STRAWBERRY FRUIT DETECTION METHOD BASED ON IMPROVED YOLOv8N / 基于改进 YOLOv8N 的草莓果实检测方法

Zhenwei LI [1], Suyun LI [1], Wenting LAN [2], Shide LI [3], Yanguan CHEN [1], Pengcheng LV [4,*]
[1] Hezhou University, School of Artificial Intelligence, HeZhou, China
[2] Hezhou University, Academic Affairs Office, HeZhou, China
[3] Guangxi Medical University, School of Information and Management, NanNing, China
[4] Shandong University of Technology, College of Agricultural Engineering and Food Science, ZiBo, China
Tel: +86 13964379460; E-mail: wslpc1999@163.com
Corresponding author: Pengcheng LV

DOI: https://doi.org/10.35633/inmateh-76-59

Keywords: *Deep learning; attention mechanism; YOLOv8n; detection system*

ABSTRACT

As an economic crop of Rosaceae family, strawberry has the advantages of short reproductive cycle, wide ecological adaptability and significant economic benefits, and its planting industry has been rapidly developed in recent years. Aiming at the low efficiency and high labor cost of traditional manual picking detection methods in the intelligent transformation of strawberry industry, this study innovatively proposes a strawberry fruit intelligent detection system based on YOLOV8N. By introducing RFAConv dynamic sensory field convolution, SENet channel attention mechanism and InceptionNeXt lightweight structure, combined with Wise-IoU loss function and DIoU-NMS post-processing algorithm, the synergistic enhancement of detection accuracy and computational efficiency is realized. The ablation experiments show that the improved model has a precision rate of 95.92%, a recall rate of 95.45%, and a mAP50 of 98.29% on the strawberry dataset, which are 4.14%, 3.31%, and 1.55% higher than that of the baseline model, respectively, while the number of model parameters is compressed to 5.17 M (a reduction of 12.96%). This research can provide technical support for intelligent strawberry picking.

摘要

草莓作为蔷薇科经济作物，具有生育周期短、生态适应性广及经济效益显著等优势，其种植产业近年来得到了快速发展。本研究针对草莓产业智能化转型中传统人工采摘检测方法存在的效率低下、人工成本高等痛点，创新性提出一种基于 YOLOV8N 的草莓果实智能检测系统。通过引入 RFAConv 动态感受野卷积、SENet 通道注意力机制及 InceptionNeXt 轻量化结构，结合 Wise-IoU 损失函数与 DIoU-NMS 后处理算法，实现了检测精度与计算效率的协同提升。消融实验表明，改进后模型在草莓数据集上精确率达 95.92%、召回率为 95.45%、mAP50 达 98.29%，较基线模型分别提升 4.14%、3.31%和 1.55%，同时模型参数量压缩至 5.17M（减少12.96%），该研究可为草莓智能化采摘提供技术支持。

INTRODUCTION

As a Rosaceae cash crop, strawberry has become the preferred crop for facility agriculture by virtue of its short reproductive cycle, wide ecological adaptability and significant economic benefits. Its fruit is generally small and dense. Dense canopy structure is easy to cause visual masking, coupled with the maturity gradient distribution phenomenon, resulting in the traditional artificial observation and picking both time-consuming and laborious, significantly increasing the difficulty of selective harvesting. Nowadays, the level of intelligence in China's strawberry industry is constantly improving, and the development of a high-precision fruit ripeness detection algorithm can not only accurately find out the distribution area of fruits with different ripeness levels, but also monitor and optimize the picking process.

With the iterative upgrading of computer vision technology, many innovative research results have emerged in the field of agricultural target detection, providing important technical support for intelligent crop management. In the direction of fruit multi-category recognition, *Wan et al., (2021)* constructed a multi-spectral visual analysis system based on the optimization of Faster R-CNN architecture, which effectively solved the feature confusion problem of apple, pear and peach fruits under the complex background of the orchard by reconfiguring the parameter distribution of the convolution kernel with the improvement of the downsampling strategy.

The experimental data show that the classification performance of the model for the above fruits reaches 92.51%, 88.94%, and 90.73%, respectively, and the overall mean average precision (mAP) is improved to 90.72%, which validates the generalization ability of deep networks in cross-species recognition tasks.

Notably, *Sharma et al., (2022),* developed a pineapple maturity analysis model based on the YOLOv5 framework for tropical crop characteristics, and constructed a three-stage discrimination system containing the green ripening stage, the color-turning stage, and the complete ripening stage by integrating spectral reflectance features and morphological parameters, and the classification accuracy of the final ripening stage discrimination exceeded 95%, which provided a new paradigm for the intelligent assessment of fruit quality.

In terms of feature enhancement technology path, *Lu et al., (2021),* researchers designed an attention-guided multi-scale feature fusion mechanism, which significantly improved the spatial localization accuracy of small-scale fruits by implementing cross-layer semantic enhancement to the shallow feature map of SSD network. The improved model mAP metrics improve by 29.2 percentage points from the baseline, especially the leakage rate in the fruit-dense region decreases by 18.6%, demonstrating excellent scene adaptability.

As for mobile deployment, the MobileNet lightweight architecture proposed by *Howard et al., (2019),* pioneered the use of depth-separable convolution instead of standard convolution operation, and compressed the model computation to about one-eighth of the traditional network by decoupling the feature learning process between the channel dimension and the spatial dimension, laying down a basic engineering framework for the development of embedded vision systems for orchard inspection UAVs.

The improved YOLOv5 model developed by *Peng et al., (2014)* introduces the Ghost module to reconfigure the feature extraction backbone network, combines the coordinate attention mechanism to strengthen the feature response in the key regions of the fruit, and also uses the SIoU loss function to optimize the bounding box regression process. After testing, the scheme achieves 94.8% mAP for strawberry target detection, the model volume is reduced to 67% of the original structure, and a real-time processing capability of 23.6 frames per second (FPS) is realized on the Jetson Nano embedded platform. These technological breakthroughs not only validate the application potential of the lightweight model in agricultural scenarios, but also provide key algorithmic components for the construction of a "cloud-edge-end" synergistic smart agriculture sensing system *(Elsayed et al., 2024; Hu et al., 2018; Hosna et al., 2022; Li et al., 2020).*

In this paper, a multi-source heterogeneous dataset for strawberry fruit detection under varying ripeness levels, lighting conditions, and complex backgrounds is first constructed, and a complete deep learning research framework is established. In the backbone network, RFAConv is used to replace the standard convolutional Conv to improve the performance of the network; the SENet attention mechanism is introduced to strengthen the learning ability of the backbone network on the color and shape features of strawberries, and at the same time, the design of the decoupling head is optimized to better adapt to the needs of classification and regression tasks. The InceptionNeXt neural network structure is introduced as the feature extraction network, and the convolution in Bottleneck in the Cf2 module of the neck network is replaced with the InceptionNeXt convolution, which effectively reduces the amount of computation. During the training process, the Wise-IoU loss function with dynamic nonmonotonic focusing mechanism is used to address the impact of low-quality samples on model performance in the target detection task. In the post-processing stage, DIoU-NMS is used to replace the traditional NMS algorithm to reduce the false deletion of overlapping target frames.

MATERIALS AND METHODS

Data Acquisition and Pre-processing

The data used in this study are mainly derived from online public data and strawberry garden collection data. This online public dataset was captured inside a greenhouse and contains 3,500 images. The advantage of strawberry orchard collection data is that these data reflect the actual farmland environment and crop growth conditions, with the advantage of authenticity and credibility.

The image acquisition device is the camera that comes with the Xiaomi 14 cell phone, and the effective pixel of the camera is 50 million. The shooting test data were images of strawberries with different growth cycles, such as front, side and shade, which were taken on February 4, 2024.

Figure 1 shows the strawberry image dataset.

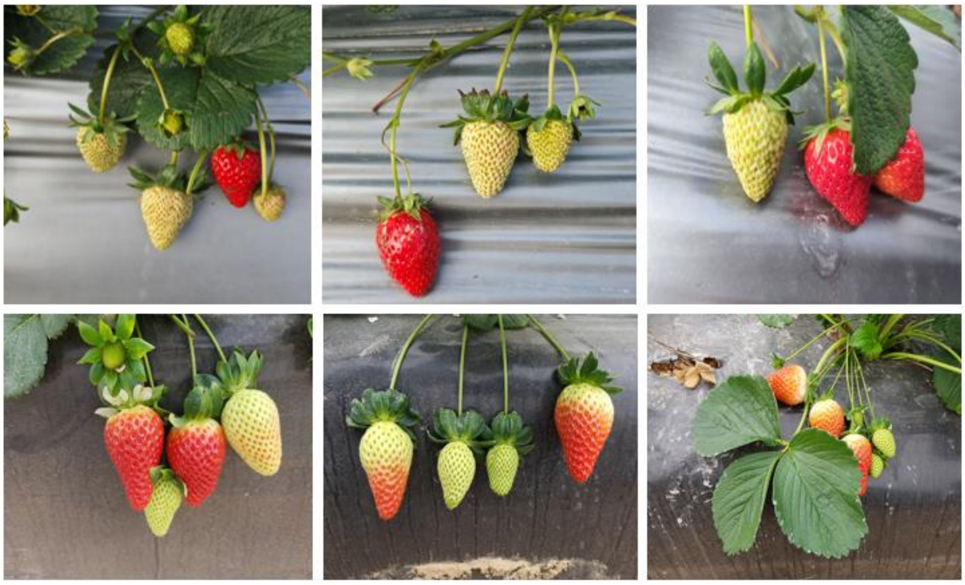

Fig. 1 – Example graph of the strawberry dataset

In order to increase the speed of model training, the collected raw images are preprocessed with uniformity. At the same time, 500 typical samples were selected from a database of 3,500 annotated samples using a stratified random sampling method to implement a data augmentation strategy, and the dataset was expanded by pre-processing the images by rotating, panning, mirroring and adding noise. After augmentation, the total number of images increased to 5,500, with the enhanced portion contributing 2,000 images. The dataset contains 24619 labeled strawberry objects, including 19162 unripe strawberries, 4364 semi-ripe strawberries, and 5021 ripe strawberries. The statistical data are shown in Figure 2.

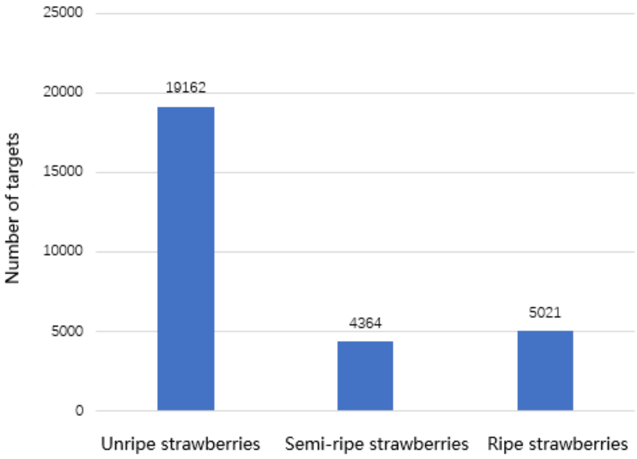

Fig. 2 – Strawberry fruit ripening statistics

The dataset was divided into training, validation and test sets according to the ratio of 8:1:1, i.e., 4400 images for the training set and 550 images each for the validation and test sets.

YOLOv8 Network Architecture

YOLOv8, an integrated and enhanced iteration of the YOLO series, incorporates a core feature of an anchor-free detection mechanism. This mechanism directly predicts the center position of the target without reliance on predefined anchor boxes, thereby simplifying the training process and accelerating the post-processing steps of non-maximum suppression (NMS). Depending on the depth and width of the network, different versions of YOLOv8n, YOLOv8s, YOLOv8m, YOLOv8l and YOLOv8x are used. Although the overall architecture of these models remains consistent, the number of modules and the configuration of the convolutional layers within each version differ significantly when implementing the network training process *(Vaswani et al., 2017; Wang et al., 2024; Zhu et al., 2024)*.

This difference mainly stems from considerations of model size, complexity, total number of parameters, computational burden, and training methods. By adjusting these factors, different versions of YOLOv8 are able to strike a unique balance between performance and computational resource consumption. The comparison data of different versions for detection on the public dataset COCO are shown in Table 1. In this study, YOLOv8n, which has a smaller weight file, faster inference speed, and is suitable for deployment to edge devices, is chosen as the base network.

Table 1

Comparison of data of different versions of YOLOv8

Mold	Depth	Width	Px	mAP	M	B
YOLOv8n	0.33	0.25	640	37.3	3.2	8.7
YOLOv8s	0.33	0.5	640	44.9	11.2	28.6
YOLOv8m	0.67	0.75	640	50.2	25.9	78.9
YOLOv8l	1.0	1.0	640	52.9	43.7	165.2
YOLOv8x	1.0	1.25	640	53.9	68.2	257.8

The structure of YOLOv8 consists of a backbone network, a neck network, and a predictive head network as shown in Figure 3.

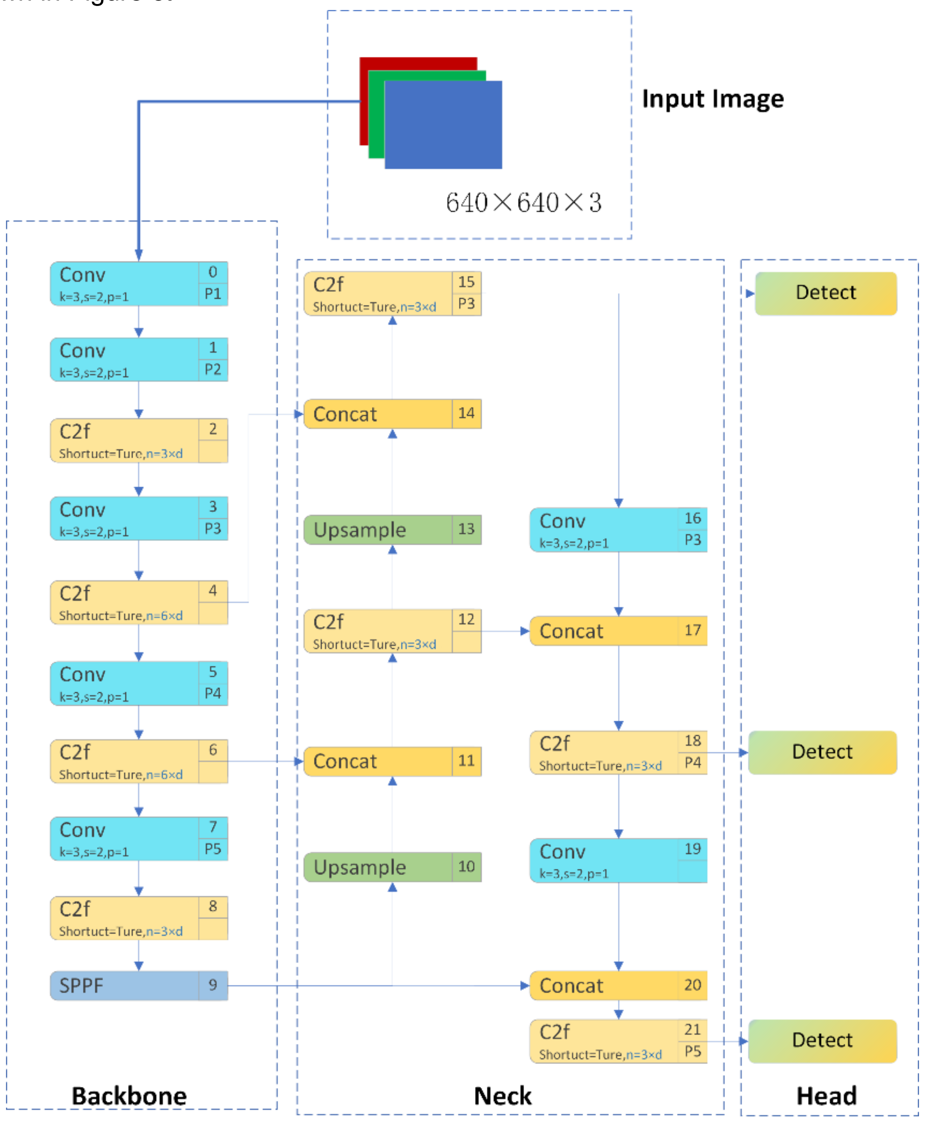

Fig. 3 – YOLOv8 network structure diagram

Study of Improved YOLOv8n Algorithm

For the strawberry fruit detection task, this study proposes a strawberry fruit detection method based on improved YOLOv8n to enhance the performance of the model in specific scenes. The use of receptive-field aware convolution RFAConv instead of standard convolution Conv in the backbone network improves the network performance, and the detection performance is significantly improved; the introduction of the SENet attention mechanism strengthens the learning ability of the backbone network for strawberry color and shape features, and the design of the decoupling head is also optimized to better adapt to the requirements of the classification and regression tasks. The InceptionNeXt neural network structure is introduced as the feature extraction network, and the convolution in Bottleneck in the Cf2 module of the neck network is replaced with the InceptionNeXt convolution, which effectively reduces the amount of computation. During the training process, the Wise-IoU loss function with dynamic nonmonotonic focusing mechanism is used to address the impact of low-quality samples on model performance in the target detection task. In the post-processing stage, DIoU-NMS is used to replace the traditional NMS algorithm to reduce the false deletion of overlapping target frames. The improved YOLOv8n network structure is shown in Fig. 4.

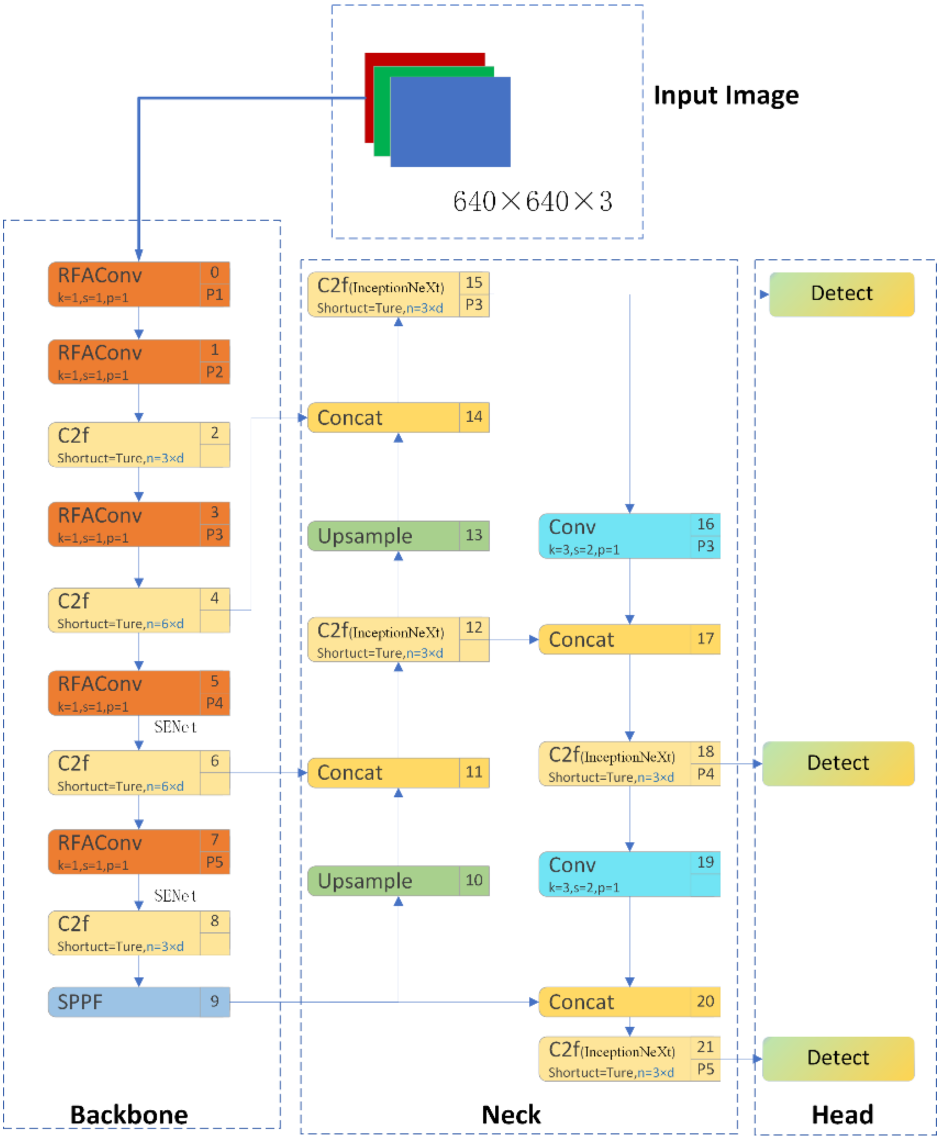

Fig. 4 – Improved YOLOv8n network architecture

Traditional convolution operation extracts local features through a fixed convolution kernel, and its receptive field is usually limited and fixed, which may lead to poor adaptability to multi-scale targets. RFAConv enables the convolution operation to adaptively adjust the size of the receptive field by introducing a dynamic receptive field mechanism and an attentional mechanism, so that it can better capture the target features at different scales. The principle of spatial feature transformation of the receptive field is shown in Fig. 5.

The original spatial feature has a size of C×H×W and is divided into non-overlapping sliding windows. When a 3×3 convolution kernel is applied, each 3×3 window in the receptive field corresponds to a local region of the input feature map. As a result, the converted spatial feature is expanded by a factor of three in both height and width, increasing its size to 3C×3H×3W. This enlargement effectively broadens the receptive field, enabling the capture of multi-scale information.

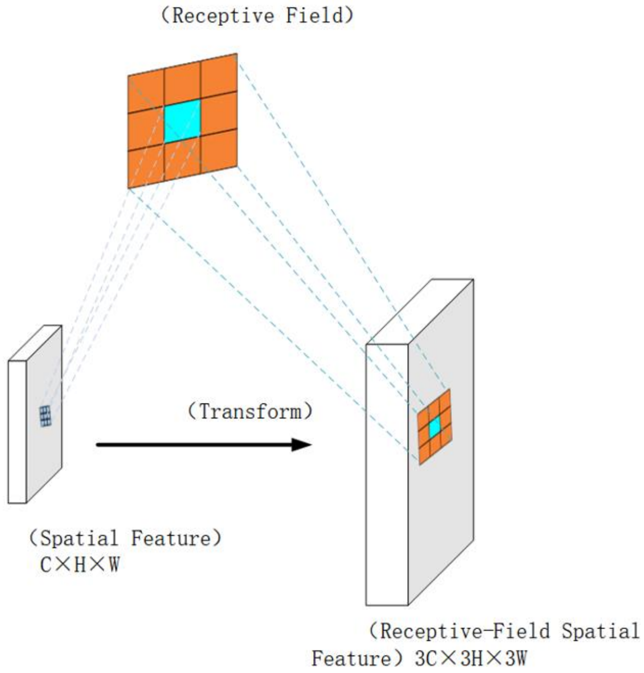

Fig. 5 – Principle of spatial feature transformation in the sensory field

Spatial attention mechanisms that focus on spatial features in the sensory field are combined with convolution to eliminate the problem of sharing convolution parameters. Current spatial attention mechanisms already consider long-range information, which can be obtained globally through global average pooling or global maximum pooling. In order to focus on feeling the wild spatial features, a k × k convolution operation with stride of k is used to extract feature information. Its specific structure is shown in Fig. 6.

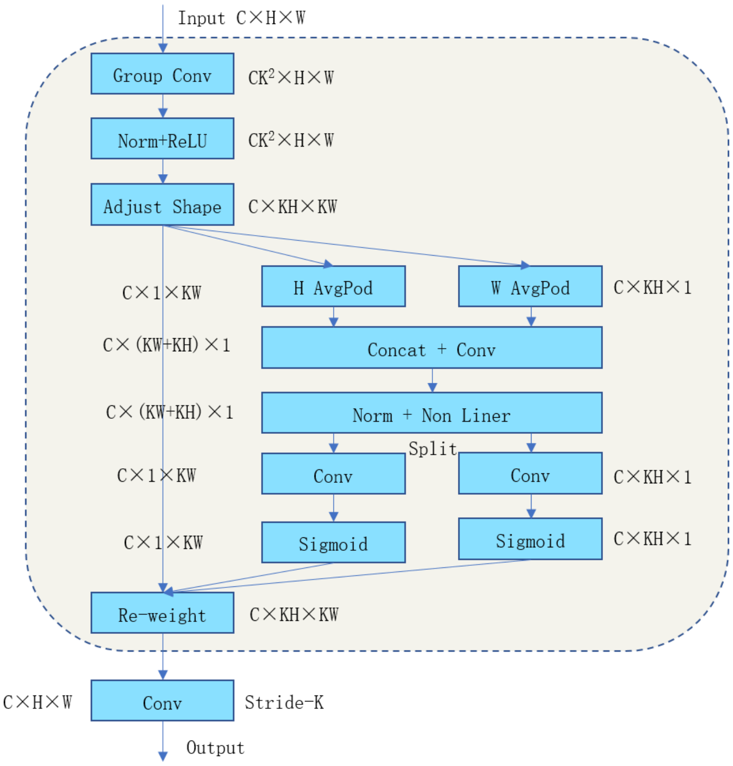

Fig. 6 – RFAConv structure

After the input feature map size of C×H×W enters the RFAConv module, the number of channels is extended to CK2 by a group convolution operation, generating a new feature map size of CK2×H×W. Subsequently, the feature map is further reshaped to C×KH×W after normalization with a nonlinear activation function, ReLU. To extract information along the spatial dimensions, average pooling is applied separately in the height and width directions, resulting in two intermediate feature maps with dimensions C × 1 × KW and C × KH × 1, respectively. These two feature maps are subsequently spliced together and fused by a convolutional layer to generate a new feature map with dimensions C × (KW + KH) × 1. After normalization and nonlinear transformation, this feature map is split into two parts corresponding to the information in the height and width directions, and each of them is passed through a convolution operation as well as a Sigmoid function to generate the final attention weight map. The generated attentional weight map is used to reweight the original input feature map through a final convolutional layer using a step size of K setting to generate the final output feature map.

SENet (Squeeze-and-Excitation Networks) is a convolutional neural network architecture that introduces an attentional mechanism in the channel dimension, aiming to improve the representation ability of the network by capturing the interdependencies between feature channels. It introduces an attention mechanism that significantly improves the image classification performance without significantly increasing the computational complexity. The core idea is to learn a weight for each feature channel to enhance useful features and weaken irrelevant features. Compared with traditional CNNs, SENet optimizes performance by introducing dynamic feature tuning through the SE module, which consists of two key operations: Squeeze and Excitation. In the compression phase, the SE module captures the global context information by compressing the spatial information of each channel into a single value through global average pooling. In the Excitation phase, a fully connected layer and a nonlinear activation function are utilized to learn the inter-channel dependencies and generate weights for each channel. These weights are used to recalibrate the responses of the feature channels, enhancing useful features and suppressing irrelevant features. This mechanism allows SENet to adaptively tune the feature representation and improve the performance of the model. The basic structure of SENet is shown in Fig. 7.

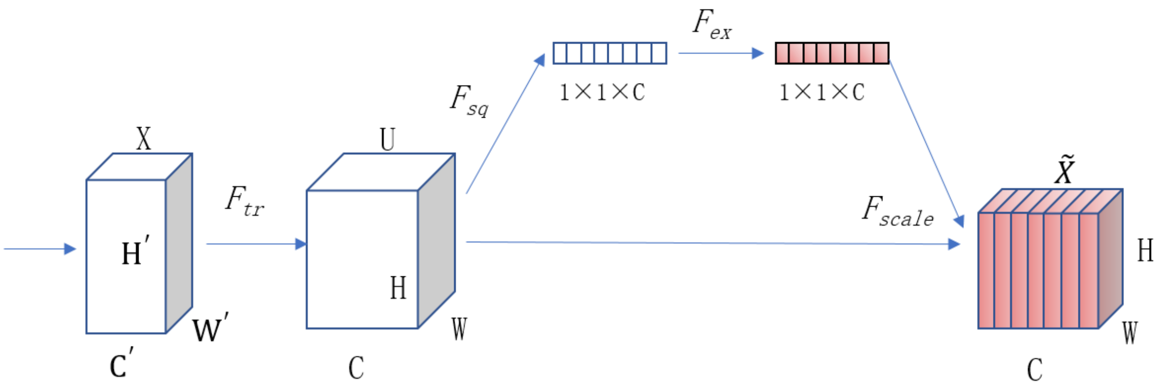

Fig. 7 – SENet module structure

For any transformation F_{tr} that maps an input X to a feature map U, where U ∊ RH×W×C, the corresponding SE block can be constructed for feature recalibration. First, the feature map U is aggregated in the spatial dimension (H × W) by the Squeeze operation to generate a channel descriptor. This descriptor is used to embed the global distribution of the channel-level feature responses, enabling the layers of the network to utilize the global sensory field information. The subsequent Excitation operation employs a self-selecting pass mechanism to generate per-channel modulation weights using the embedding as input. These weights are applied to the feature map U to generate SE block outputs that are passed directly to subsequent network layers.

Squeeze operation: the SE module generates a descriptor for each channel by aggregating the spatial dimensions (H×W) of the input feature map through a global average pooling operation. This operation compresses the global spatial information into channel vectors, capturing the global distribution of the channel feature responses and providing critical global information for subsequent recalibration. In the Squeeze operation, the input feature map is globally average pooled to compress the feature map with dimensions W × H × C into a 1 × 1 × C feature vector.

This process aggregates the spatial information of each 2D channel into a single value with a global receptive field. In this way, each channel is represented as a numerical value, removing the spatial distribution information and thus utilizing the inter-channel correlation more efficiently. Let ZC be the output vector of the Cth channel in the feature map after the Squeeze operation, and Fsq(UC) denotes the Squeeze operation performed on the input feature map UC of the Cth channel, where H and W denote the height and width of the feature map, respectively.

Excitation operation: after the Squeeze step, an Excitation mechanism is introduced, which is essentially a self-gating mechanism consisting of two fully connected layers and a nonlinear activation function. The first fully-connected layer processes the channel descriptors by dimensionality reduction and imposes the ReLU activation function; the second fully-connected layer maps them back to the original channel dimensions. This process effectively captures the nonlinear relationships between channels and generates a set of channel weights.

The SENet attention mechanism is able to fit the correlation between channels more efficiently and capture the dependencies and interactions between channels, thus optimizing the expression of feature weights. In addition, SENet is less parametric and computationally intensive, adding only a simple fully-connected layer and activation function after the convolutional layer, and thus does not incur significant additional overhead. Integrating it into YOLOv8n's backbone network enhances the model's ability to recognize strawberry features at the channel level and strengthens the extraction of key features while weakening the interference of irrelevant content and complex background. With the SENet attention mechanism, YOLOv8n can more fully utilize the inter-channel relationships to enhance feature differentiation and thus improve the accuracy of target detection.

The Backbone network structure contains a complex structure of ten layers, which is designed to extract features efficiently. Introducing the attention mechanism in each layer may significantly increase the computational burden and model complexity, which may be undesirable for resource-limited application scenarios, and the performance enhancement brought by the attention mechanism needs to be weighed against the computational cost in practical applications. In order to balance the accuracy and efficiency of the model, this paper designs three schemes to add the SENet attention module in the first and third layers, the third and fifth layers, and the fifth and seventh layers respectively, and selects the optimal introduction method as the basis for subsequent model improvement after comparison tests. The experimental results are shown in Table 2.

Table 2

Comparison of data of different versions of YOLOv8

Framework	Precision	Recall	mAP50
Unchanged	91.78%	92.14%	96.74%
Add in the first and third layers	91.25%	91.93%	96.82%
Add in the third and fifth layers	91.93%	92.32%	96.39%
Add in fifth and seventh layers	92.33%	92.87%	97.57%

The experimental data show that in the shallow layer structural adjustment, the improvement of the network module for the first and third layers makes the mAP50 rise slightly by 0.12%, but the Precision and Recall show a drop of 0.53% and 0.21%, respectively, which indicates that the shallow parameter adjustment, although it can improve the comprehensive detection performance to a limited extent, leads to negative fluctuation of the basic detection indexes. In the middle-layer network optimization experiments, enhancements to the five-layer structure increased Precision and Recall by 0.15% and 0.18%, respectively. However, mAP decreased by 0.35%, indicating that while this adjustment improved single-sample discriminative ability, it substantially weakened the model’s overall detection performance on multi-scale targets. In contrast, the deep-network optimization scheme involving the fifth and seventh layers showed significant advantages. Precision and Recall improved by 0.55% and 0.73%, respectively, while mAP increased by 0.83%, verifying that optimizing deep feature extraction modules effectively enhances the model’s multidimensional detection capability. Overall, the results demonstrate a clear depth-dependent characteristic of network structural optimization: the closer the adjustment is to the output layer, the more significant the enhancement in comprehensive performance. Based on these findings, architectural improvements were applied to the deep network module, with the SENet attention mechanism added to the fifth and seventh layers.

InceptionNeXt is an innovative convolutional neural network architecture that combines the multi-scale feature extraction idea of Inception with the large kernel deep convolution design feature of ConvNeXt. In recent years, although large kernel convolution can effectively expand the sensory field and improve the performance, its high memory access cost limits the efficiency of its application on high-performance devices. InceptionNeXt proposes a new deep convolutional structure that achieves a good balance of speed and performance by decomposing the large kernel into multiple small kernels and keeping some of the channels unchanged. This structure divides the channel into three parts, which are processed by 3×3, 1×k and k×1 convolutional kernels respectively, and the rest is passed directly through constant mapping, thus reducing the memory access cost while maintaining a large sense field. This design not only inherits the advantages of the classical Inception module, but also improves the efficiency by reducing the computational cost, solves the bottleneck of the traditional CNN in terms of speed and performance, and becomes the preferred model for efficient image categorization tasks in resource-constrained scenarios. Its structure is shown in Fig. 8.

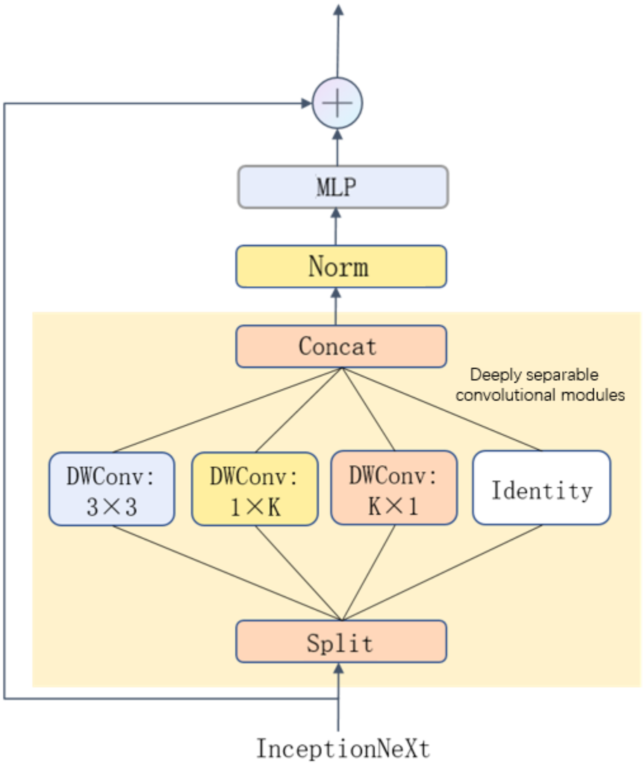

Fig. 8 – InceptionNeXt structure diagram

Distance Intersection over Union - Non Maximum Suppression (DIoU-NMS) is an improved non-maximum suppression algorithm for removing redundant bounding boxes in target detection tasks. In traditional NMS methods, IoU is the only suppression criterion, which may lead to false suppression in the case of occlusion. DIoU-NMS uses DIoU as a criterion for NMS and also considers the distance between the centroids of the two boxes as a judging criterion. Let the Euclidean distance between the centroids of the two bounding boxes be denoted as *d*, and the length of the diagonal of the smallest closure rectangle containing these two bounding boxes as *c*. Although the DIoU computation is more complex compared to the normal IoU, the performance gain it delivers far outweighs the impact of the additional computational cost, ensuring that the overall computational efficiency remains efficient. In terms of implementation details, DIoU-NMS typically consists of four steps: sorting, iterative selection, calculation of DIoU values, and filtering.

First, the prediction frames are sorted in descending order based on their confidence scores. Next, the best candidate frames that are currently unsuppressed are sequentially selected starting from the highest score and added to the final result set. For all the remaining candidate frames, their DIoU values are calculated separately with respect to the selected best frames. If the DIoU between a candidate and the selected best frames exceeds a predefined threshold, the candidate is suppressed; otherwise, it is retained for subsequent comparisons. This method is highly integrable and can be easily embedded with only minor code modifications.

RESULTS

The tests were conducted using a 64-bit Microsoft Windows 11 operating system. The central processor CPU was AMD Ryzen 7 6800H with Radeon Graphics at 3.2 GHz; the running memory (RAM) capacity was 16 GB, and the graphics processor (GPU) is NVIDIA GeForce RTX 3060 with 6 GB of video memory. The software platform was developed using the open-source deep learning framework PyTorch 2.2.1, with Python 3.9.12 as the programming language. The development environment utilized CUDA 11.7 (Compute Unified Device Architecture) and cuDNN 8.7.0 (CUDA Deep Neural Network library) for GPU acceleration. PyCharm and VScode served as the primary development tools.

Parameter settings

The YOLOv8n detection model training parameter settings are shown in Table 3. Because of the changes involved in the model, the model was uniformly constructed from scratch using the yaml file. High-resolution images, such as those with 50 million pixels, are captured to preserve details, while YOLOv8 selects 640×640 input images to improve inference speed and computing efficiency while ensuring detection accuracy.

Table 3

YOLOv8n test parameter settings

Parameter name	Parameter value
Image Input Size	640×640
Training batch	100
Batch size	8
Initial learning rate	0.001
Optimizer	Adam
Attenuation weight	0.0005
Number of processes	4
Learning rate warming	3

Loss Function Comparison

In the training process of deep learning models, the loss function plays a crucial role, which is not only used to quantify the gap between the model prediction results and the actual values, but also provides a direction guide for the optimization and adjustment of model parameters. For YOLOv8n, the design of its loss function plays a decisive role in improving its detection accuracy and computational efficiency. During model training, the loss function helps determine the model parameters that need to be adjusted and the magnitude of their adjustment by evaluating the deviation between the predicted output and the real label. In order to verify the performance of WIOU loss in strawberry target detection, EIoU, GIoU, DIoU, SIoU, and WIOU are used as the loss functions of YOLOv8n for comparative experiments on the strawberry dataset, respectively, and the test results are shown in Table 4.

Table 4

Comparison results of experiments with different loss functions

Mold	Loss function	Precision	Recall	mAP50
YOLOv8n	CIoU	91.78%	92.14%	96.74%
	GIoU	91.72%	92.04%	96.69%
	DIoU	91.93%	92.43%	96.89%
	SIoU	91.91%	92.36%	96.87%
	EIoU	90.94%	91.25%	95.63%
	WIoU	92.36%	93.19%	97.23%

The experimental comparison results show that WIOU using the dynamic non-monotonic focusing mechanism performs best over the other loss functions on this data, with an improvement of 0.58%, 1.05%, and 0.49% in precision, recall, and mean average precision, respectively, compared to the CIoU loss function used by default. WIOU enables YOLOv8n to ensure high-speed processing while also providing more accurate detection results, thus significantly improving overall performance.

Ablation experiment

In this paper, three main optimizations have been made in the structure of the YOLOv8n model. Firstly, the feeling field attention convolution RFAConv is used in the backbone network to replace the standard convolution Conv. Secondly, the SENet attention mechanism is introduced to strengthen the learning ability of the backbone network for strawberry color and shape features. Thirdly, the InceptionNeXt neural network structure is introduced as the feature extraction network, replacing the convolution in Bottleneck with InceptionNeXt convolution in the Cf2 module of the neck network. In order to verify the effectiveness of each improvement, ablation experiments are performed on the strawberry dataset, and the experimental environment as well as the parameters are consistent for each experiment. The experimental results are shown in Table 5.

Table 5

Results of ablation experiments

Model	RFAConv	SENet	Inception NeXt	Precision	Recall	mAP50	Size
YOLO v8n	×	×	×	91.78%	92.14%	96.74%	5.94M
	√	×	×	94.82%	94.15%	97.69%	7.84M
	×	√	×	93.73%	93.47%	97.57%	5.94M
	×	×	√	92.08%	92.27%	96.83%	4.22M
	√	√	×	95.79%	94.83%	97.91%	7.84M
	√	×	√	94.96%	94.33%	97.32%	5.17M
	×	√	√	95.24%	94.71%	97.78%	5.17M
	√	√	√	95.92%	95.45%	98.29%	5.17M

The triple co-optimization model proposed in this study demonstrates significant advantages in both performance and efficiency. By deeply integrating the feature enhancement module, dynamic attention module, and lightweight decoding module, the model achieves 95.92% precision, 95.45% recall, and 98.29% mAP50 value on the baseline dataset. Compared with the baseline model, the three core metrics are improved by 4.14%, 3.31% and 1.55% respectively, while the model size is compressed to 5.17M, which is 12.96% less than the original model. The experimental results show that this architectural innovation effectively solves the contradictory relationship between the number of parameters of the traditional model and the detection accuracy while maintaining the high-precision feature extraction capability, which provides support for the development of the subsequent detection system.

Visualization of detection results

As shown in Figure 9, the confusion matrix provides valuable information about the model's performance across different strawberry-related categories. The dataset contains three main categories: green, half_ripened, and fully_ripened, as well as a background category.

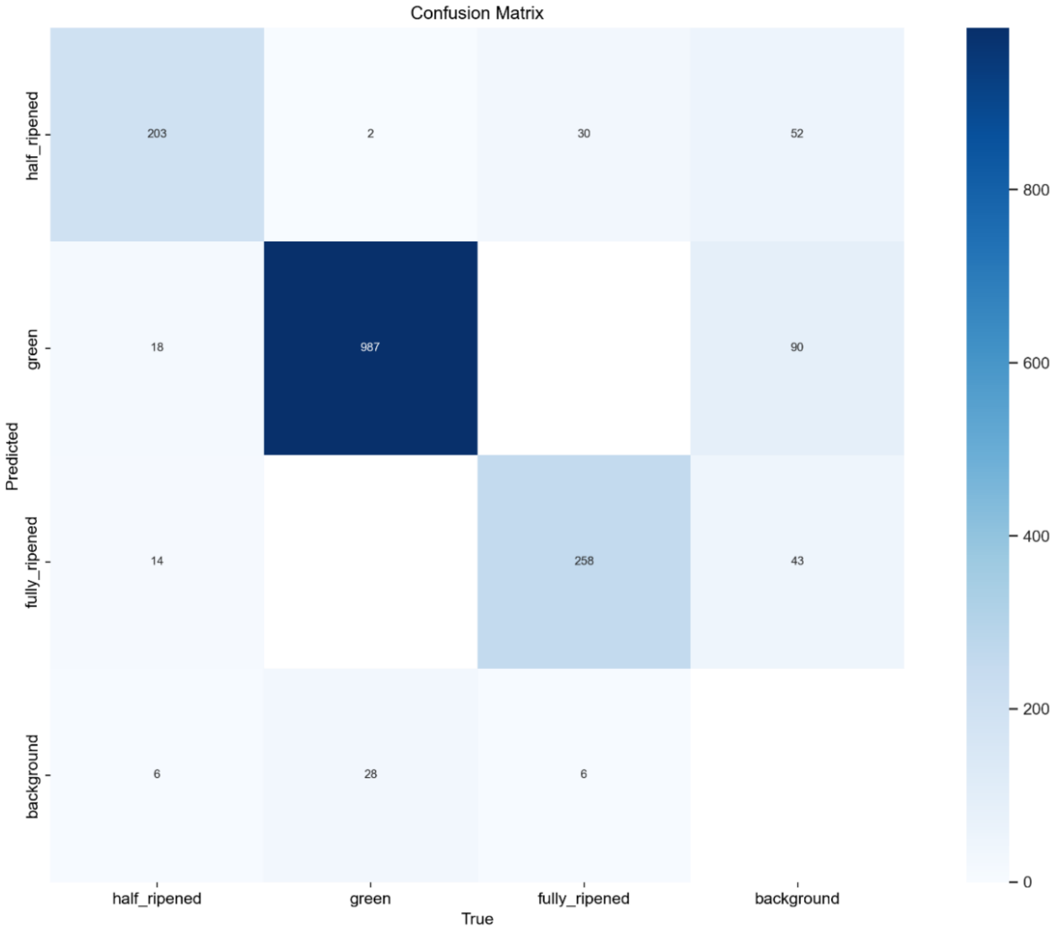

Fig. 9 –Confusion matrix after YOLO8N training

For the half-ripened category, the model demonstrated reasonable recognition capabilities, correctly classifying 203 samples. However, there were also some misclassifications, with six strawberries incorrectly labeled as background. This indicates that although the model can effectively recognize the characteristics of strawberries, there is still some confusion, possibly due to overlapping characteristics or visually similar elements between the strawberries and the background.

The green category achieved the highest recognition accuracy, with a total of 987 samples correctly identified. This high true positive rate indicates that the model effectively learned how to distinguish flowers. However, there are still a small number of misclassifications: 2 were misclassified as the half_ripened category, and 28 were classified as background. The proportion of strawberries misclassified as background is relatively high, indicating that in some cases, the model may have difficulty distinguishing between strawberries and non-strawberry elements, possibly influenced by environmental noise or scene complexity.

From the fully_ripened category, the model's recognition performance was moderate, correctly identifying 256 samples. There were a few misclassifications: 30 samples were misclassified as half_ripened, none were misclassified as green, and 6 were labeled as background. These errors indicate that while the model can identify strawberry ripeness to a certain extent, there is still room for improvement when dealing with ambiguous or unclear cases.

The strawberry dataset is validated using the optimized training model in this paper and the detection results are shown in Fig. 10. It can be seen that it performs well in detecting strawberries of different maturity levels.

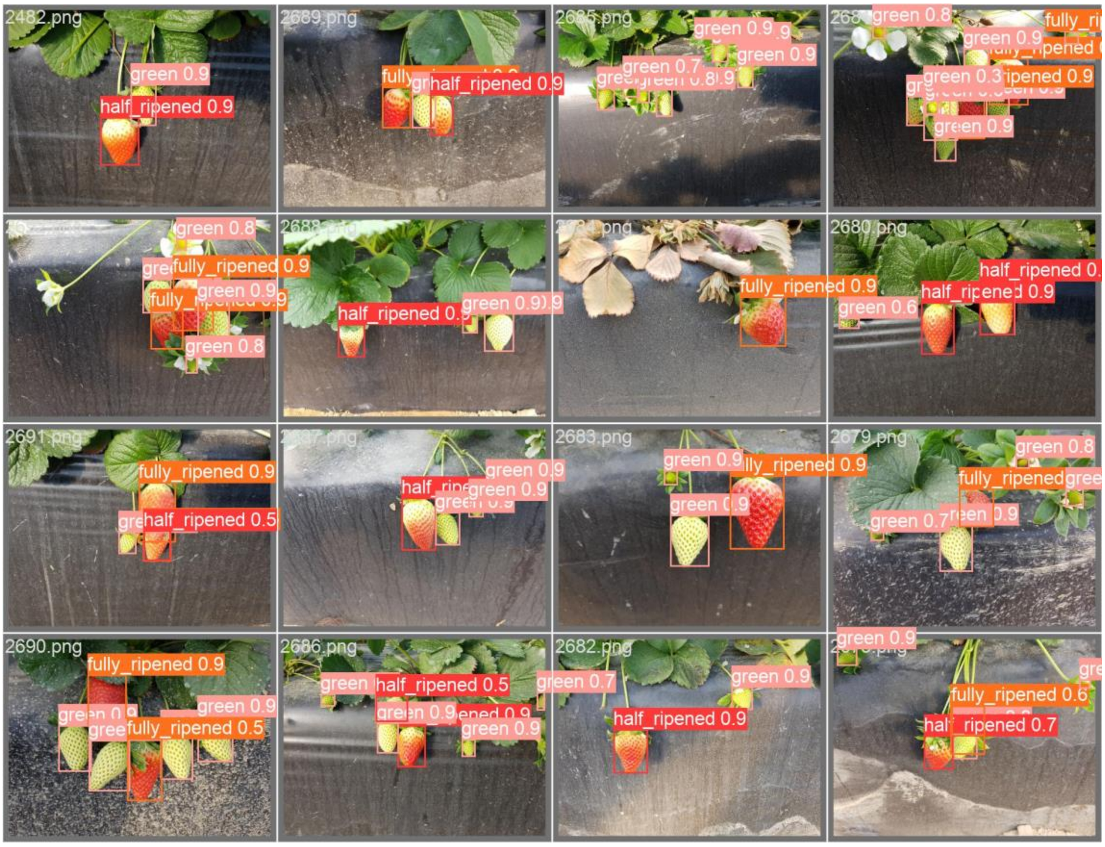

Fig. 10 –Visualization of detection results

CONCLUSIONS

The study introduces RFAConv dynamic sensory field convolution, SENet channel attention mechanism and InceptionNeXt lightweight structure. It combines these with Wise-IoU loss function and DIoU-NMS post-processing algorithm. The result is a synergistic enhancement of detection accuracy and computational efficiency. The experimental findings demonstrate that the enhanced model exhibits a precision rate of 95.92%, a recall rate of 95.45%, and an mAP50 of 98.29% on the strawberry dataset. These metrics represent enhancements of 4.14%, 3.31%, and 1.55%, respectively, when compared to the baseline model. Additionally, the model's parameter count is reduced to 5.17M, marking a 12.96% decrease, which is of considerable importance in fostering the advancement of precision agriculture.

ACKNOWLEDGEMENT

This work was supported by the National Natural Science Foundation of China, grant no. 62366013

REFERENCES

[1] Elsayed M, Reda M, Mashaly A S. (2024) LERFNet: An enlarged effective receptive field backbone network for enhancing visual drone detection. *The Visual Computer,* 4: 1-14.

[2] Hu J, Shen L, Sun G. (2018). Squeeze-and-excitation networks. *IEEE conference on computer vision and pattern recognition*, 8: 7132-7141.

[3] Howard A, Sandler M, Chu G. (2019). Searching for mobilenetv3. *IEEE/CVF international conference on computer vision*, 4: 1314-1324.

[4] Hosna A, Merry E, Gyalmo J. (2022). Transfer learning: a friendly introduction. *Journal of Big Data*, 9(1): 102.

[5] Li W, Qi F, Tang M. (2020). Bidirectional LSTM with self-attention mechanism and multi-channel features for sentiment classification. *Neurocomputing*, 387: 63-77.

[6] Lu X, Ji J, Xing Z. (2021). Attention and feature fusion SSD for remote sensing object detection. *IEEE Transactions on Instrumentation and Measurement*, 70: 1-9.

[7] Peng H, Li B, Xiong W. (2014). RGBD salient object detection: A benchmark and algorithms. *Computer Vision–ECCV 2014: 13th European Conference*, 9: 92-109.

[8] Sharma A K, Nguyen H H C, Bui T X. (2022). An approach to ripening of pineapple fruit with model Yolo v5. *IEEE 7th international conference for convergence in technology (I2CT)*, 2: 1-5.

[9] Vaswani A, Shazeer N, Parmar N. (2017). Attention is All You Need. *Conference of Neural Information Processing Systems*,3:6000-6010.

[10] Wan L, Zhang J, Dong X. (2021). Unmanned aerial vehicle-based field phenotyping of crop biomass using growth traits retrieved from PROSAIL model. *Computers and Electronics in Agriculture*, 187: 106304.

[11] Wang J, Zhang Z, Dai B. (2024). Cow-YOLO: Automatic cow mounting detection based on non-local CSPDarknet53 and multiscale neck. *International Journal of Agricultural and Biological Engineering*, 17(3): 193-202.

[12] Zhu J, Hu T, Zheng L. (2024). YOLOv8-C2f-Faster-EMA: an improved underwater trash detection model based on YOLOv8. *Sensors*, 24(8): 2483.