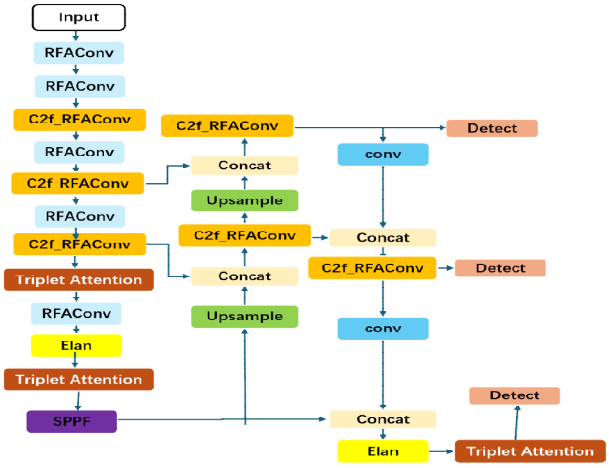


Figure 2. Structure of the improved YOLOv8 algorithm

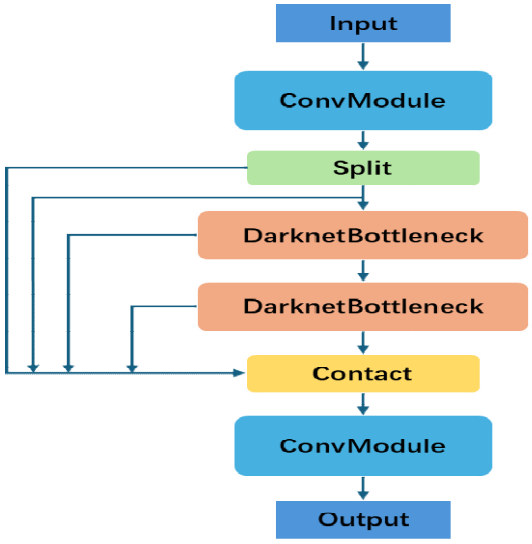


Figure 3. C2f schematic diagram

ELAN module with lower gradient information loss was selected to replace the C2f module in YOLOv8. The structure diagram of ELAN module was shown in Figure 4. The ELAN (Efficient Layer Aggregation Networks) network model in YOLOv7 is divided into several independent modules, and several groups of convolution modules are added to the original gradient transmission path to increase the cardinality of the original image features. The features of different images are reorganized and combined to enhance the image features. ELAN features extraction has a total of four paths, layer by layer aggregation of these modules, to obtain the final network model.

Improved convolutional module: In the backbone network of YOLOv8, conventional convolutional modules and C2f are mainly used to extract network features, and the importance of key feature maps cannot be highlighted. The receptive field convolution attention mechanism RFAConv not only pays attention to the spatial features of the receptive field, but also gives different weights to the features of each region to increase the importance of the target features and enhance the anti-interference ability of the model. Its schematic diagram is shown in Figure 5.

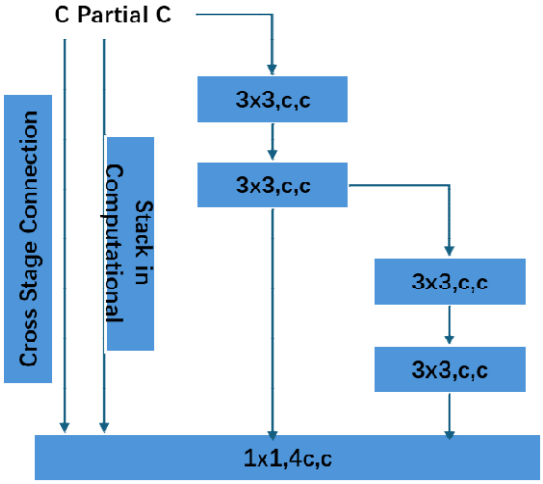


Figure 4. ELAN schematic diagram

Firstly, the feature map of the whole space is average pooled to aggregate the global information of each receptive field feature, and then the group convolution operation is used to carry out information exchange. Finally, softmax is used to emphasize the importance of each receptive field feature, grade the importance of different features in the receptive field slider, and rank their importance levels. This ensures that important features are extracted first. The formula for calculating the receptive field can be:

$$F = Soft\max(g^{1\times1}(AvgPool(X)))\times \\ \mathrm{ReLU}(Norm(g^{k\times k}(X))) = A_{rf}\times F_{rf}, \quad (1)$$

F is obtained by multiplying A_{rf} (attention map) and F_{rf} receptive field space features, which helps the network to better process the information in the receptive field space according to the importance weight of the features and improve the adaptability of the network. $g^{1\times1}$ represents a grouping convolution of size 1×1, X represents the input feature map, k represents the size of the convolution kernel, Norm represents normalization, and ReLU is the activation function.

Adds the Triplet Attention module: In recent years, attention model has been widely used in the field of deep learning algorithms. The attention model mimics the human visual attention system, and the algorithm model can focus more on the object features of target detection, which can better help realize the target detection task. Triplet Attention is a triplet channel data attention mechanism, by establishing the relationship between the three dimensions of space, height, and width, the model can accelerate data processing and quickly find the target object. Triplet Attention not only establishes the weight relationship between height and channel, width and channel, but also establishes the final attention weight relationship between the three dimensions through rotation, arrangement, convolution and other operations, so that the model pays more attention to the multi-dimensional features. The principle of Triplet Attention three-channel is shown in Figure 6. The Triplet Attention constructed by triplet Attention has three branches. The a

branch directly compacts the original tensor by Z-pooling, screening the key channel features and reducing the complexity.

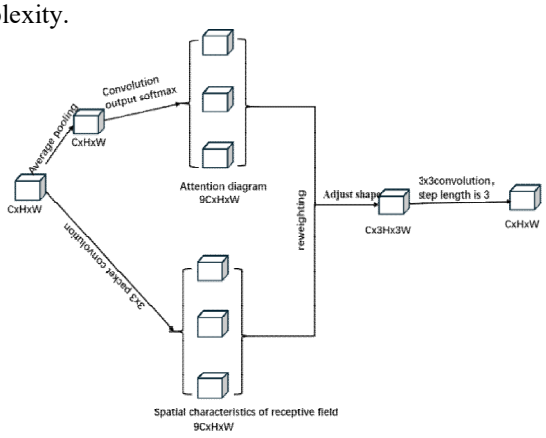


Figure 5. RFA Conv structure diagram

The weights are generated by convolution, batch normalization and Sigmoid, emphasizing important primitive features and acting on the original tensor output. The b branch is a rotation tensor formed after rotating the tensor 90 counterclockwise along the W axis, making it easy to discover the connection between horizontal features. Important horizontal features and data are obtained and optimized by Z-pooling compression, convolution and batch normalization. Sigmoid generates weighted focus keyhorizontal features that act on the rotation tensor output to enhance horizontal orientation processing.

For the c branch, the tensor after the channel of the original input tensor is compressed by Z-pooling is ($2 \times H \times W$), and the input tensor rotates along the height axis, transforming the viewing Angle to capture the vertical feature relationship. The dimensionality of the rotation tensor is reduced by Z-pooling compression and key vertical features are extracted. The features and data distribution are optimized by convolution and batch normalization. The attention weight generated by Sigmoid acts on the rotation tensor output to enhance vertical focus. The tensor ($C \times H \times W$) generated by each branch is aggregated by simple averaging. The algorithm model in this paper chooses to add two Triplet Attention to the backbone network.

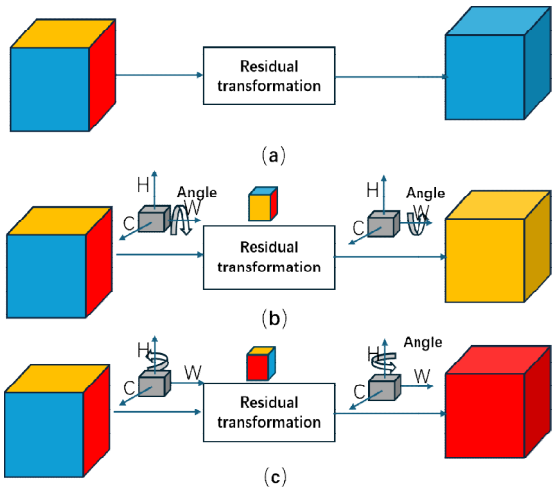


Figure 6. Schematics of multiple attention mechanisms

Experimental Setting and Analysis of Results

Experimental Environment: The data set mainly includes on-site shooting and Internet acquisition, with a total collection of more than 21,000 images, which are divided into training set, test set and verification set, and the ratio is 8:1:1. The dataset covers 13 fruits, including lemon, watermelon, pear, banana, orange, grape, apple, pineapple, blueberry, cantaloupe, peach, dragon fruit and strawberry. The experimental environment of this paper is: the host system Ubuntu21.04 system, the computer GPU is NVIDIA RTX 3090, the running memory is 16G, the development language Python3.16, PyTorch2.0.0.

Data sets and assessment indicators: In this experiment, mAP is used to evaluate the detection performance of the algorithm. mAP@0.50 represents the average accuracy of all categories when the threshold of IOU is 0.5, and mAP@0.50:0.95 represents the average accuracy when the threshold of IOU is from 0.5 to 0.95. If the IOU of the real box and the predicted box is greater than the IOU threshold, it is TP; otherwise, it is FP; FN is the number of real boxes minus TP. Model evaluation indicators include Precision and Recall, and parameter calculation formulas are as follows:

$$\text{mAP} = \frac{1}{N}\sum_{n \in N} AP(n)\text{,} \tag{2}$$

$$\text{Precision} = \frac{TP}{TP + FP}\text{,} \tag{3}$$

and

$$\text{Recall} = \frac{TP}{TP + FN}\text{.} \tag{4}$$

Comparison with Baseline Model: In order to verify the superiority of the proposed model over the benchmark model, we placed the improved YOLOv8s model and the original YOLOv8s model under the same parameters for 150 epochs training respectively, and then carried out tests on the verification set. As can be seen from Figure 7, the yellow lines in the figure represent the data of the model in this paper, and the blue lines represent the data of the original YOLOv8s model. Through experiments, it is found that with the increase of epochs, the yellow lines gradually surpass the data of blue lines, which indicates that the average accuracy and accuracy of the model in this paper are higher than that of the benchmark model.

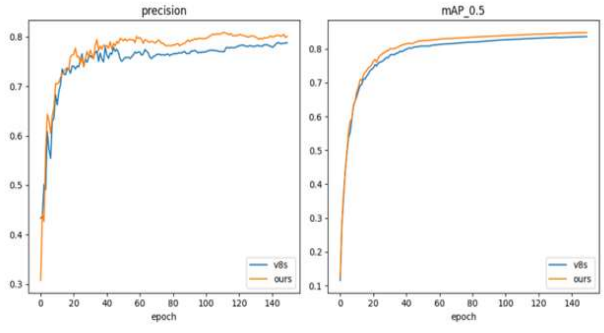


Figure 7. Comparison chart of experimental results

Table 1. Ablation experiment data

Comparison of each parameter of ablation experiment target detection				
Models	mAP@0.50	mAP@0.50:0.95	Parameters	GFLOPs
YOLOv8s	83.6	63.7	11.13M	28.5
YOLOv8s-ELAN	84.3	64.7	9.79M	27.3
YOLOv8s-ELAN- RFAConv	84.5	65	9.79M	28.3
YOLOv8s-ELAN- RFAConv -Triplet Attention	84.8	65.5	9.79M	28.3

Table 2. Comparative experimental data

Models	P	R	mAP@50	mAP@0.50:0.95
Faster-RCNN	77.5	76.5	77.6	50.6
SSD	57.35	76.84	73.48	46
RetinaNet	89.29	68.04	80.32	57.3
YOLOv5s	79	77.8	83	62.5
YOLOv8s	78.8	78.8	83.6	63.7
YOLOv10s	80.1	77.3	83.6	63.1
Ours	80.5	79.5	84.8	65.5

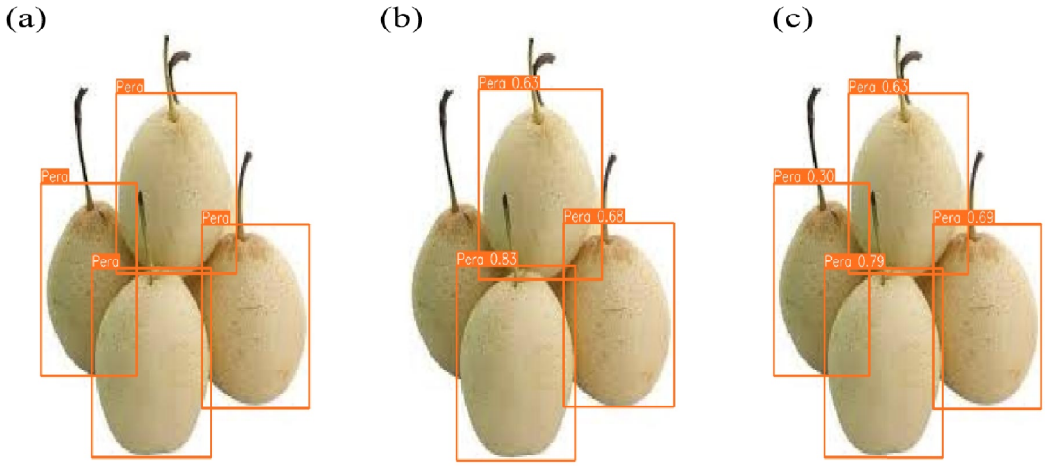


Figure 8. In the comparison diagram of improved experimental results in the detection scenario of a single type of fruit, group A is the target to be found in the original image, Group b is the detection result of YOLOv8s, and group c is the detection result of the model in this paper

Figure 9. Comparison diagram of target detection effect of fruit classification in complex scene, Group b is the detection result of YOLOv8s, and group c is the detection result of the model in this paper

Ablation Experiment: In order to verify the effectiveness of the improved modules, YOLOv8 is taken as the benchmark model, and its backbone network, neck network, feature extraction module, convolutional layer and other aspects are improved. The third to fifth lines in Table 1 indicate that the feature extraction module of backbone network and neck network is replaced by ELAN, convolution module and feature extraction module integrating receptive field Attention, and Triplet Attention is introduced into the model, respectively. According to the experimental data in the table, firstly, after replacing ELAN module, mAP@0.50 of the model increases by 0.7%, mAP@0.50:0.95 increases by 1% and GFLOPs decreases by 1.2. Secondly, C2f_RFAConv module and RFAConv module are introduced into the model. The mAP@0.50 of the model is increased by 0.2%, and the mAP@0.50:0.95 of the model is further increased by 0.3%. Finally, two triplet attention modules are introduced into the backbone network and a single triplet attention module is introduced before the small object detector. The mAP@0.50 of the model is increased by 0.3%. mAP@0.50:0.95 with an increase of 0.5%. Compared with the previous YOLOv8s, the accuracy of the modified model in this paper has significantly increased.

Comparison experiment: In order to verify the effectiveness of the improved YOLOV8 model, this paper carried out comparative experiments with other mainstream models. These experiments were compared in the same fruit classification data set collected in this paper to evaluate the performance indicators of each model. The models compared include Faster-RCNN(16), SSD(17), RetinaNet(18), YOLOv5s(19), YOLOv8s, YOLOv10(20), etc., as well as the improved YOLOv8 model proposed in this paper, as shown in Table 2. It is clear from the comparative experiments in the table that compared with other models, the proposed model has improved in Precision, Recall, mAP@0.50, mAP@0.50:0.95, and has improved recall and precision. Faster-RCNN and SSD belong to two-stage detection algorithms, and the feature extraction of small targets is not sufficient and fine, and the overall structure is complex and the calculation amount is large, resulting in poor detection performance. Although the YOLOv5s model has improved over the two-stage detection, it has more subsampling operations and is easy to lose the features of small targets. Meanwhile, due to the limited scale level, it is not precise and comprehensive enough in feature extraction and fusion, resulting in poor detection effect. Finally, compared with mainstream one-stage algorithm models such as YOLOv5s, YOLOv10s and YOLOv8s, the proposed model has the highest accuracy, detection accuracy and recall rate, and has excellent comprehensive performance.

Ablation Experiment: In the specific scene of fruit classification, the detection results obtained by the algorithm are further compared and analyzed. It can be clearly seen from Figure 8 that in these three groups of pictures, four pears were originally needed to be found, but the original YOLOv8s algorithm model only detected three pears, while the algorithm model proposed in this paper successfully detected all four pears, significantly improving the problem of missing detection.

Figure 9 shows the comparison of target detection effects of fruit classification in a complex scene. In the complex fruit detection scene, it was originally necessary to find 4 complete Limons, but the YOLOv8s model only found 3 Limons, while our model detected 4 limons, and the confidence scores of two limons detected were higher than those of group b pictures. Our model also has good detection results when the detection background is complex and the target is fuzzy. Through the observation and comparison of the two groups of pictures, it can be seen that the original YOLOv8s has problems of missing detection, low confidence score, and weak anti-interference ability. In contrast, the improved model proposed in this paper can detect all targets completely and correctly, and still gives a higher confidence score than the basic model under difficult situations. This comparison clearly shows the difference in precision.

Summarize

Target detection of fruit classification will be of great significance in the future application of fruit picking robots. In this paper, YOLOv8s model is mainly used for fruit detection and classification. The data set in this paper contains most types of daily fruits, which can be widely used in various recognition scenarios. In this paper, a fruit classification and recognition algorithm based on multiple attention and receptive field attention is proposed. On the basis of the baseline model, the convolutional module of the backbone network and the neck network is replaced with receptive field attention convolutional module, and the feature extraction module of the network is replaced with ELAN. The participation-free attention mechanism is introduced to make the model more effectively capture the relationship between the three channels and improve the detection accuracy. In the future, based on this model, we will optimize the network architecture and simplify the model to improve the performance of the model in the fruit classification detection task. In the future, we will continue to study and improve on the basis of the algorithm in this paper.

Authors' contributions: C. Wang, Q. Zhang, andJ. Huang contributed equally to each part of this work.

REFERENCES

1. Chen X, Zhou G, Chen A, *et al*. The fruit classification algorithm based on the multi-optimization convolutional neural network. Multimedia Tools and Applications, 2021, 80: 11313-11330.
2. Girshick R, Donahue J, Darrell T, *et al*. Rich feature hierarchies for accurate object detection and semantic segmentation. Proceedings of the IEEE conference on computer vision and pattern recognition. 2014: 580-587.
3. Liu W, Anguelov D, Erhan D, *et al*. Ssd: Single shot multibox detector. Computer Vision-ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part I 14. Springer International Publishing, 2016: 21-37.

4. Redmon J, Divvala S, Girshick R, *et al*. You only look once: Unified, real-time object detection. Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 779-788.
5. Sa I, Ge Z, Dayoub F, *et al*. Deepfruits: A fruit detection system using deep neural networks. sensors, 2016, 16(8): 1222.
6. Yamamoto K, Guo W, Yoshioka Y, *et al*. On plant detection of intact tomato fruits using image analysis and machine learning methods. Sensors, 2014, 14(7): 12191-12206.
7. Zhang W, Wang J, Liu Y, *et al*. Deep-learning-based in-field citrus fruit detection and tracking. Horticulture Research, 2022, 9: uhac003.
8. Koirala A, Walsh K B, Wang Z, *et al*. Deep learning for real-time fruit detection and orchard fruit load estimation: Benchmarking of 'MangoYOLO'. Precision Agriculture, 2019, 20(6): 1107-1135.
9. Apolo-Apolo O E, Martínez-Guanter J, Egea G, *et al*. Deep learning techniques for estimation of the yield and size of citrus fruits using a UAV. European Journal of Agronomy, 2020, 115: 126030.
10. Wan S, Goudos S. Faster R-CNN for multi-class fruit detection using a robotic vision system. Computer Networks, 2020, 168: 107036.
11. Akiva P, Dana K, Oudemans P, *et al*. Finding berries: Segmentation and counting of cranberries using point supervision and shape priors. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. 2020: 50-51.JOCHER G. Ultralytics YOLOv8(EB/OL). (2023.01.10.). https://github.com/ultralytics/ ultralytics,2023.01.10.
12. Wang C Y, Bochkovskiy A, Liao H Y M. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2023: 7464-7475.
13. Zhang X, Liu C, Yang D, *et al*. Rfaconv: Innovating spatital attention and standard convolutional operation. arxiv preprint arxiv:2304.03198, 2023.
14. Misra D, Nalamada T, Arasanipalai A U, *et al*. Rotate to attend: Convolutional triplet attention module. Proceedings of the IEEE/CVF winter conference on applications of computer vision. 2021: 3139-3148.
15. Ren S, He K, Girshick R, *et al*. Faster R-CNN: Towards real-time object detection with region proposal networks. IEEE Transactions on Pattern Analysis & Machine Intelligence, 2017, 39(6):1137-1149.
16. Liu W, Anguelov D, Erhan D, *et al*. Ssd: Single shot multibox detector. Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part I 14. Springer International Publishing, 2016: 21-37.
17. Lin T Y, Goyal P, Girshick R, *et al*. Focal loss for dense object detection. Proceedings of the IEEE international conference on computer vision. 2017: 2980-2988.
18. JOCHER G. YOLOv5(EB/OL). (2022.09.05.). https:// github.com/ ultralytics/yolov5
19. Wang A, Chen H, Liu L, *et al*. Yolov10: Real-time end-to-end object detection. arxiv preprint arxiv:2405.14458, 2024.

****