Fruit Target Recognition and Maturity Detection Based on Improved YOLOv7

Qi Chen, Renjie Li , Lianjun Hu and Yong Zhang

School of Information Engineering, Tianjin University of Commerce, Tianjin 300134, China, chenqi@tjcu.edu.cn, li18326653378@163.com, hulianjun@tjcu.edu.cn, zhangyong@tjcu.edu.cn

Corresponding author: Yong Zhang, zhangyong@tjcu.edu.cn

Abstract. Aiming at the problems of low detection accuracy and long detection time in real-time detection of traditional convolutional networks, a lightweight model R-SE-YOLOv7 was designed in this paper. First, in the feature enhancement module, the max-pooling layer and deep convolution are applied to retain as many of the important features of the image as possible. The designed module encodes the image with the local region extremum and full image information to improve the module's classification accuracy. Secondly, the design of residual modules is optimized using SE modules and layer normalization to reduce model complexity, enhance generalization capabilities, and mitigate overfitting. The cross-entropy loss function is used to improve the recognition rate under multi-classification tasks. Finally, The R-SE- YOLOv7 model is applied to the recognition and ripeness detection of apples, oranges, and bananas. Comparative analysis with the original model reveals an 8% increase in precision, an 11.1% increase in recall, an increase in mAP by 8.2 %, and an increase in F1 score by 0.05. The experimental results indicate that the designed R-SE-YOLOv7 model achieves accurate fruit object recognition and ripeness detection, and the parameter scale and detection speed have advantages that provide technical support for real-time fruit detection research.

Keywords: Fruit Recognition; Deep Learning; YOLOv7; Feature Enhancement; Residual Modules
DOI: https://doi.org/10.14733/cadaps.2024.S25.156-170

1 INTRODUCTION

In recent years, with the development of artificial intelligence technology and digital image processing technology, target recognition and detection technology based on deep learning has become a research hotspot. Image feature extraction is used to identify and detect targets or areas of interest in images, including fruit recognition and it is a popular application in many fields such as detection. The research of traditional fruit recognition mainly uses three features: color feature, shape feature, and position feature to identify [1], but there are some problems: low average accuracy, slow recognition speed, and low practicability, making it impossible to put it into production. Compared with the problems existing in traditional fruit recognition technology, the continuous development of in-depth learning techniques has brought new directions to fruit recognition. Deep learning is a learning algorithm with multi-level representation [2-3]. The target recognition algorithm based on deep learning realizes feature extraction through a convolutional neural network, which abandons the traditional method of manually designing features. This learning ability makes it possible to improve the performance of target recognition algorithms. [4-6]. Some scholars have applied it to fruit recognition research. However, the deep learning algorithm has many uncontrollable factors such as complex background environment, large fruit density, large differences and more illumination when performing fruit recognition [7-8]. In the meantime, the algorithm models have complex network structures, and there are problems such as excessive occupation of computing resources and poor real-time performance, which affect the accuracy and efficiency of fruit and vegetable recognition [9-10].

The single-stage recognition algorithm and the dual-stage recognition algorithm are two kinds of target recognition algorithms based on deep learning [11]. The representative of the dual-stage recognition algorithm is the Region with CNN Features (RCNN) target recognition algorithm proposed by R. Girshick et al. [12] in 2014. The dual-stage algorithm needs to generate candidate areas in advance and then use the network for classification and regression, so the recognition speed is slow [13-14]. Although many scholars have made improvements, there are still certain problems. The single-stage recognition algorithm does not need to generate candidate areas in advance; it only performs feature extraction once and can directly perform classification and regression [15]. Compared with the dual-stage recognition algorithm, the recognition speed of the single-stage algorithm has been greatly improved. Its representative algorithm is the YOLO algorithm, which was proposed by R. Joseph et al. [16] in 2016. In YOLO algorithm, the input original image is divided into grids, and feature classification and regression prediction are directly performed on the cells of each grid. Although the single-stage recognition algorithm based on YOLO is simpler to implement than the two-stage algorithm, it directly uses fully connected layer prediction, which will lose position information and small target information. Therefore, the accuracy of the single-stage recognition algorithm is lower than that of the two-stage algorithm based on YOLO. To improve the accuracy of target recognition tasks, since the YOLO algorithm was proposed in 2016, a series of improved algorithms of YOLO have been proposed. YOLOv2 [17] was proposed in 2017, and YOLOv3 [18] was proposed in 2018. The YOLOv4 [19] algorithm was proposed in 2020, YOLOv5 was released in June of the same year [20], and Chien-Yao Wang proposed YOLOv7 in July 2022.

2 PROBLEM STATEMENT

Traditional algorithms mostly rely on artificially designed features, which require a large workload, have a great impact on algorithm performance, and do not have good robustness [21]. In recent years, YOLO series has been widely used in fruit detection. The detection accuracy of YOLOv1 is low, and the target location is not accurate [22]. YOLOv2 improves on the YOLOv1 network structure but has lower detection accuracy for small targets [23]. On the basis of YOLOv2, YOLOv3 introduces residual structure, anchor box selection mechanism, multi-scale training method, and other operations. While improving the detection speed, the background false detection rate is low, and the versatility is strong, but the detection position accuracy and recall rate are low [24]. The YOLOv5 model consists of a feature extraction layer, a feature fusion layer, and a detection layer [25]. In YOLOv5, the feature fusion network uses the ideas of FPN and PAN. The model size is small, making it easier to deploy and run. Due to changes in network structure and feature extraction methods, the detection accuracy is slightly lower in some complex scenes. YOLOv7 algorithm introduces strategies such as extended efficient long-range attention network, model scaling, and convolution reparameterization based on concatenation-based models in order to achieve a balance between speed and accuracy [26]. Although YOLOv7 has achieved good object detection speed, it has deficiencies in object detection accuracy, and there will be some false detections or missed detections [27]. Its performance in small target detection is relatively poor, and some target positioning will be inaccurate. In this case, the generalization ability is weak.

M. Lawal et al. [28] used YOLOv3 algorithm to build the YOLO-tomato model by adding a spatial pyramid pooling module and a Mish activation function, which improved the recognition rate of tomatoes to a certain extent. Cao Qiuyang [29] and others replaced the original bounding box regression loss function GIOU with CIOU in YOLOv5, which overcomes the problem of overlap between the prediction frame and the target frame of fruit recognition, and considers the relationship between the aspect ratio and the center point of the target frame and the prediction frame, so as to reduce the difference between the fruit prediction frame and the real frame and make the prediction more accurate. Song Huaibo [30], addressing the challenge of highly similar colors between apple fruit and leaf in the early fruit stage, proposed an improved YOLOv7 model (YOLOv7-ECA) with an efficient channel attention mechanism. It is inserted into the model's three reparameterization paths and achieves local cross-channel interaction between adjacent channels without reducing channel dimensions. Wang Xiaorong [31] improved the YOLOv7 model, creating a safflower sample data set to establish real picking complex environmental data. The model increases the Swin Transformer attention mechanism to improve the robustness of the detection of each classification sample and improves the Focal Loss function to improve the recognition accuracy of unbalanced samples under multi-classification tasks.

With the advantages of the above models, this article essentially analyzes the effectiveness of the convolution parameters, through the improvement of YOLOv7, the parameter quantity and calculation amount of the model are further reduced, and improved the robustness of the target detection system. The improved YOLOv7 is applied to the processing and analysis of fruit images to realize the automatic identification of fruit types and maturity and help the picking and technological upgrading of agricultural products.

3 NETWORK MODEL AND IMPROVEMENTS

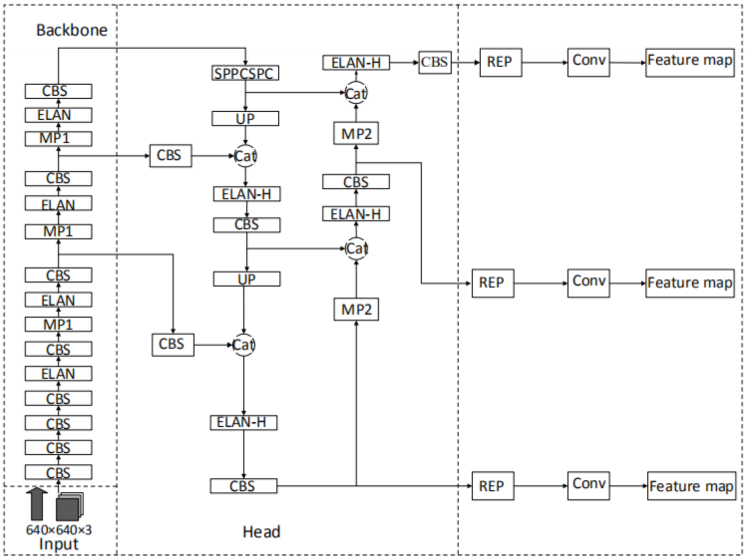

Figure 1: YOLOv7 structure diagram.

3.1 YOLOv7 Target Detection Model

The YOLOv7 network model structure mainly consists of the input end, backbone architecture, head architecture and prediction end. The model structure is shown in Figure 1 [32].

In order to meet the training requirements of the backbone network, the input end scales the input image to the same size. The backbone architecture consists of several cross-stage binary modules, efficient layer aggregation network modules, and MP modules. The cross-stage binary module is composed of a convolutional layer, a batch normalization layer, and a SiLU activation function. The efficient layer aggregation network module is composed of some convolution modules. The MP module is composed of a max-pooling layer and several convolution modules. It has two branches: the feature map is down-sampled respectively so that the image parameters and the number of channels are reduced, and then the feature fusion is performed to optimize the feature extraction ability of the model [32].

3.2 Improvement of the YOLOv7 Network Model

3.2.1 Feature augmentation module design

To suppress the matching noise caused by local similarity between classes and enhance the intra-class correlation in naive correlation, the Feature Augmentation (FA) module is designed.

The most intuitive function of pooling is dimensionality reduction, and the pooling layer does not require training parameters. Therefore, in the design, the local information is obtained by maximum pooling to better retain the features on the feature map, and the parameter quantity is optimized to improve the operation speed. If only the local maximum information of the image is extracted, the problem of excessive loss of the original image information will be caused. Deep convolution can extract all the information from the original image and reduce the feature map. However, because part of the original image is redundant information, such as background information, if all the information of the image is retained, it will be difficult for the module to analyze the data because the amount of information is too complex. If the local extremum information after maximum pooling and the global information after deep convolution are concatenated and used as the input of the next layer of convolution, the images that originally belong to different categories will have the matching results of their local extremum and full image information. Therefore, the module design uses deep convolution together with maximum pooling to retain as many important features as possible and encode the image with local area extreme values and full image information to improve the classification effect of the module and accuracy. The resulting increase in computing time will be solved in the following module design. The feature augmentation module is shown in Figure 2.

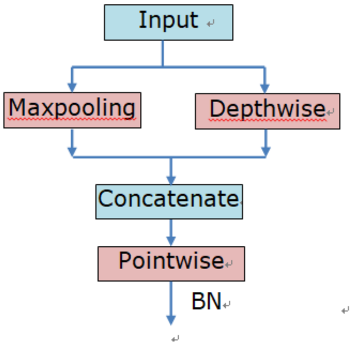

Figure 2: FA-Module.

3.2.2 Residual module design

The residual module is defined as R-SE block. Module design is based on the SE module and layer regularization. A SE module has an attention mechanism, which can automatically obtain the weight of each feature channel through learning, and then use this weight to enhance useful features and suppress features that are less useful for the current task. The Squeeze operation compresses the feature map by global average pooling. The Excitation operation contains two fully connected layers. The SE module uses two fully connected layers, the amount of calculation is not large compared to the convolution layer, but parameters will increase significantly.

Suppose the input and output channels are d , the channel number reduction ratio is K , the parameter amount is P , the calculation amount is $FLOP_S$, P_{LN} is the parameter amount using layer normalization as shown in Equation (1), P_{TWOFC} uses the parameter amount of two fully connected layers as shown in Equation (2), $FLOP_{SLN}$ is the calculation amount of layer normalization as shown in Equation (3), $FLOP_{STWOFC}$ is the calculation amount of two fully connected layers which is showed in Equation (4).

$$P_{LN} = 2d \tag{1}$$

$$P_{TWOFC} = \frac{2d^2}{K} \tag{2}$$

$$FLOP_{SLN} = 4d \tag{3}$$

$$FLOP_{STWOFC} = \frac{4d^2}{K} \tag{4}$$

It can be seen from equations (1)-(4) that if $d > K$, then $P_{TWOFC} > P_{LN}$; if $d > 2K$, then $FLOP_{STWOFC} > FLOP_{SLN}$. Usually, the K of the SE module is 16, and the number of channels of the classic model is more than a few hundred. Therefore, in the design, LN (Layer Normalization) will be used to replace the two fully connected layers in SE, and defined as R-SE. While retaining the different characteristics of the same samples, the parameter amount and calculation amount of the module can be significantly reduced. Therefore, LN (Layer Normalization) is used to replace the two fully connected layers in SE, which is defined as R-SE. While retaining the different characteristics of the same sample, the parameter quantity and calculation amount of the module can be reduced.

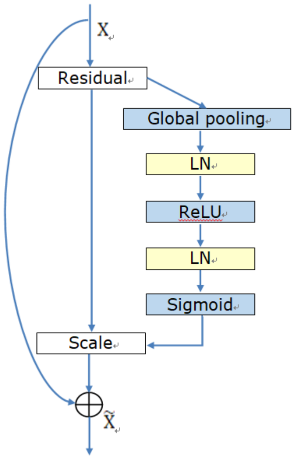

Figure 3: Residual learning unit.

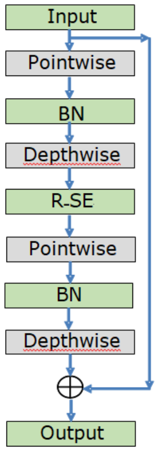

Figure 4: R-SE residual module.

Global average pooling compresses the feature map; two LNs reduced the amount of calculation; the scale operation is to multiply the weight values of each channel calculated by the R-SE module with the two-dimensional matrix of the corresponding channel of the original feature map, and output the results. The residual learning unit is shown in Figure 3, and the residual module introducing R-SE is shown in Figure 4. This design improves the adaptability of the model while reducing over-fitting.

3.2.3 Loss function

In YOLOv7 model, the CIOU loss function is used for bounding box regression. The CIOU loss function introduces the aspect ratio of the prediction boxes and the real boxes in the calculation of the loss value, which avoids the problem of providing the moving direction of the bounding box when the prediction box and the real box do not overlap. However, the CIOU loss function calculates all loss variables as a whole and does not consider the mismatch between the actual target and the prediction box, resulting in slow convergence and instability.

The design converts the output of the model into a probability distribution through the softmax function and then calculates the cross-entropy loss function. Therefore, the cross entropy loss function can be used to deal with multi-classification problems and multi-label problems. When the number of samples is not balanced and the number of samples in some categories is small, the model is easy to overfit a large number of samples, and the calculation method of the cross entropy loss function makes the model pay more attention to the samples with fewer categories, thus improving the generalization ability of the model. Cross-entropy softmax loss are shown in Equation (5), (6), and (7).

$$\text{Cross Entropy} \qquad L=\frac{1}{n}\sum_{i=1}^{n}(-\sum_{j=1}^{C}y_{i,j}\log p_{i,j}) \tag{5}$$

$$\text{Softmax Function} \qquad p_{i,j}=\frac{e^{l_{i,j}}}{\sum_{j=1}^{C}e^{l_{i,j}}} \tag{6}$$

$$\text{Softmax Loss} \qquad L=\frac{1}{n}\sum_{i=1}^{n}(log\frac{e^{l_{i,Y_{(i)}}}}{\sum_{j=1}^{C}e^{l_{i,j}}}) \tag{7}$$

In this size is represented by n; the number of classes is represented by C ; $y_{i,j}$ represents the true label of the i-th sample on the j-th class is represented by; $p_{i,j}$ represents the prediction of the j-th class by the i-th sample Probability; $l_{i,j}$ represents the output logit of the neural network for the i-th sample for the j-th category; $Y_{(i)}$ is the category to which the i-th sample belongs.

3.3 Model Integration

The improved YOLOv7 model first performs feature enhancement on the original image at the input end. Secondly, the model uses the R-SE module to replace the last convolutional layer in the backbone architecture. If the R-SE module is used in the front of the backbone architecture, the calculation amount of the self-attention mechanism is squared to the resolution of the feature map, and the calculation amount will be greatly increased, thereby reducing the detection speed of the model. The R-SE module can show that the interdependence between the convolutional feature channels improves the representation ability of the network, enabling the network to achieve feature recalibration and selectively enhance information features. Third, the use of cross-entropy loss softmax loss can measure the prediction accuracy of the model, which can make the model converge faster, and can be updated online to make the model more optimized. Figure 5 is the improved YOLOv7 structure.

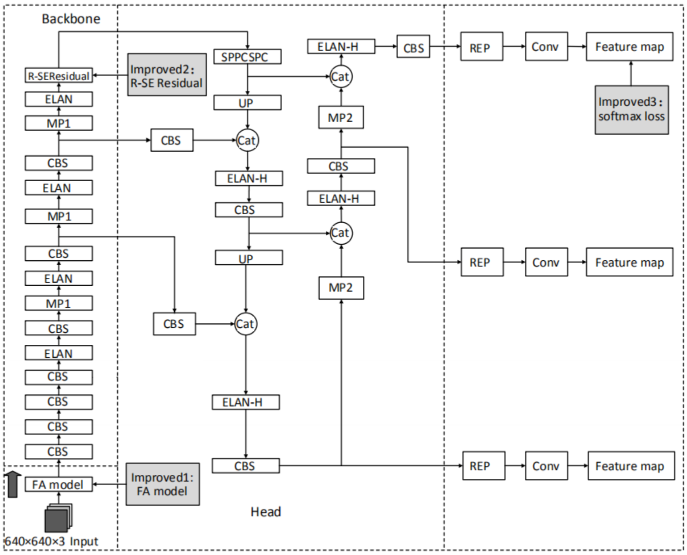

Figure 5: Improved YOLOv7 structure.

4 EXPERIMENTS AND RESULT ANALYSIS

4.1 Experimental Environment

To obtain objective experimental results, all experiments were conducted under the same conditions: the operating system version of the experimental platform was Windows 10 (64-bit), and the GPU was GeForceRTX2080Ti. The deep learning framework used in the experiment is Pytorch, version 2.0, and the programming language is Python3.9.

4.1.1 Data set production

To train the model, many well-labeled data sets are needed so that the neural network can learn the characteristics of the samples. These labeled data sets need to cover fruit samples of various shapes, colors, and sizes to ensure the generalization ability of the model. The fruit recognition and detection data set established in this article is based on the Python web crawler. The corresponding images are obtained on the Baidu image platform through the keywords "ripe apples", "immature apples", etc., and then the images that are too blurry and do not meet the detection category are removed and finally get 1,000 pictures, and the data were uniformly named in .jpg format, named after Apple, Banana, and Orange category labels + picture numbers.

For the labeling of apples and oranges, a single whole frame of the fruit needs to be selected. For bananas since the shape of bananas is curved and bananas often appear in the form of clusters, a whole cluster of bananas needs to be selected. In addition, because the ripe characteristics of apples, oranges, and bananas can be easily seen with the naked eye, the maturity can be judged and labeled based on the color characteristics of each fruit and labeling the data set with the labeling tool Labelimg.

The label format obtained after the data set annotation is completed is VOC, but the data set reading method in YOLO is not an XML file in VOC format but rather a txt file in YOLO format. Therefore, a Python script is used to convert the obtained label format while converting the YOLO format, 800 images were randomly selected from the data set as a training set, and both the validation set and the test set selected 100 images in a ratio of 8:1:1.

4.1.2 Evaluation index selection

In order to verify the effectiveness of the model, precision, recall, mean average precision (mAP) and F1-score were used to quantitatively evaluate the experimental results. The equations are shown in (8), (9), (10) and (11).

$$Precision = {TP}/{(TP+FP)} \tag{8}$$

$$Recall = {TP}/{(TP+FN)} \tag{9}$$

$$mAP = \sum_{i=1}^{C} {AP_i}/{C} \tag{10}$$

$$F1-score = {2\times Precision\times Recall}/{(Precision+Recall)} \tag{11}$$

In formulas, TP is the correctly classified positive samples, FP is the misclassified positive samples, FN is the misclassified negative samples, and C is the total number of categories.

4.2 Results and Analysis

This paper trains four network models: YOLOv3, YOLOv5, YOLOv7, and improved YOLOv7 on the same fruit data set. The input image size is 640×640, the batch size is 16, the number of iterations Epoch is 150, the optimizer is Adam, and the initial learning rate is set to 0.01.

According to the log files generated during the training process, the results of the YOLOv3, YOLOv5, YOLOv7, and improved R-SE-YOLOv7 models are drawn, which are shown in Figures 6, 7, 8, and 9.

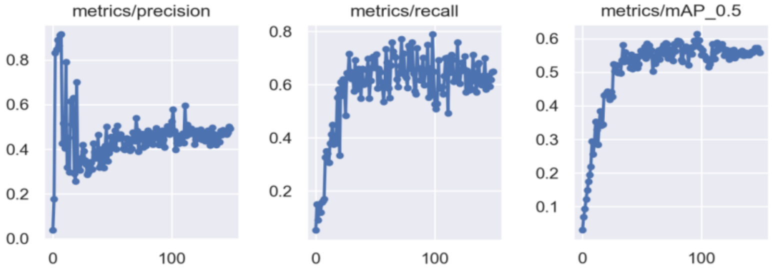

Figure 6: YOLOv3 training results.

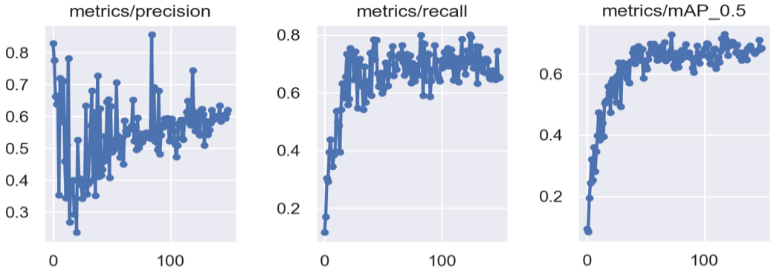

Figure 7: YOLOv5 training results.

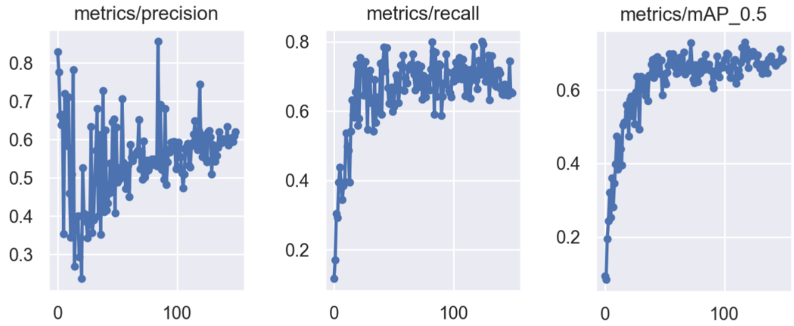

Figure 8: YOLOv7 training results.

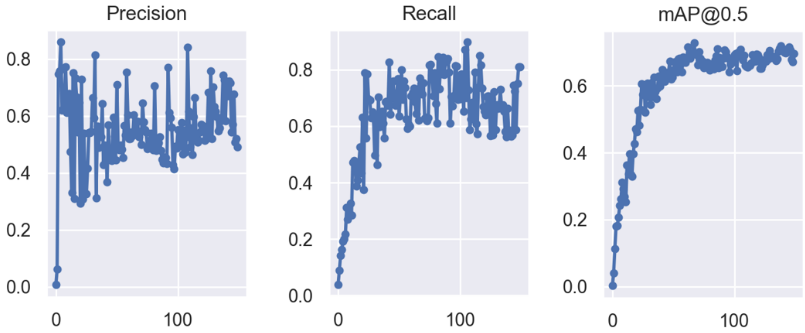

Figure 9: R-SE-YOLOv7 training results.

From the curves in Figures 6-9, it can be observed that the accuracy of YOLOv3, YOLOv5, YOLOv7, and improved R-SE-YOLOv7 models continuously improves. Although the fluctuations of each model in the first 50 rounds were large, they all converged in the end within a small range; the model's recall rate fluctuates less and shows an overall upward trend; the model's average accuracy continues to improve, indicating that the model's learning ability is gradually increasing; From the F1-score calculation considering both accuracy and recall, it can be seen that the F1 scores of the four models from low to high are 0.56, 0.67, 0.70, and 0.75 respectively, that is, the improvement from YOLOv3 to YOLOv5 is the highest, an increase of 0.11, the improvement from YOLOv5 to YOLOv7 is relatively stable, an increase of 0.03, and the improvement from YOLOv7 The improvement of YOLOv7 has been improved by 0.05, which generally shows that the indicators of the model is showing an optimization trend.

Before training, use Labelimg to annotate pictures, as shown in Figure 10. In the picture, fruits are divided into low, medium, and high maturity according to color, and are associated with the corresponding fruit categories. After comparing the four models of YOLOv3, YOLOv5, YOLOv7, and improved R-SE-YOLOv7 in terms of fruit category and ripeness detection, it was found that the four models have their own characteristics. For the identification of fruit categories and ripeness, the performance of the four models is relatively stable and can meet most detection needs. From the comparison of Figures 11, 12, 13, and 14, when it comes to overlapping targets, the YOLOv3 model will have certain problems in the recognition of bananas. It is difficult to select accurately, and the confidence level is relatively low. For citrus, YOLOv3 performs well, can accurately identify overlapping targets, and can accurately identify targets at different distances. In contrast, YOLOv5 has deviations in the detection of ripeness in apples, and there is also the problem of inaccurate frame selection, but the recognition of the other two fruits is relatively accurate. YOLOv7 can accurately identify bananas, and the detection of apple ripeness is also stable, but there is a problem of missed detection in the recognition of distant targets. The improved R-SE-YOLOv7 performs stably in both the recognition of near and far targets and the recognition of fruit maturity.

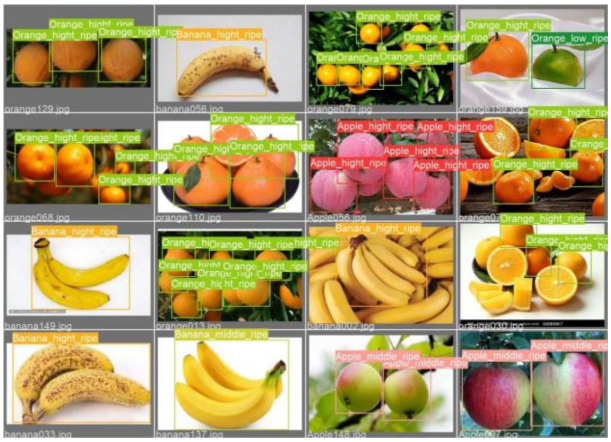

Figure 10: Pictures annotated using labeling before training.

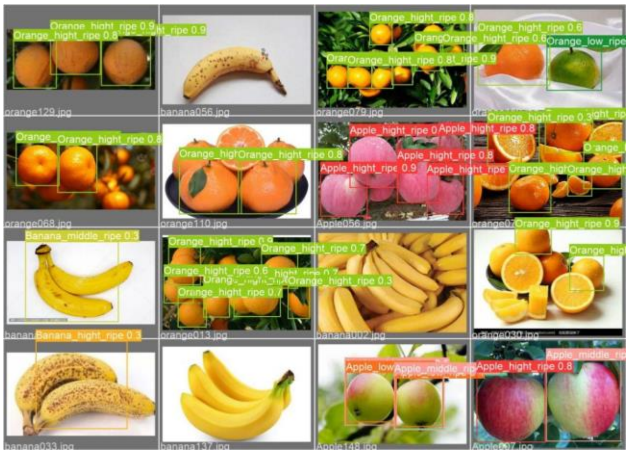

Figure 11: YOLOv3 detection results.

In order to verify the feasibility of the improved YOLOv7 model, some pictures are selected from the data set to further compare the detection effects of YOLOv7 and the improved R-SE-YOLOv7.From the comparison of Figures 15 and 16, we can see that both YOLOv7 and improved R-SE-YOLOv7 can detect the category and maturity of fruits. However, after comparison, it is found that the improved R-SE-YOLOv7 model is more accurate in the frame selection of fruits, and the confidence level in identifying multiple types of fruits is also lower. There has been a huge improvement, almost all reaching more than 90%. Therefore, it is proved that the improved R-SE-YOLOv7 model can normally identify fruit types and maturity, and the detection accuracy is also good.

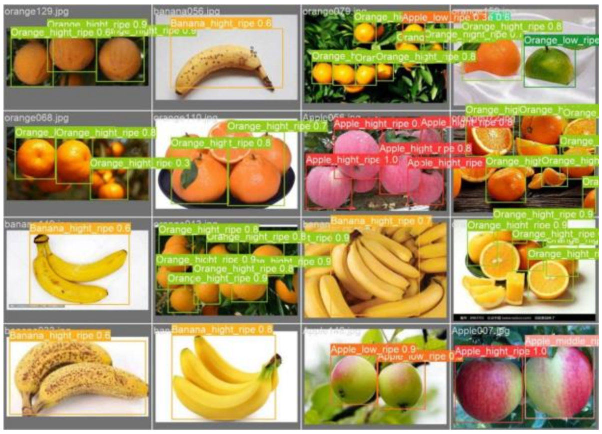

Figure 12: YOLOv5 detection results.

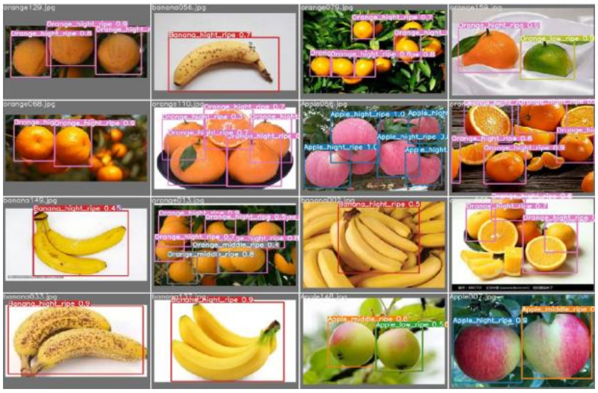

Figure 13: YOLOv7 detection results 1.

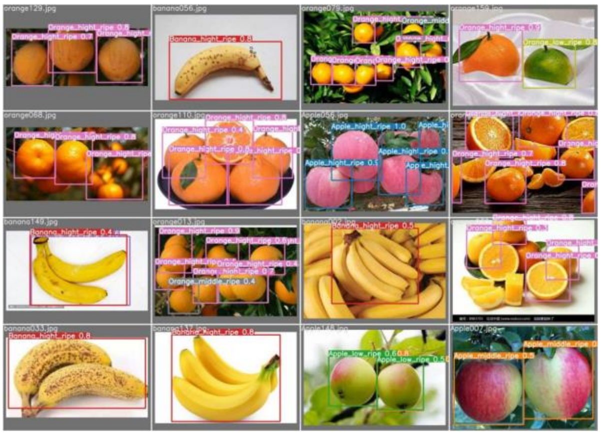

Figure 14: R-SE-YOLOv7 detection results 1.

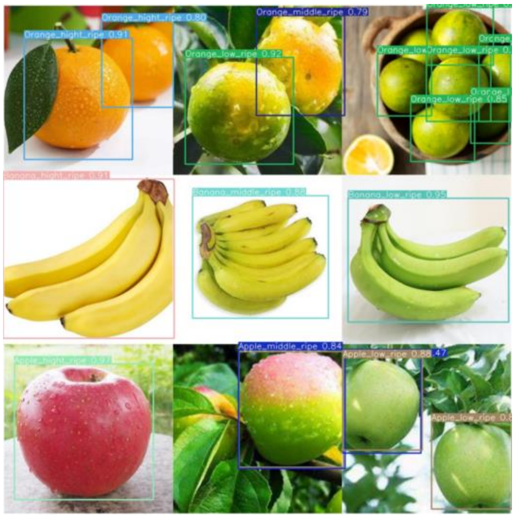

Figure 15: YOLOv7 Detection Results 2

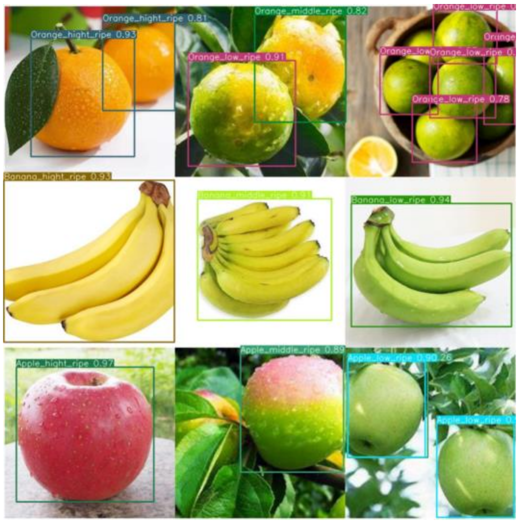

Figure 16: R-SE-YOLOv7 Detection Results 2

The evaluation index data are shown in Table I. According to the comparison of table data, within the same number of iterations, the recognition accuracy of YOLOv7 is improved by 8 % compared with that before improvement by introducing the FA module to encode local region extremum and whole image information. The R-SE residual module was introduced, which improved the recall rate of YOLOv7 by 11.1%, mAP by 8.2%, and F1-score by 0.05 while retaining different features of the same sample. Considering the above parameters, the improved R-SE-YOLOv7 model performs better, which verifies the effectiveness of the improved model.

Model	*Precision/%*	*Recall/%*	*mAP/%*	*F1-core*	*Params*	*TestingTime（ms）*

YOLOV3	51.2	61.3	58.2	0.56	8,685,172	482
YOLOV5	62.0	72.4	65.0	0.67	7,034,398	629
YOLOV7	72.1	68.1	73.1	0.70	6,029,244	531
R-SE-YOLOv7	80.1	79.2	81.3	0.75	6,078,643	535

Table 1: Evaluation Index Data for Four Models

5 CONCLUSIONS

To improve the level of fruit intelligence detection, this paper proposes an improved YOLOv7 fruit target recognition and maturity detection algorithm. This algorithm first performs feature enhancement on the original image at the input end, applies maximum pooling to retain important features of the image, applies deep convolution to prevent information loss, enhances the intra-class correlation, and the model classification accuracy is improved; then uses Layer Normalization replaces the fully connected layer in the SE mechanism and applies this to the residual module design to reduce the increase in model parameters and calculations caused by the feature enhancement module, while improving the accuracy of detection. Finally, the cross entropy loss function is designed in the model to accelerate the convergence speed and improve the accuracy of fruit recognition. YOLOv3, YOLOv5, YOLOv7 and improved R-SE-YOLOv7 were used to detect the category and maturity of apple, citrus and banana on the same data set. The experimental results show that although the improved YOLOv7 is not as good as YOLOv3 in terms of detection time, it is better than YOLOv5 and close to YOLOv7. In terms of Precision, Recall, mAP, and F1-core, the improved R-SE-YOLOv7 performed better, achieving the goal of improving model recognition accuracy and verifying the effectiveness of model improvements.

The proposed algorithm can provide more accurate and efficient solutions for the field of fruit intelligence detection. Future work could explore further optimizations and extensions of the proposed algorithm to address additional challenges and scenarios in the realm of object detection and recognition.

Qi Chen, https://orcid.org/0009-0003-8210-7978
Renjie Li, https://orcid.org/0009-0005-3488-1146
Lianjun Hu, https://orcid.org/0000-0001-5463-2685
Yong Zhang, https://orcid.org/0000-0001-8854-1319

ACKNOWLEDGEMENT

This article was funded by The Natural Science Foundation of Tianjin (grant number 23JCQNJC00840).

REFERENCES

[1] Tian Y.; Yan C. G.; Wnag Z.; et al.: Instance segmentation of apple flowers using the improved Mask R-CNN model, Biosystems Engineering, 193, 2020, 264-278. https://doi.org/10.1016/j.biosystemseng.2020.03.008

[2] Lecun Y.; Bengio Y.; Hinton G: Deep learning, Nature, 521(7553), 2015,436 - 444. https://doi.org/10.1038/nature14539.

[3] Zhao Y. Q.; Rao Y.; Dong S.; et al.: Survey on deep learning object detection, Journal of Image and Graphics, 25(4), 2020, 629 - 654. http://dx.doi.org/10.11992/tis.202004033.

[4] Yong Z.; Renjie L.; Fenghong W.; Weijing Z.; et al.: An Autonomous Navigation Strategy Based on Improved Hector SLAM With Dynamic Weighted A* Algorithm, IEEE Access, 11, 2023, 79553-79571. https://doi: 10.1109/ACCESS.2023.3299293.

[5] Yong Z.; Renjie L.; Qi C.; Derui Z.; et al.: An improved bicubic interpolation SLAM algorithm based on multisensor fusion method for rescue robot, International Journal of Sensor Networks, 42(2), 2023, 125-136. https://doi.org/10.1504/ijsnet.2023.131656

[6] Yong Z.; Xinyue L.; Li W.; et al.: An Autocorrelation Incremental Fuzzy Clustering Framework Based on Dynamic Conditional Scoring Model, Information Sciences, 648(11), 2023,119567. https://doi.org/10.1016/j.ins.2023.119567

[7] Kamilaris A.; PrenafetaF. X.: Deep learning in agriculture: a survey, Computers and Electronics in Agriculture, 147, 2018,70 -90. https://doi.org/10.1016/j.compag.2018.02.016

[8] Li Y.; Feng Q. C.; Li T.; et al.: Advance of target visual information acquisition technology for fresh fruit robotic harvesting: A review, Agronomy, 12(6), 2022, 1336. https://doi.org/10.3390/agronomy12061336

[9] Zhu L.; Xie Z.; Luo J.; et al.: Dynamic object detection algorithm based on lightweight shared feature pyramid, Remote Sensing, 13(22), 2021, 4610. https://doi.org/10.3390/rs13224610

[10] Tong K.; Wu Y.; Zhou F.: Recent advances in small object detection based on deep learning: a review, Image and Vision Computing, 97, 2020, 103910. https://doi.org/10.1016/j.imavis.2020.103910

[11] Fang W.; Wang L.; Ren P. M.: Tinier-YOLO: A real-time object detection method for constrained environments, IEEE Access, 8, 2020, 1935-1944. https://doi.org/10.1109/ACCESS.2019.2961959

[12] 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. https://doi.org/10.48550/arXiv.1311.2524

[13] Zaidi S. S. A.; Ansari M. S.; Aslam A.; et al.: A survey of modern deep learning based object detection models, Digital Signal Processing, 126, 2022, 103514. https://doi.org/10.1016/j.dsp.2022.103514

[14] Wu X.; Sahoo D.; Hoi S. C. H.: Recent advances in deep learning for object detection, Neurocomputing, 396; 2020, 39 -64. https://doi.org/10.1016/j.neucom.2020.01.085

[15] Xiao Y.; Tian Z.; Yu J.; et al.: A review of object detection based on deep learning, Multimedia Tools and Applications, 79(33), 2020, 23729-23791. https://doi.org/10.1007/s11042-020-08976-6

[16] Redmon J.; Divvala S.; Girshick R.; et al.: You Only Look Once: Unified, Real-Time Object Detection, Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition, 2016,779-788. https://doi.org/10.1109/CVPR.2016.91

[17] Redmon J.; Farhadi A.: YOLO9000: Better, Faster, Stronger, Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition, 2017, 6517-6525. https://doi.org/10.1109/CVPR.2017.690

[18] Redmon J.; Farhadi A.: YOLOv3: An Incremental Improvement, Computer Science, 4(1), 2018, 1-6. https://doi.org/10.48550/arXiv.1804.02767

[19] Bochkovskiy A.; Wang C.; Liao H.: YOLOv4: Optimal Speed and Accuracy of Object Detection, arXiv 2020, 2004, 10934. https://doi.org/10.48550/arXiv.2004.10934

[20] Cao M.; Fu H.; Zhu J.; et al.: Lightweight tea bud recognition network integrating GhostNet and YOLOv5, Mathematical Biosciences and Engineering, 19(12), 2022,12897-12914. https://doi.org/10.3934/mbe.2022602

[21] Mo Y.; Wu Y.; Yang X.; et al.: Review the state-of-the-art technologies of semantic segmentation based on deep learning , Neurocomputing, 493, 2022, 626-646. https://doi.org/10.1016/j.neucom.2022.01.005

[22] Redmon J.; Divvala S. K.; Girshick R.; et al.: You Only Look Once: Unified, Real-Time Object Detection, Computer Vision and Pattern Recognition, 2016, 779-788. https://doi.org/10.48550/arXiv.1506.02640

[23] Zhao J.; Qu J.: A detection method for tomato fruit common physiological diseases based on YOLOv2, 10th international conference on Information Technology in Medicine and Education, 2019, 559-563. https://doi.org/10.1109/ITME.2019.00132

[24] Liu G. X.; Nouaze J. C.; Mbouembe P. L. T.; et al: YOLO-Tomato: A Robust Algorithm for Tomato Detection Based on YOLOv3, Sensors, 20(7), 2020, 1-20. https://doi.org/10.3390/s20072145

[25] Lin G. C.; Tang Y. C.; Zou X. J.; et al.: In-field citrus detection and localisation based on RGB-D image analysis, Biosystems Engineering, 186, 2019, 34-44. https://doi.org/10.1016/j.biosystemseng.2019.06.019

[26] Wang C. L.; Tang Y. C.; Zou X. J.; et al.: A robust fruit image segmentation algorithm against varying illumination for vision system of fruit harvesting robot, Optik, 131, 2017, 626-631. https://doi.org/10.1016/J.IJLEO.2016.11.177

[27] Zhao Z.; Wei H. F.; Huang Y.; et al.: Online real-time detection system for cracked eggs using improved YOLOv7, Transactions of the Chinese Society of Agricultural Engineering (Transactions of the CSAE), 39(20), 2023,255-265. http://dx.doi.org/10.11975/j.issn.1002-6819.202305228

[28] Olarewaju L. M.: Tomato Detection Based on Modified YOLOv3 Framework, Scientific Reports, 11(1), 2021, 1-11. https://doi.org/10.1038/s41598-021-81216-5

[29] Cao Q. Y.; Shao Y. Q.; Yin H.: Automatic fruit recognition based on attention YOLOv5 model, Computer System & Applications, 31(07), 2022, 333-340. http://dx.doi.org/10.15888/j.cnki.csa.008576

[30] Song H. B.; Ma B. L.; Shang Y. Y.; Wen Y. C.; Zhang S. J.: Detection of Young Apple Fruits Based on YOLO v7-ECA Model, Transactions of the Chinese Society for Agricultural Machinery, 54(6), 2023,233-242. http://dx.doi.org/10.6041/j.issn.1000-1298.2023.06.024

[31] Wang X. R.; Xu Y.; Zhou J. P.; Chen J. R.: Safflower picking recognition in complex environments based on an improved YOLOv7, Transactions of the Chinese Society of Agricultural Engineering (Transactions of the CSAE), 39(6), 2023,169-176. http://dx.doi.org/10.11975/j.issn.1002-6819.202211164

[32] Ni C. S.; Li L.; Luo W. T.; Qin Y.; Yang Z.; Fu Y.: Disease Detection of Asphalt Pavement Based on Improved YOLOv7, Computer Engineering and Applications, 59(13), 2023, 305-316. https://doi.org/10.3778/j.issn.1002-8331.2301-0098