Journal of Agricultural Science and Technology A 14 (2024) 46-66
doi: 10.17265/2161-6256/2024.02.002

An Efficient and Lightweight YOLOv8s Strawberry Maturity Detection Model

Yiming Cheng[1], Guohao Feng[2] and Chunchang Zhang[3]

1. Merchant Marine College, Shanghai Maritime University, Shanghai 200135, China

2. College of Information and Technology, Shanghai Ocean University, Shanghai 201306, China

3. Merchant Marine College, Shanghai Maritime University, Shanghai 200135, China

Abstract: The manual picking of strawberries is inefficient and costly, limiting scalability and economic benefits. Mechanizing this process reduces labor demands, improves working conditions, and modernizes the strawberry industry. Target detection technology, crucial for mechanized picking, must accurately determine strawberry maturity. This study presents an enhanced YOLOv8s model addressing current machine learning issues like accuracy, parameters, and complexity. The improved model replaces the Bottleneck structure in C2f with the FasterNet network, integrates an efficient multi-scale attention mechanism, and uses the Ghost module in the backbone to reduce computational load while maintaining performance. It also introduces Wise-IoU for bounding box regression loss, improving recognition accuracy. The YOLOv8s-FEGW model achieves a 93.8% mAP in detecting strawberry ripeness, with significant reductions in parameters (36.8%), complexity (34.6%), and model size (37.7%), alongside a 12.7% Frames Per Second (FPS) boost. These enhancements result in excellent detection capabilities, supporting agricultural automation and intelligence.

Key words: Automation equipment, artificial intelligence, efficient and lightweight, YOLOv8s.

1. Introduction

The swift advancement in computer vision technology has established object detection as one of its most promising application areas. This technology is now extensively applied in numerous aspects of everyday life, including surveillance systems, autonomous vehicles, and drone-based scene analysis [1]. Strawberries, being nutrient-dense fruits packed with beneficial phytochemicals, offer vital nourishment to humans. The utilization of computer vision for strawberry identification demonstrates significant potential for real-world application [2] and offers a theoretical basis for robotic strawberry harvesting [3]. Object detection is a technology that uses algorithm models to identify and track targets in real time. Traditional object detection methods involve using several sliding windows to analyze image datasets, extract features, and train classifiers to detect target images. Chen et al. [4] introduced a technique for identifying corn diseases using computer vision technology. The integration of sophisticated image processing and machine learning methods offers a reliable approach to enhancing the precision in identifying corn diseases. In challenging settings for corn leaf disease image identification, Bachhal et al. [5] suggested an automated system utilizing the PRF-SVM model. The system employs PSPNet, ResNet50, and a Fuzzy Support Vector Machine (Fuzzy SVM) to boost the model's ability to discern fine visual details and manage the ambiguity and variability present in image data. The suggested approach shows strong capability in identifying and categorizing five types of corn diseases—common rust, southern rust, gray leaf spot, medium leaf blight, and turmeric leaf blight—alongside healthy leaves, utilizing the Plant Village database. This method attains an average accuracy of 96.67% and an mAP score of 0.81, affirming its effectiveness in

Corresponding author: Chunchang Zhang, Doctorate Degree, associate professor, research fields: Naval Architecture and Marine Engineering.

detecting and categorizing corn leaf diseases. Additionally, Mustafa et al. [6] presented a new method for creating a smart fruit sorting system that leverages digital image processing and artificial neural network technologies. The results showed significant improvements over prior studies. Li et al. [7] proposed a machine vision-based automatic grape size classification method designed to achieve non-destructive, automatic size grading of grapes. The process begins with capturing grape images using a machine vision device, followed by image enhancement to improve quality. Edge detection is then employed to perform double contour segmentation, ensuring accurate extraction of the grape outline. Experiments showed that the method achieves a classification accuracy of nearly 90%, confirming its effectiveness and feasibility for automatic grape grading. Qiaohua et al. [8] developed an innovative ellipse fitting method based on iterative minimum median squares. The method involves four key steps: first, removing grape stems from RGB and NIR images captured by a 2-CCD camera; second, extracting grape edges using edge detection, image binarization, and morphological processing; third, integrating image segmentation with ellipse fitting to determine the minor axis length for size measurement; and fourth, grading grapes according to the 15% downgrade principle, ensuring re-evaluation of more than 15% of the cases. When tested on 38 cases of red ball grapes, the method achieved an accuracy of 92.1%, correctly grading 35 cases. The results confirm the method's ability to meet the high standards required for accurate and fast automatic online grape detection, providing an efficient and reliable solution for the automatic grading of agricultural products. Reis et al. [9] addressed the low adoption of automated harvesting technology in precision agriculture and precision grape cultivation in the Douro River demarcation area of Portugal. Their system focuses on detecting and locating grape clusters in natural color images, particularly addressing the challenge of distinguishing between white and red grapes, which often blend in with leaves. The developed system not only distinguishes between white and red grapes but also calculates their locations. Experiments showed an accuracy of 97% in red grape classification and 91% in white grape classification, demonstrating the system's potential and effectiveness in practical applications. This innovation supports grape harvesting tasks in winemaking, improving efficiency and accuracy, reducing labor intensity, and promoting the application and development of precision agriculture technologies in the region. Arefi et al. [10] proposed an innovative segmentation algorithm to identify and locate ripe tomatoes, a key challenge in the field of automated tomato picking. The algorithm aims to guide the robot arm to accurately pick tomatoes in combination with a machine vision system to address the time-consuming, tedious, and costly problems of manual picking. The research team collected 110 color images of tomatoes under greenhouse lighting conditions and developed a recognition algorithm that adapts to these lighting changes. The proposed algorithm achieves efficient and accurate tomato recognition in two stages: first, background removal is performed in the RGB color space, and the RGB, HSI, and YIQ color spaces are fused to extract the features of ripe tomatoes; second, the morphological features of the image are used to locate ripe tomatoes. After testing, the algorithm accomplished a high accuracy of 96.36% in the recognition and location of ripe tomatoes.

In the context of contemporary large-scale data processing tasks, vision algorithms that rely on conventional feature extraction methodologies have demonstrated their inherent limitations. This challenge is particularly pronounced in the domain of strawberry ripeness detection. The variability in the growth orientation and positioning of strawberries, mutual occlusion among them, interaction with surrounding foliage, and the instability of lighting conditions present significant challenges to the accurate detection capabilities of traditional algorithms. The assessment

of strawberry ripeness encompasses multifaceted changes in color, shape, size, and texture. Traditional methods are often unable to capture and analyze these complex features at the same time, thus affecting the accuracy of detection. In addition, the appearance characteristics of strawberries at different growth stages and between different varieties are significantly different, which results in an extremely complex data distribution of strawberry maturity, making it difficult for traditional feature extraction and classification methods to cope with and achieve accurate identification. The introduction of deep learning neural networks offers a transformative solution to these challenges. Contrary to traditional algorithms, deep learning models learn the intrinsic patterns of data through training, rather than relying on intricate programming logic. This approach significantly reduces the dependence on extensive expert analysis and meticulous parameter tuning. Deep learning technology has demonstrated remarkable flexibility and potential in maturity detection and a multitude of other fields, enabling the attainment of superior detection outcomes even in the absence of extensive horticultural expertise. Consequently, this study opted for a deep learning-based approach to maturity detection, leveraging its powerful data processing capabilities and feature learning mechanisms to surmount the limitations of traditional algorithms and achieve efficient and accurate detection of strawberry ripeness. By leveraging deep learning techniques, we can achieve a more profound insight and analysis of the intricate features related to strawberry ripening, thus enabling smarter approaches in agriculture. The swift progress in deep learning, especially with the advent of Convolutional Neural Networks (CNNs), has transformed object detection methodologies, marking a significant advancement in the field. Since LeCun and others laid the foundation in 2015, to the breakthrough contributions of researchers such as Girshick, Ren, Redmon, and Liu in the following years, deep learning has demonstrated unprecedented capabilities in image recognition and processing. This technological advance is particularly widely used in agriculture. Deep learning not only optimizes critical aspects such as crop disease detection, maturity assessment, and yield prediction but also provides substantial technical support for the implementation of precision agriculture. Through the powerful feature extraction and pattern recognition capabilities of deep learning models, agricultural inspections have become more automated and intelligent, significantly enhancing the efficiency and sustainability of agricultural production. As technology continues to advance, we can foresee that deep learning will continue to exert its huge potential and value in agricultural inspection and more fields. Goyal et al. [11] built a fruit recognition and quality assessment model based on YOLOv5. Experimental results demonstrate that the average accuracy (mAP) of the model in the initial stage reached 92.80%. In the subsequent stage, the quality detection models for apples and bananas attained mAP values of 99.60% and 93.1% respectively, and the quality detection models for oranges and tomatoes also achieved mAP values of 96.70% and 95% respectively. Ma et al. [12] implemented advanced training methods and applied transfer learning to integrate features of strawberry diseases into the training dataset. These features were then classified and recognized to accomplish the goal of disease identification. Javanmardi et al. [13] proposed a novel approach using a deep Convolutional Neural Network (CNN) for extracting general features. Following this, various classifiers such as Artificial Neural Networks (ANN), Cubic Support Vector Machines (SVM), Quadratic SVM, Weighted k-Nearest Neighbor (kNN), Boosted Tree, Bagged Tree, and Linear Discriminant Analysis (LDA) were employed to analyze the extracted features. When trained with features derived from the CNN, the model demonstrated superior accuracy in classifying corn seed types compared to a model that relied solely on fundamental features. Among the classifiers, the CNN-ANN combination was the most effective, categorizing

2,250 test samples in 26.8 s, achieving a classification accuracy of 98.1%, precision of 98.2%, recall of 98.1%, and an F1 score of 98.1%. Ashtiani et al. [14] and their colleagues detailed the creation and evaluation of a computer vision-based application using Convolutional Neural Networks (CNN) for classifying mulberry ripening stages. Transfer learning was utilized to optimize the CNN model, reducing training expenses and enhancing classification accuracy. The tested CNN models included DenseNet, Inception-v3, ResNet-18, ResNet-50, and AlexNet. Transfer learning was applied to refine these models and improve their classification performance. The AlexNet and ResNet-18 networks achieved the highest accuracy rates of 98.32% and 98.65%, respectively, in the classification of white and black mulberry maturity. Jeong et al. [15] created an advanced automated system for precise measurement of strawberry size and AI-predicted weight by integrating computer vision with LiDAR sensor data. The system utilizes the deep learning model HRNet for keypoint detection, which minimizes human error and increases measurement accuracy. Experimental data show that the relative errors of strawberry width and height are as low as 3.71% and 5.42% respectively, and the standard deviations are even more negligible, 0.19% and 0.24% respectively. Through regression analysis, combined with width and height data, the system was able to accurately predict strawberry weight with a relative error of 10.3%. Chen et al. [16] proposed an optimized lightweight variant of the YOLOv5 model specifically designed for real-time detection of strawberry diseases. By incorporating the Ghost Convolution (GhostConv) module and the Convolutional Block Attention Module (CBAM), the model enhances its capacity to extract and identify strawberry disease features, while simultaneously reducing the number of parameters and floating-point operations (FLOPs). Furthermore, the Content-Aware Feature Rearrangement (CARAFE) lightweight upsampling operator is utilized to replace the traditional upsampling module, further improving the accuracy of feature extraction. Testing on the strawberry disease data set showed that the model achieved 94.7% average precision (mAP) at 0.5 with only 3.9 million parameters and 360 million FLOPs. Ridho [17] used deep neural network and computer vision technology to realize the feasibility of real-time detection of strawberry quality through single-board computers (SBC), aiming to help strawberry growers automate the harvest process and ensure that only high-quality strawberries are picked. In the study, the robot software was constructed based on the Robot Operating System (ROS) framework, integrated with a target detection algorithm, and a monocular camera utilized as an input device. Through the training of a deep learning model, the robot was able to successfully differentiate between high-quality and defective strawberries with an accuracy of 90%, while maintaining a high frame rate, thereby fulfilling the requirements for real-time processing. Kim et al. [18] proposed a model to simultaneously learn the ripeness of strawberries and the coordinates of the stem through a dual-path model of semantic segmentation. The dual-path model can be viewed as a model that performs two tasks simultaneously. The model's accuracy in detecting strawberry maturity was 90.33%, and its accuracy in identifying stem coordinates was 71.15%. Wang et al. [19] improves appearance quality and identification efficiency. This method initially employs the ResNeXt network to deeply extract the features of strawberry images, which are subsequently passed to a Support Vector Machine (SVM) for appearance quality recognition. Experimental results demonstrate that the accuracy of the ResNeXt-SVM method in identifying strawberry appearance quality reaches 98.56%, representing an improvement of 1.42% over the original ResNeXt model. Moreover, the total computation time for this method is only 21.27 s, indicating a faster processing speed compared to alternative models.

In summary, in the agricultural field, convolutional neural networks have broad application prospects for

target detection, especially in the detection of fruit maturity. Traditional manual judgment of maturity has many shortcomings, such as strong subjectivity, inconsistent standards, and high labor intensity. Using computer vision technology to detect fruit maturity can improve detection accuracy, avoid waste caused by improper picking time, and improve product quality and market competitiveness.

Aiming at the challenges of existing fruit maturity detection research, such as low maturity detection accuracy, lack of lightweight and efficient features of the model, and poor effect in complicated environments, this study proposes an improved YOLOv8s-FEGW model based on the YOLOv8s model. This model combines the advantages of FasterNet-EMA, Ghost, and WIoU modules, and achieves accurate detection of strawberry maturity through data enhancement, feature fusion, and improvement of the residual network.

The main contributions of this paper include:

(1) A comprehensive dataset covering a variety of strawberry varieties, angles, and densities is established. Compared with existing datasets, the data types are richer and more versatile.

(2) An efficient and lightweight model YOLOv8s-FEGW is proposed. By adopting the FasterNet network structure and integrating the EMA attention mechanism to replace C2f, integrating the Ghost module, replacing the loss function, and adopting WIoU, efficient and lightweight detection of strawberry maturity is achieved.

(3) The data from ablation and comparative experiments indicate that, relative to existing algorithms, YOLOv8s-FEGW exhibits significant advantages in model accuracy, complexity, and computational efficiency. Specifically, across key performance indicators including mean Average Precision (mAP), Frames Per Second (FPS), number of Parameters (Params), and GigaFLOPs (GFlops), enhancements of 2.7% and 12.7%, along with reductions of 36.8% and 34.6%, have been noted, respectively. This enables the accurate detection of strawberry maturity in diverse and complex growth environments, with good compatibility on mobile platforms.

Through these innovations and improvements, this study provides a new strategy for strawberry maturity detection, which helps to improve the operating efficiency of fruit-picking robots in complex agricultural environments, ensure the quality of picked fruits, and advance the progress of agricultural automation and intelligence.

2. Datasets

2.1 Acquisition of Datasets

The dataset used in this research includes three distinct strawberry cultivars: Suizhu, Jingzangxiang, and Falandi. These samples were collected from a specialized strawberry farm located in Nanning, Guangxi Zhuang Autonomous Region, between February 1 and March 1, 2024. During the data acquisition phase, a German Basler industrial camera was utilized for image capture, maintaining a camera-to-strawberry distance of 0.5 to 1.4 m to ensure optimal image quality. Data collection occurred from 7:00 AM to 7:00 PM, encompassing a variety of meteorological conditions, including overcast, partly cloudy, and clear skies, as shown in Fig. 1, to ensure the diversity of the dataset. A total of 2,000 images were taken during this period, with each image resized to 640 × 640 pixels to form the foundational image dataset for detecting strawberry maturity at different growth stages.

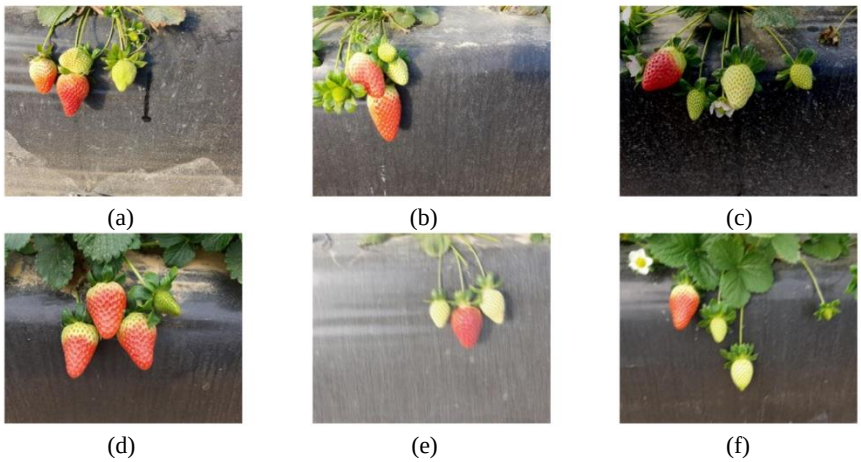


Fig. 1 Representative sample datasets in various conditions: (a) sunny, (b) cloudy, (c) overcast, (d) enhanced, (e) rainy, (f) blurred.

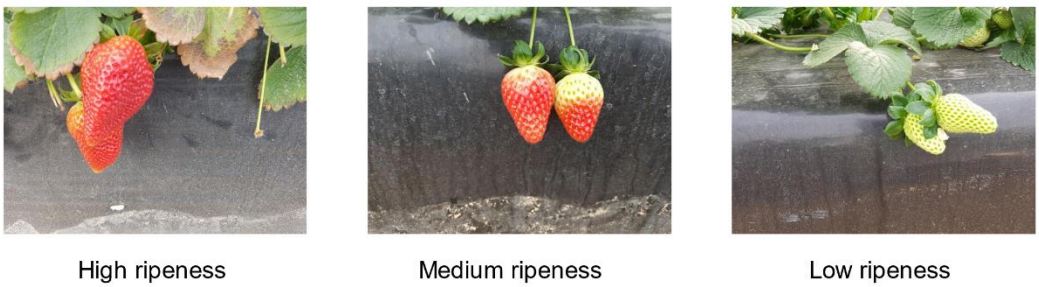


Fig. 2 Strawberries at different degrees of maturity.

2.2 Annotation of Datasets

Conventional strawberry maturity classification can be divided into red ripening stage, color change stage, white ripening stage, and green ripening stage according to the extent of coloration on the fruit surface, and the coloring area accounts for about 100%, 75%, 50%, 25% and less of the fruit surface respectively. Since it is difficult to accurately judge the degree of coloring of strawberry peel in the natural environment, this study divides strawberry maturity into three levels based on the coloring area of the peel and market demand: high ripeness, the strawberry peel is basically all red; medium ripeness, the strawberry peel is red and green; low ripeness, the peel is mainly green [20].The three types of strawberries are marked separately, as shown in Fig. 2.

2.3 Augmentation of Datasets

Data augmentation possesses considerable benefits and is aimed at improving the learning efficiency and performance of machine learning models by creating synthetic data. This method reduces the need for raw data collection and annotation while improving the adaptability and robustness of the model to input data changes (such as noise, size differences, and lighting conditions) [21]. In this study, to improve the model robustness and generalization ability to individual strawberry size changes and lighting conditions, we used four image processing techniques for data augmentation, as shown in Fig. 3, namely AverageBlur, GaussianBlur, WithBrightness Channels, and image occlusion. These methods were flexibly combined and applied to expand the dataset, ultimately generating a training set comprising 5,000 images, a validation set of 625 images, and a test set of 625 images. Table 1 illustrates the changes in the number of category labels before and after augmentation.

(a) (b) (c) (d)

Fig. 3 Effects of different data augmentations. (a) AverageBlur, (b) GaussianBlur, (c) WithBrightnessChannels, (d) Image occlusion.

Table 1 Number of category labels before and after data augmentation.

Catagory	Original	Data enhancement
High ripeness	1,764	7,056
Medium ripeness	1,578	6,312
Low ripeness	5,427	16,281

3. Methodologies

3.1 YOLOv8 Overview

You Only Look Once (YOLO) is a One-Stage object detection algorithm proposed by Joseph Redmon et al. in 2015. After several years of development, it has now been updated to the YOLOv10 version [22-25]. YOLOv8 is a real-time object detection model released by Ultralytics in January 2023. It is currently open sourced on GitHub and can be used for scene tasks such as image classification, object detection, and instance segmentation. Compared to previous versions of Yolo, YOLOv8s has further improved the performance of the model by introducing various powerful features and improvement points. The network framework is shown in Fig. 4-5. The specific improvement points are as follows:

(1) In the backbone structure, the C3 module (shown in Fig. 6a) in the model was improved and the C2f module (shown in Fig. 6b) was proposed, which further achieved lightweight compared to the C3 module,

(2) In the head structure, the current mainstream Decoupled Head structure has been replaced, separating the classification and detection heads, and the design concept has been changed from Anchor Based to Anchor Free.

(3) In the design of the loss function, VFL Loss is used as the classification loss (BCE Loss is used in actual training); Use DFL Loss+IOU Loss as the regression loss.

(4) In the label classification task, the traditional IoU allocation or one-sided proportional allocation method has been abandoned, and the Task Assigned positive and negative sample allocation strategy has been adopted [26].

3.2 YOLOv8s Model Improvement Strategy

As a lightweight network, YOLOv8s has shown impressive performance, but there is still room for improvement. The effectiveness of attention mechanisms has been well established in various computer vision tasks. To enhance the model's ability to accurately recognize strawberries in complex environments and identify various irregularly distributed fruit, this study refines the C2f module and introduces the FasterBlock component to augment the Bottleneck structure, leading to the construction of the C2f-Faster model. By further integrating and refining the EMA attention mechanism, the C2f-Faster-EMA model is developed, which significantly boosts the model's accuracy in detecting strawberries at different stages of maturity. To create a lightweight model and enhance its performance, this study introduces the C2f-Ghost and Conv-Ghost models by incorporating the Ghost module into the backbone. The architecture of the C2f-Faster model is depicted in Fig. 6c, while the C2f-EMA (CFE) model architecture is shown in Fig. 6d. The choice of loss

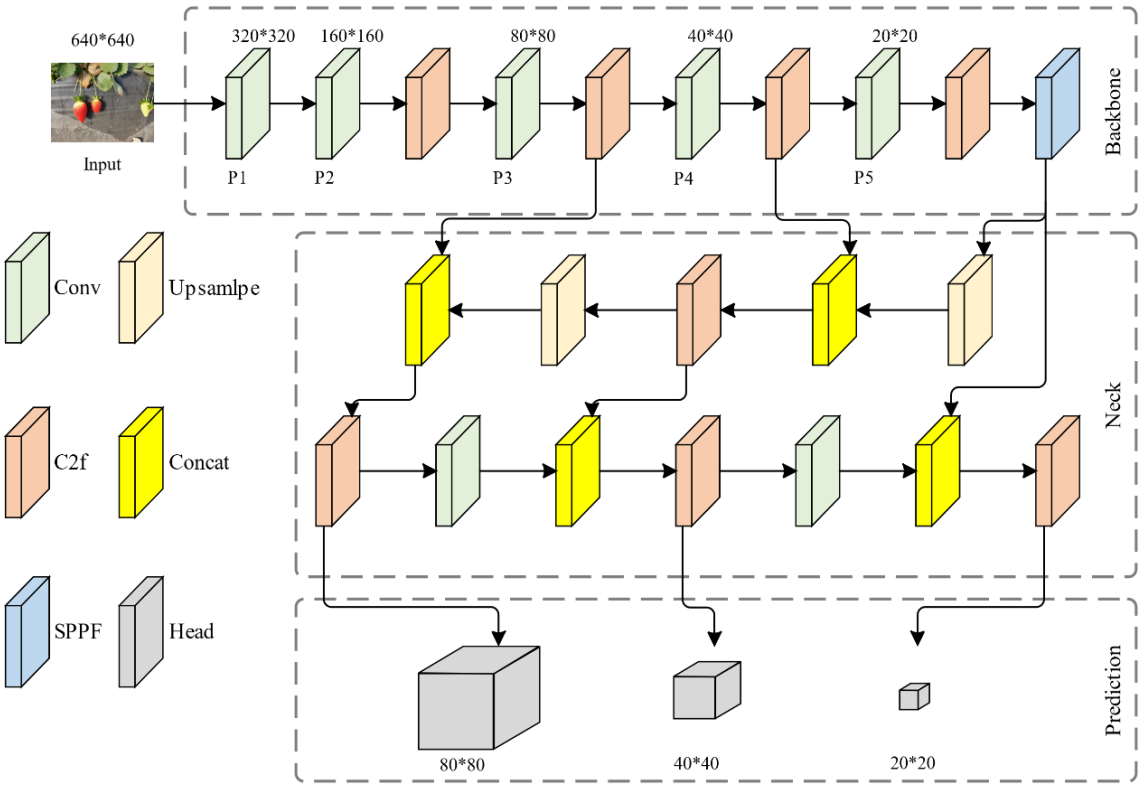


Fig. 4 YOLOv8 network framework.

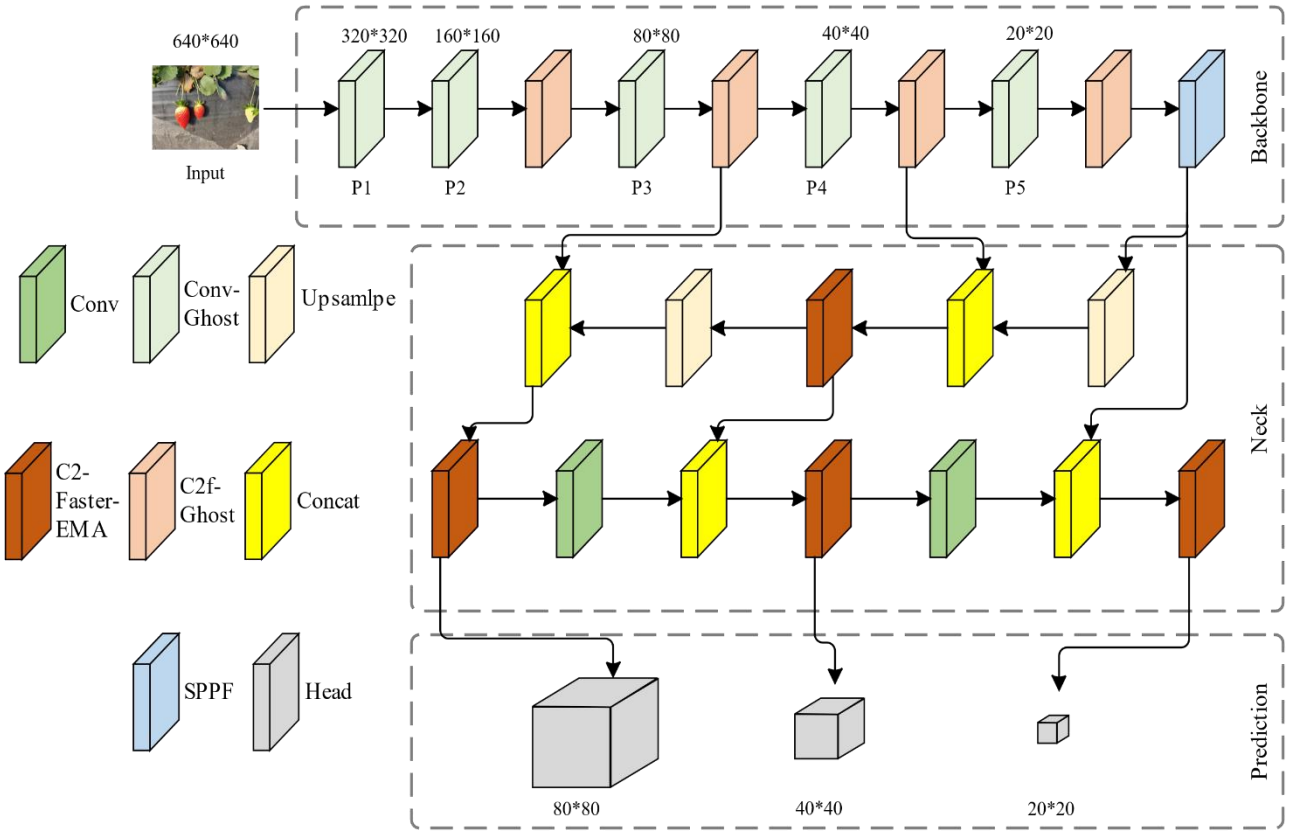
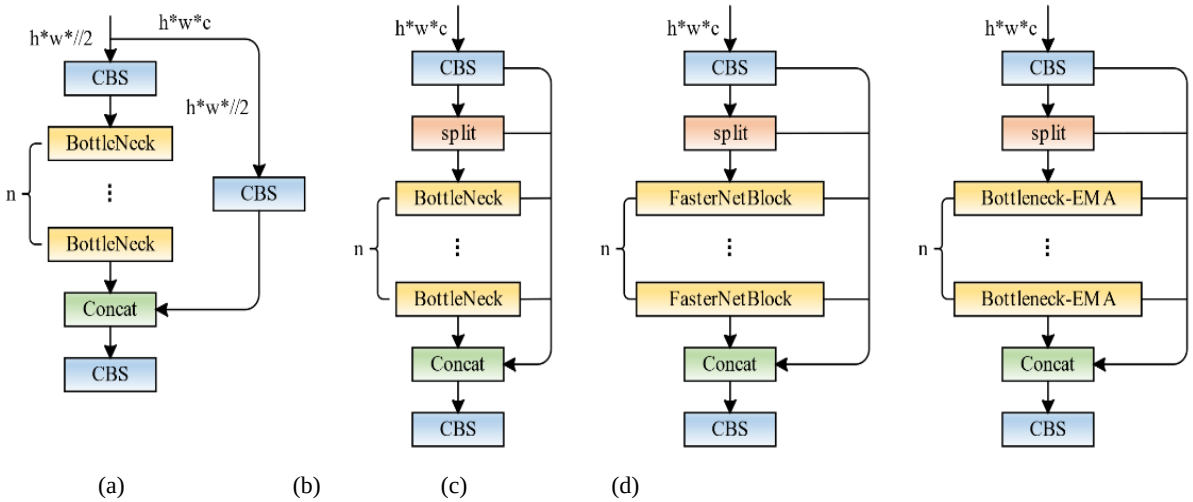


Fig. 5 The improved YOLOv8s-FEGW network framework.



(a) (b) (c) (d)

Fig. 6 Partial module architectural diagram: (a) C3, (b) C2f, (c) C2f-Faster, (d) C2f-EMA.

function is crucial for improving detection accuracy. To tackle the challenges of detecting small targets and handling occlusions in the strawberry detection process, this study adopts the WIoU loss function instead of the CIoU loss function used in the original YOLOv8 network. This change effectively enhances the model's accuracy in identifying strawberries in complex growth conditions. The detailed architecture of the YOLOv8s-FEGW network designed in this study is presented in Fig. 5. In-depth theoretical analysis and specific implementation details will be covered in Sections 3.3 to 3.5.

3.3 C2f-Fasten-EMA Module

3.3.1 FastenNet Block

To enhance the computational efficiency of deep learning algorithms, we innovatively proposed the C2f-Faster module by integrating FastenNet Block into C2f, which can significantly cut down the number of model parameters and floating-point operations, thereby reducing the complexity of the model. C2f-Faster is derived by adopting FasterNet Block [27] instead of Bottleneck in C2f and has substantially improved the efficiency of feature extraction. To optimize costs by utilizing feature map redundancies, as depicted in Fig. 7a, this study proposes a novel partial convolution (PConv) to extract spatial features while reducing redundant calculations and memory access. For sequential memory access, PConv evaluates only a segment of the input channel c_p representative of the whole feature map, applies standard convolution to extract spatial features, and retains the rest of the channels as is assuming that the input and output feature maps have the same number of channels, denoted as c, the FLOPs of PConv is $h\times w\times k^2\times c_p^2$. When the partial ratio is $r=c_p/c=1/4$, the FLOPs of PConv are only $1/16$ of that of ordinary convolution. Simultaneously, PConv's memory utilization is computed as $h\times w\times 2c_p+k^2\times c_p^2\approx h\times w\times 2c_p$, which is 1/4 of the original. Fig. 7b illustrates the configuration of the FasterNet Block. Constructed around PConv as its core component, the FasterNet Block boosts processing speeds while maintaining visual task accuracy across different devices. This block includes a PConv and two 1×1 Conv layers arranged as inverse residual blocks, incorporating shortcut connections to recycle key functionalities. Normalization and activation layers are strategically placed post the central 1×1 Conv layer to maintain feature variety and improve response times.

3.3.2 Efficient Multi-scale Attention (EMA)

In recent years, within computer vision, the use of channel and spatial attention mechanisms has shown efficacy in producing more distinct and detailed feature representations. Addressing the issue where the reduction

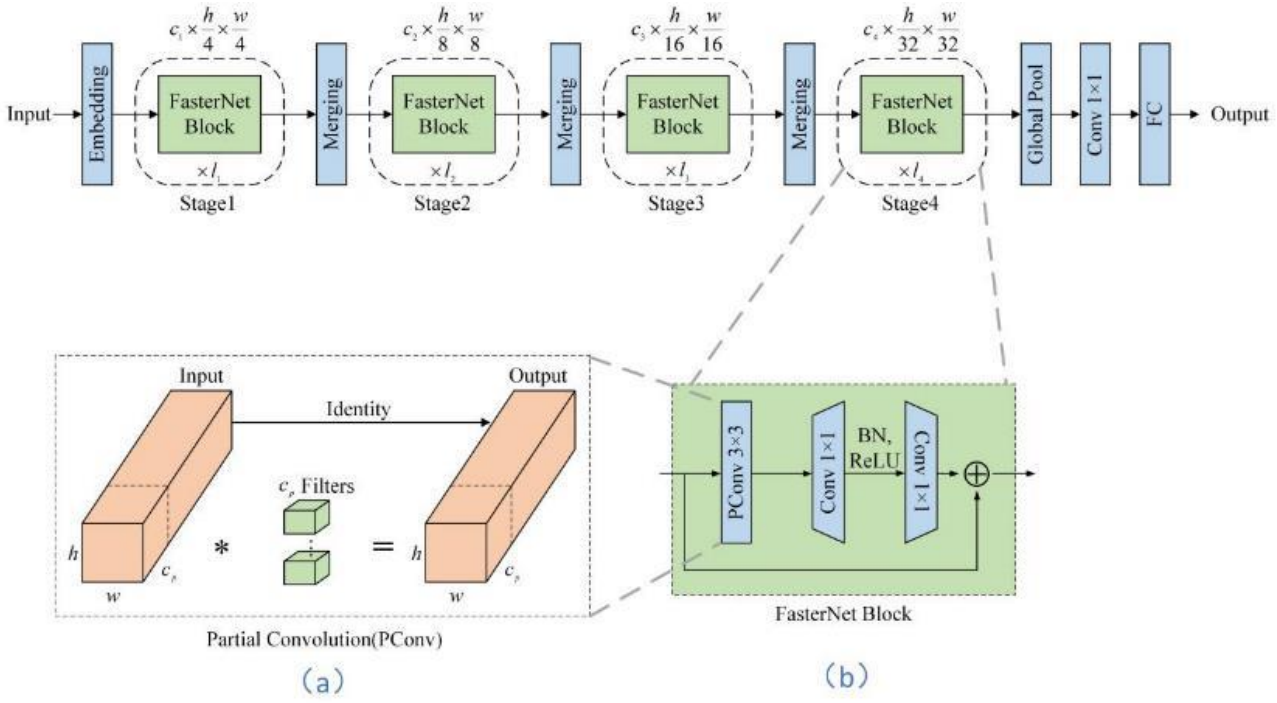


Fig. 7 FastenNet Block schematic.

of channel dimensions in traditional attention mechanisms could result in data loss, Ouyang and colleagues [28, 29] have introduced an innovative efficient multi-scale attention (EMA) module, building upon the CA module. This EMA module acquires precise channel descriptions during convolution by grouping features, utilizing parallel sub-networks, and employing cross-space learning tactics, thus maintaining channel dimensions intact and enhancing pixel-level attention for the creation of superior feature maps.

Specifically, any input feature map $F \in R^{c\times h\times w}$ is divided into $g(g \square c)$ sub-feature sets $F=\left[F_0,F_1,\ldots,F_{g-1}\right]$, each is used to capture different semantic information. The EMA module derives attention weight descriptors for each grouped feature map via three parallel routes; two paths involve 1×1 convolutions, while the third employs a 3×3 convolution. This configuration facilitates cross-channel communication along the channel axis, adeptly identifying dependencies across all channels. It leverages parallel substructures to manage computational resources, thereby diminishing reliance on sequential operations and curtailing the depth of the network.

Furthermore, the EMA module introduces a unique approach for aggregating cross-spatial information, facilitating diverse feature integration across various spatial dimensions. Unlike conventional attention mechanisms that depend on reducing channel dimensions to model inter-channel interactions, EMA exhibits superior computational performance and enhanced generalization capabilities.

Taking advantage of EMA with its flexibility and lightweight characteristics, we integrated it into the FasterNet module and innovatively proposed the FasterNet-EMA module. The modular design of the model network is shown in Fig. 8, which further optimizes the computational efficiency and model performance, making it particularly suitable for real-time application scenarios with limited resources. These improvements not only enhance the model functionality but also improve its practicality and accuracy in advanced visual tasks.

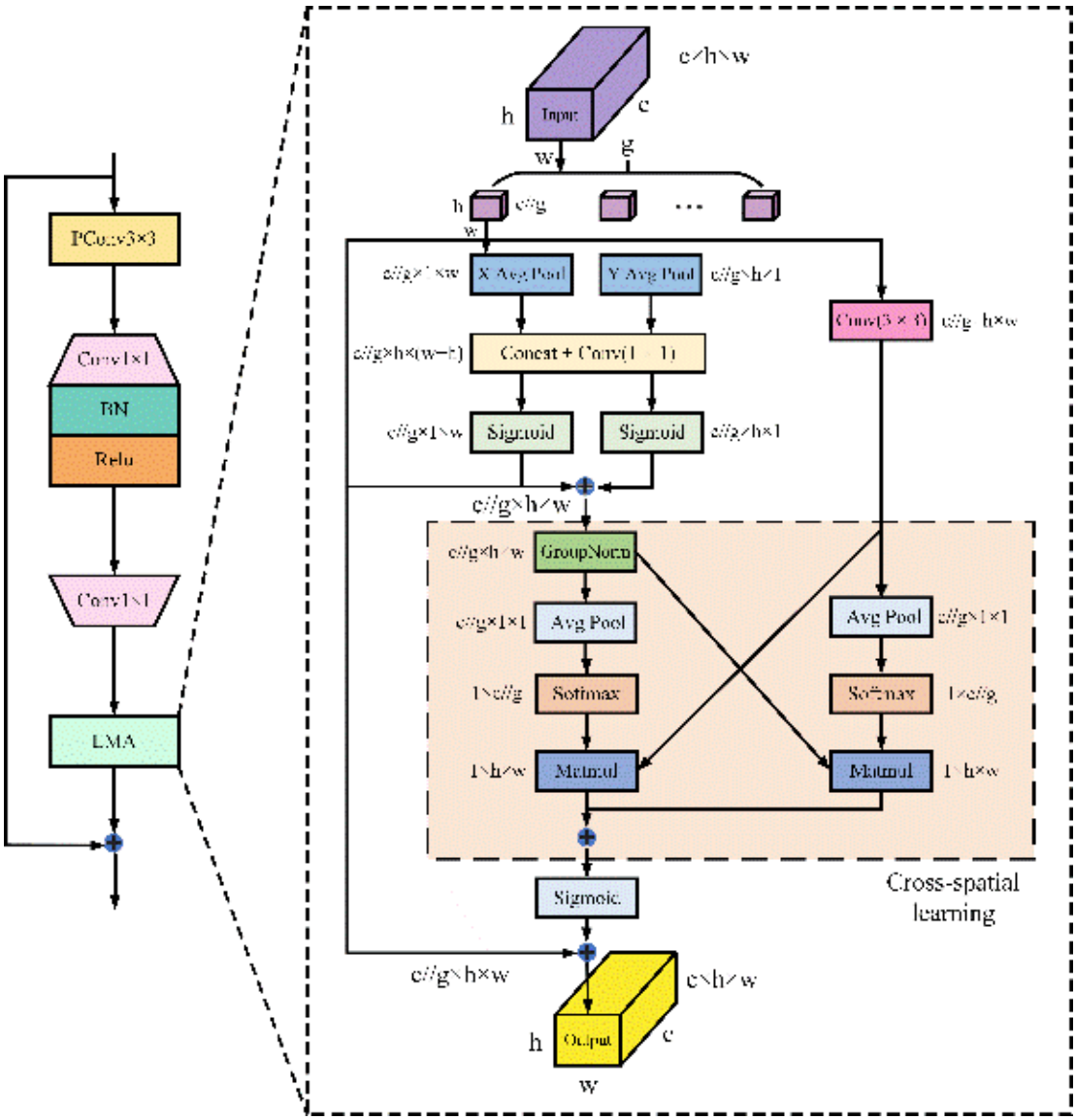


Fig. 8 EMA mechanism.

3.4 Ghost Module

To improve the efficiency of pre-trained networks, pruning, quantization, and other techniques are usually used to reduce parameters and computation. In the above field, Han et al. [30] proposed a novel Ghost module, which aims to generate rich feature maps with fewer parameters. The aforementioned model is predicated on recognizing the redundancies present in feature maps within deep neural networks. Although redundancy may be a key factor in the success of deep learning networks, it is not necessary to retain all of them.

Fig. 9 illustrates the function of a standard convolutional layer, the dimension of the input data is $c \times h \times w$, and convolving it with n groups of $k \times k$ convolution kernels, the dimension of the output data is $n \times h' \times w'$. Here, c, h, w, k, n, h' and w' represent the number of channels, height, width, convolution kernel size of the input data, and the number of output channels, output height, and output width, respectively.

The Ghost module aims to reduce computational requirements by minimizing the number of convolution operations. Its structure is shown in Fig. 4. In the Ghost

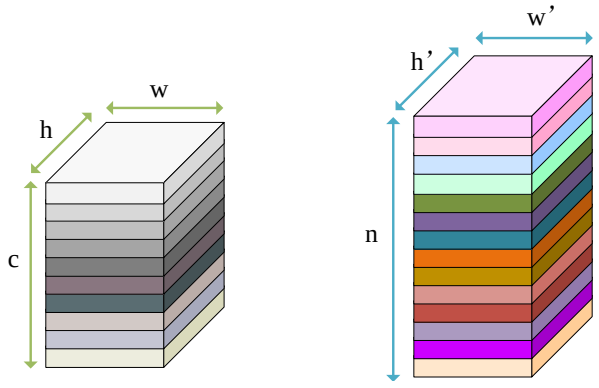


Fig. 9 Convolution module.

module, the output feature map consists of n channels, including one group of identity transformations using conventional convolution and $s-1$ groups of transformations using linear convolution, where s is much smaller than c. Assuming that the convolution kernel size used in each group of linear convolution is $d \times d$, the computational speedup ratio of Conv compared to Ghost can be expressed by formula:

$$r_s = \frac{n \times c \times h' \times w' \times k \times k}{n \times c \times h' \times w' \times k \times k + (s-1) \times \frac{n}{s} \times h' \times w' \times d \times d}$$

Because $d = k$, and $s << c$, simplifying equation:

$$r_s \approx \frac{s \times c}{s + c - 1} \approx s$$

It can be inferred that the Ghost module, as shown in Fig. 10, markedly lowers computational complexity relative to traditional Convolution (Conv), thereby simplifying the model and enhancing processing speed.

In the YOLOv8s model, the backbone part performs a large number of convolution operations, which enhances the expressiveness of features but also leads to a significant rise in parameter count. Because of the memory limitations and response speed requirements in the deployment of automated equipment in this study, the Conv module and C2f module of the backbone part were modified and integrated with the Ghost module to obtain Conv-Ghost and C2f-Ghost, intended to decrease the count of model parameters and boost the model's computational speed. The configuration is depicted in Fig. 11.

3.5 WIoU Bounding Box Regression Loss Function

In machine vision, object detection tasks revolve around employing computational algorithms and mathematical frameworks to precisely identify and pinpoint regions of interest within images. The success of object detection models heavily hinges on the architecture of their loss functions, particularly the Bounding Box Regression Loss function, which plays a vital role in fine-tuning the coordinates of predicted bounding boxes to closely match the true locations. This process is essential for reducing discrepancies between predicted and actual bounding boxes, thereby boosting the model's accuracy.

Nonetheless, an excessive focus on refining the boundaries of low-quality anchor boxes can lead to significant errors in object positioning. To counteract this, Zhang et al. [31] proposed the Focal-EIoU v1 Bounding Box Regression Loss function. This function integrates a static focusing mechanism designed to equilibrate the gradients between high-quality and low-quality anchor boxes, optimizing gradient allocation and enhancing the model's capacity to generalize across various scenarios.

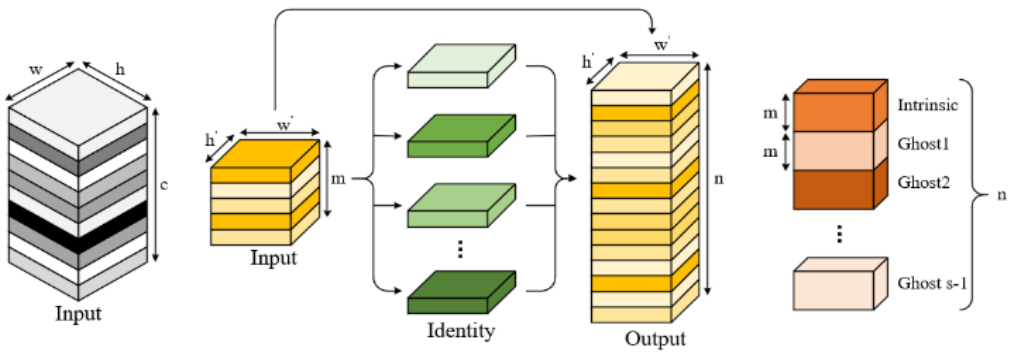


Fig. 10 Ghost module.

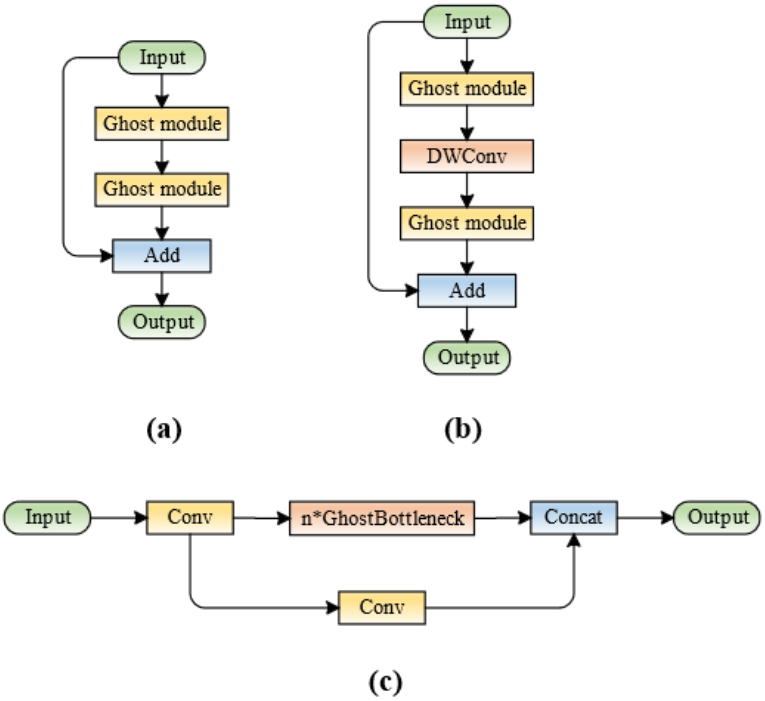


Fig. 11 Ghost-related building blocks. (a) GhostBottleneck, stride = 1; (b) GhostBottleneck, stride = 2; (c) C2f-Ghost module.

Building on this concept, Tong et al. [32] introduced a dynamic, non-monotonic focusing mechanism called WIoU, which utilizes the intersection over union (IoU) metric. This innovative loss function differentiates the quality of anchor boxes by leveraging outliers—defined as the extent of deviation between the predicted anchor box and the ground-truth box. The WIoU loss function dynamically adjusts the gradient distribution, assigning smaller gradients to high-quality anchor boxes with minor deviations, while allocating larger gradients to low-quality anchor boxes with significant deviations. This dynamic approach mitigates the impact of adverse gradients from low-quality boxes, reduces competitive interference among high-quality boxes, and emphasizes medium-quality boxes, collectively enhancing the overall performance of the detection model. Moreover, WIoU addresses additional challenges such as occlusion, small object detection, and inconsistent annotation quality within datasets. Compared to traditional bounding box loss functions, WIoU demonstrates superior performance. The calculation of the WIoU loss function is given by following formula:

$$L_{WIoU} = r \times (1 - IoU) \times \exp\left(\frac{(x - x_{gt})^2 + (y - y_{gt})^2}{\left(W_g^2 + H_g^2\right)^*}\right)$$

where IoU represents the intersection between the prediction box and the target box in object detection, x and y represent the horizontal and vertical coordinates of the center point of the prediction box, x_{gt} and y_{gt} represent the horizontal and vertical coordinates of the center point of the target box, and W_g and H_g represents the width and height of the minimum enclosed rectangular area of the prediction box and the target box respectively. To prevent $e^{\left((x-x_{gt})^2+(y-y_{gt})^2\right)/\left(w_g^2+h_g^2\right)}$ from generating gradients that hinder convergence, we separate W_g and H_g from the computation through superscript *. r is the non-monotonic clustering coefficient, which is calculated as shown in following formula:

$$r = \frac{\beta}{\delta \alpha^{\beta - \delta}}, \beta = \frac{L_{IoU}^*}{\bar{L}_{IoU}} \in [0, +\infty] L_{IoU} = 1 - IoU$$

where α and δ are hyperparameters, β is the outlier estimate of the prior box, $\bar{L}_{IoU}$ represents the average value of L_{IoU}, and L_{IoU}^* represents the separation value of L_{IoU}.

4. Testing Experiments

4.1 Experimental Details

To ensure the fairness and rationality of all experimental results in subsequent analysis, all experiments were conducted on the same device, and the specific hardware and software configurations are shown in Table 2. The initial parameter settings during deep learning are determined based on the complexity of the computer's hardware devices and network models. After multiple rounds of debugging, batch size, works, epochs, and image size were set to 8, 10, 100, 640 ×640, respectively. The weight files trained on the COCO2017 public dataset were used for transfer learning to improve the convergence speed and accuracy of the model. The remaining hyperparameter settings are shown in Table 3.

4.2 Evaluation Metrics

To thoroughly assess the model's performance, this study carried out two types of experiments on strawberry maturity detection: ablation experiments and comparative experiments. The complexity of the model was evaluated using parameters (Params) and gigaflops per second (GFLOPs). The detection speed was measured in FPS. To measure the accuracy of strawberry maturity detection, precision (P), recall (R), and average precision (AP) were employed as performance indicators. The definitions of these indicators are provided in the following formulas:

$$\mathrm{Pr}ecision = \frac{TP}{TP + FP} \times 100\%$$

$$\mathrm{Re}call = \frac{TP}{TP + FN} \times 100\%$$

$$AP = \int_0^1 p(r)dr$$

True Positives (TP) represent instances where the model correctly identifies target objects. Conversely, False Negatives (FN) occur when the model fails to detect existing target objects. False Positives (FP) are cases where the model incorrectly classifies non-target objects as targets. To evaluate the model's performance, the precision-recall (PR) curve is generated by plotting precision against recall across various confidence thresholds. The average precision (AP) is then computed by averaging the precision values at each recall level along the PR curve. Before assessing detection speed, it is essential to release memory and GPU resources, followed by preheating the hardware to achieve optimal performance. Finally, the inference time of the model, or FPS value, is determined by executing validation steps.

4.3 Ablation Experiments

Four improvements are made to the original YOLOv8s model (A: FasterNet Block is used to replace C2f in the neck, and the EMA attention mechanism is integrated; B: Ghost is integrated into the model to form Conv-Ghost module and C2f-Ghost module respectively; C: Wise-IoU is used as the loss function). Each improved module is integrated into the original model separately. The efficacy of the enhanced module is verified by ablation experiments, and the corresponding results are shown in Table 4.

Table 2 Hardware and software configuration.

Hardware/software	Configuration/version
Operating system	Windows 11 professional edition
CPU	AMD Ryzen 9 7945HX with Radeon Graphics
GPU	NVIDIA GeForce RTX 4060 Laptop GPU
Memory	Micron Technology 40G DDR5 5200
Hard disk	SKHynix_HFS001TEJ9X115N 1TB
integrated development environment	PyCharm 2023.3 + Anaconda3 2023.03
programming language	Python 3.12.2
Development framework	Pytorch2.3.1+cuda11.8+cudnn8700

Table 3 Training initial and hyper parameters.

Parameters	Form/value
Optimiser	SGD
Initial learning rate	0.01
Final One Cycle learning rate	0.0001
Momentum	0.937
Lr-decay-type	Cos
Weight-decay	0.0005
Intersection over Union	0.5

Table 4 Comparison of ablation experiments of different models on the test set.

A	B	C	P (%)	R (%)	mAP50 (%)	Params (M)	GFlops (G)	FPS	Model size (MB)
×	×	×	83.3	83.5	91.1	11.4	29.4	188	22.5
√	×	×	85.5	85.4	92.7	10.0	26.6	158	19.8
×	√	×	84.7	84.0	92.4	8.5	21.6	217	17.0
×	×	√	88.5	84.9	93.7	11.4	29.4	217	22.5
√	√	×	86.5	84.7	93.3	7.2	18.7	208	14.0
√	×	√	88.8	85.7	94.2	10.0	26.9	188	19.9
×	√	√	85.5	84.9	93.3	8.6	21.6	212	17.0
√	√	√	88.4	86.2	93.8	7.2	19.2	212	14.0

By substituting the Bottleneck structure in C2f with the FasterNet Block, we achieved a reduction in Params and GFlops by 12.28% and 9.5% respectively, indicating a decrease in model complexity and the number of parameters. Incorporating the EMA attention mechanism into the base model resulted in a 2.2% increase in Precision and a 2.3% increase in mAP, with minimal impact on parameter count and computational load. The integration of the attention mechanism significantly boosted the network's feature extraction capabilities and minimized the impact of irrelevant information. When the Ghost module was added to the model, there was a notable decrease in Params and GFlops by 24.9% and 26.5% respectively, further reducing the model's complexity. Additionally, substituting the CIoU loss function with the WIoU loss function, and using outliers to assess the anchor box quality while reducing the gradient's impact on low-quality samples, enhanced the model's precision and mAP by 5.2% and 1.4% respectively. The combined use of the C2f-Faster-EMA model, Ghost module, and WIoU loss function resulted in the model achieving optimal performance. Compared to the original model, the improved version reduced Params and GFlops by 36.8% and 34.6% respectively. Improvements in Precision, Recall, and mAP50 were observed at 5.1%, 2.7%, and 2.7% respectively, alongside a 12.7% increase in FPS. Fig. 12 provides a visual comparison of heatmaps from detection results before and after the YOLOv8s improvements.

In this study, the final version of the improved YOLOv8s model exhibited the highest overall detection performance among the five compared network models. Compared to the original YOLOv8s, the YOLOv8s-FEGW model showed a 5.1% improvement in *p* value and a 2.7% increase in mAP. The frame rate analysis revealed that the original YOLOv8s model operated at 188 FPS, whereas the enhanced version achieved 212 FPS. The mAP50 curves of the five models are shown in Fig. 13. The confusion matrix before and after the improvement is shown in Fig. 14.

Fig. 15 highlights the differences in recognition performance between the original YOLOv8s and the newly proposed YOLOv8s-FEGW model on various strawberry images. By examining part (a), the enhanced YOLOv8s-FEGW model offers superior performance in detecting multiple targets within

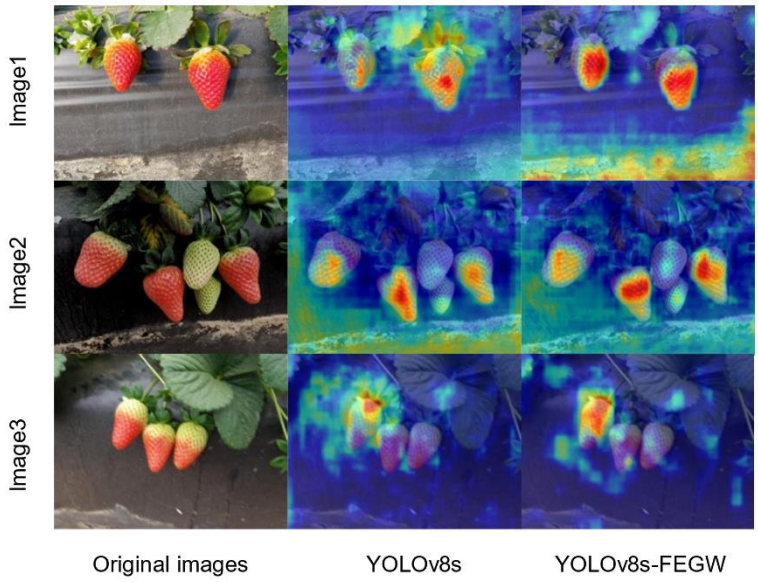


Fig. 12 Heatmap visualization of YOLOv8s before and after improvement.

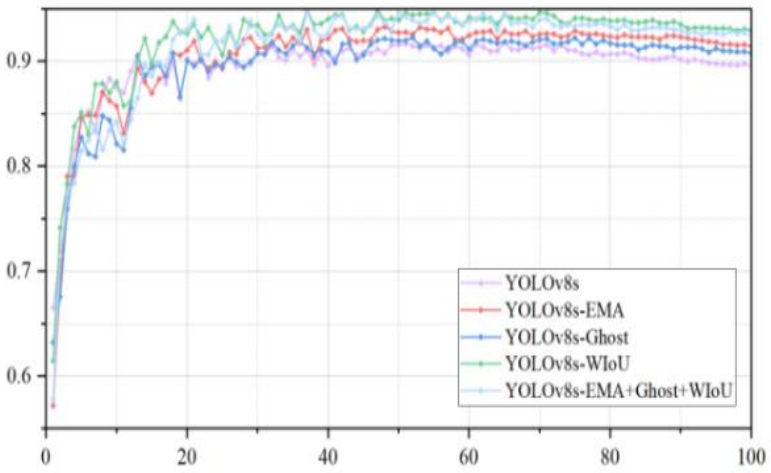
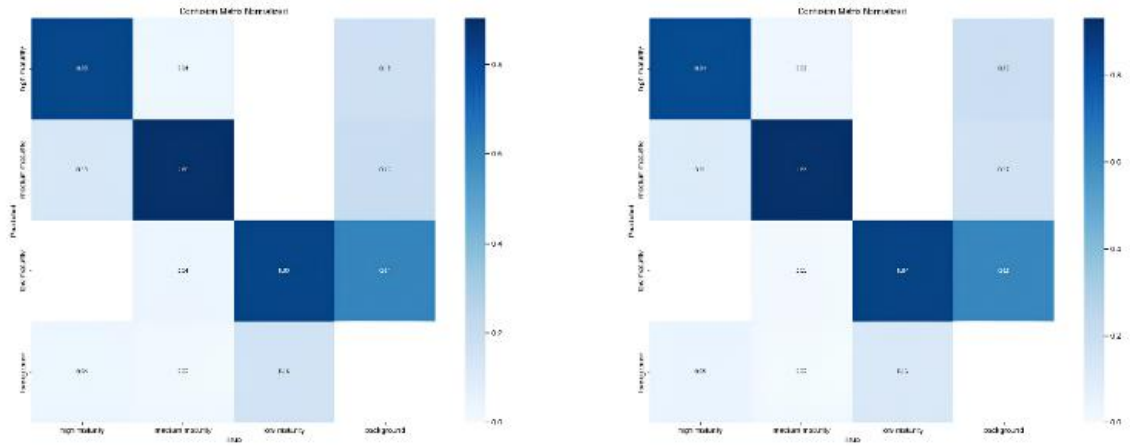


Fig. 13 mAP50 curves of different models under ablation experiment.



(a) before improvement (b) after improvement

Fig. 14 Confusion matrix before and after improvement.

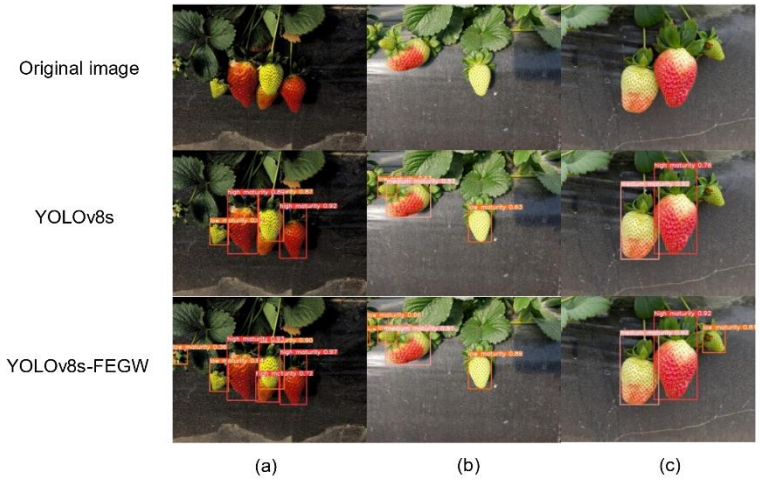


Fig. 15 Effect diagram before and after improvement.

intricate environments. Unlike the original model, this improved version can identify strawberries even at the edges of the detection zone and can recognize targets partially obscured by other objects. As shown in part (b), the model effectively detects strawberry targets concealed by branches and leaves, accurately identifying them even when they resemble their background. In part (c), there is a significant improvement in detecting small strawberry targets with the enhanced YOLOv8s-FEGW model. The addition of the Faster-EMA module significantly enhances the model's ability to detect small objects by expanding the receptive field, which improves its performance in recognizing edge cases and in complex environments. The Ghost module has been integrated to reduce the overall model complexity, facilitating easier deployment and more efficient lightweight detection. Furthermore, the use of the WIoU loss function enhances the model's ability to recognize features and improve detection accuracy for difficult-to-detect targets, effectively reducing both missed detections and false positives. In conclusion, the refined YOLOv8s-FEGW model outperforms the original YOLOv8s in terms of detection capabilities and reliability, providing a more efficient and lightweight solution for strawberry detection.

4.4 Comparative Experiments

To evaluate the performance of the enhanced algorithm described in this paper, we conducted comparative experiments involving YOLOv3s, YOLOv5s, YOLOv6s, YOLOv8s, YOLOv9s, YOLOv8s-FEGW, and RT-DETRs. These experiments were performed using identical hardware, datasets, and data augmentation techniques, ensuring that the ratio of the training set to the test set remained constant. Each experiment was repeated 100 times, and the best outcome was used for analysis. The results comparing accuracy, recall, mAP, frame rate, and model size are presented in Table 5.

Table 3 illustrates that the YOLOv8s-FEGW model introduced in this research surpasses the YOLOv3s model in terms of Params and GFlops. Compared to the YOLOv8s prior to enhancement, there has been a reduction of 36.8% in Params and 34.6% in GFlops, achieving the objective of being more lightweight and better suited for agricultural equipment deployment. While YOLOv9s and RT-DETRs exhibit higher mAP50 values, which are 0.6% and 0.4% greater than those of YOLOv8s-FEGW respectively, they fall short in recall and precision compared to YOLOv8s-FEGW. This discrepancy is due to multiple detections of the same target or misidentification of branches and leaves, as detailed in Figs. 16-18. In terms of FPS, YOLOv8s-FEGW has a notable advantage, being 12.7% higher than the pre-improvement of YOLOv8s and achieving the highest FPS among the models compared. The mAP50 of YOLOv8s-FEGW is 3.5%, 3.3%, 2.1%, and

2.7% higher than those of YOLOv3s, YOLOv5s, YOLOv6s, and the original YOLOv8s, respectively. The improved model enhances the efficiency of feature extraction and integration for detecting dense strawberry targets by leveraging the strengths of EMA and WIoU. By incorporating the Ghost model, it achieves a substantial reduction in both the number of parameters and overall model complexity. Consequently, the algorithm meets the demands of real-time detection, significantly improves accuracy, streamlines model size, and offers excellent versatility and practical application potential.

Table 5 The comparative performance of various models on the test set.

Model	P (%)	R (%)	mAP50 (%)	Params (M)	GFlops (G)	FPS (f/s)	Model size (MB)
YOLOv3s	82.4	81.7	90.3	4.1	12.2	232	8.3
YOLOv5s	82.8	82.2	90.5	9.4	24.8	204	18.6
YOLOv6s	84.6	82.7	91.7	16.3	44.3	185	33.2
YOLOv8s (Base Model)	83.3	83.5	91.1	11.4	29.4	188	22.5
YOLOv9s	89.5	87.4	94.4	26.1	106	181	52.1
YOLOv8s-FEGW (Our)	88.4	86.2	93.8	7.2	19.2	212	14.0
RT-DETRs	87.9	85.6	94.2	11.4	39.2	114	22.9

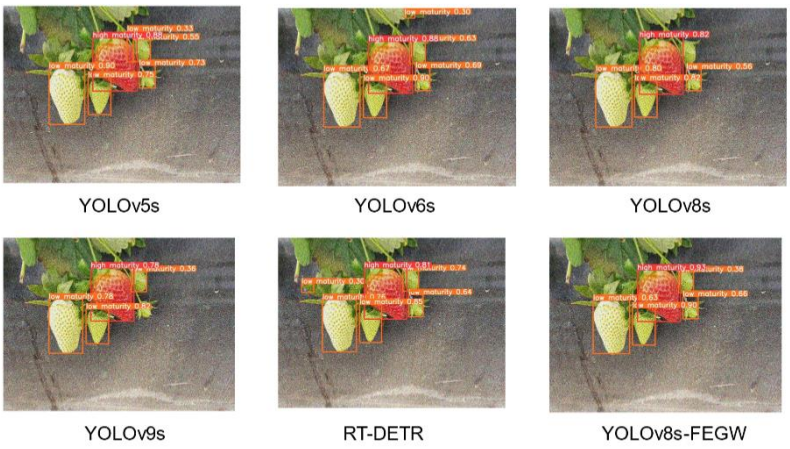


Fig. 16 Detection results of strawberry maturity under different models under dense conditions.

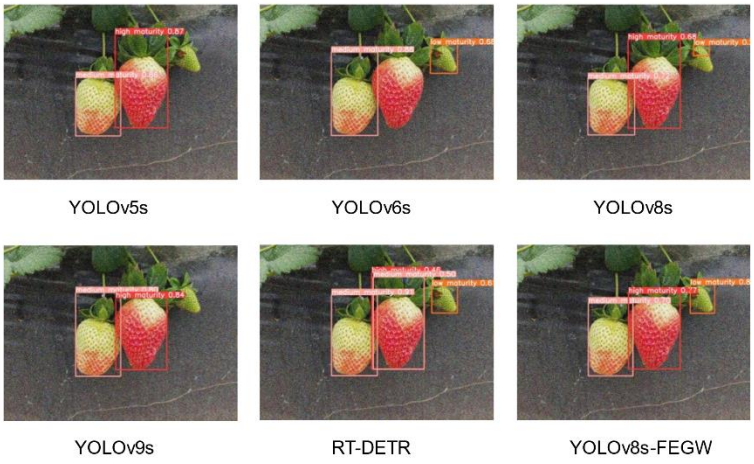


Fig. 17 Detection results of strawberry maturity under different models under cloudy conditions.

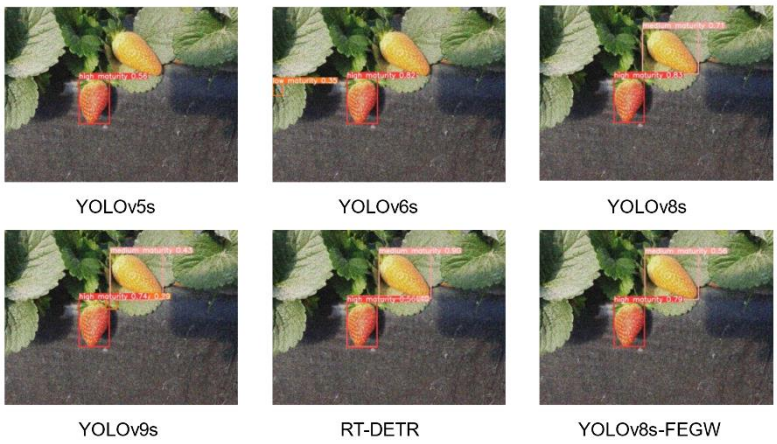


Fig. 18 Detection results of strawberry maturity under different models under sunny conditions.

5. Conclusions

YOLOv8-FEGW presents a highly efficient and lightweight approach to determining the maturity of strawberries, achieving accurate detection even in complex conditions and with limited computational capabilities. By incorporating Faster-EMA, Ghost, and WIoU methodologies into YOLOv8s, we have developed the streamlined YOLOv8s-FEGW model. The experimental data demonstrate a notable improvement in strawberry maturity detection accuracy, increasing from 91.1% to 93.8%. When benchmarked against several leading contemporary target detection algorithms, the model proposed here exhibits a marked competitive edge in precision. Additionally, there is a significant reduction in computational demands, with GFlops decreasing from 29.4 G to 19.2 G and parameters from 11.4 M to 7.2 M, representing decreases of 30.6% and 33.2%, respectively. These results indicate that the optimized YOLOv8-FEGW model, which integrates multiple techniques, offers substantial robustness and practicality for accurate detection and easy model deployment, making it ideal for use in harvesting robots. The main findings of this research are as follows:

(1) In ablation experiments, the mAP of the optimized YOLOv8s-FEGW increased by 2.7%, reaching 93.8% during tests. Compared to the original YOLOv8s architecture, the advanced YOLOv8s-FEGW model shows important improvement across all key metrics. Furthermore, the enhanced YOLOv8s-FEGW model demonstrates greater reliability in strawberry maturity detection, exhibiting a lower rate of missed detections and a higher mAP than other models.

(2) In tests using the strawberry maturity dataset, the refined YOLOv8s model was evaluated against YOLOv3s, YOLOv5s, YOLOv6s, YOLOv9s, RT-DETRs, and the original YOLOv8s model. The results show that the upgraded YOLOv8s-FEGW model achieves a well-balanced performance regarding model size, mAP, and detection frame rate. The model size is 7.2 MB, the mAP is 93.8%, and the frame rate is 212 FPS, meeting the requirements for real-time agricultural detection.

The experimental results underscore the model's substantial potential for application in strawberry maturity detection. However, this research has limitations, such as not fully accounting for strawberry varieties with unique skin colors. Future work could see the enhanced YOLOv8s model integrated with intelligent inspection and picking robots, facilitating efficient and high-quality harvesting through an AI-based strawberry maturity detection system.

Author Contributions

Yiming Cheng: Conceptualization, Data curation, Methodology, Writing—original draft, Writing—review & editing, Software. Guohao Feng: Conceptualization,

Methodology, Software, Writing—original draft, Writing—review & editing, Visualization. Chunchang Zhang: Conceptualization, Data curation, Methodology, Writing—original draft, Writing—review & editing, Software. All authors have read and agreed to the published version of the manuscript.

Funding

This research was funded by the National Engineering Research Center of Special Equipment and Power System for Ship and Marine Engineering and the Shanghai Engineering Research Center of Ship Intelligent Maintenance and Energy Efficiency Control (20DZ2252300).

Institutional Review Board Statement

Not applicable.

Informed Consent Statement

Not applicable.

Data Availability

The data that have been used are confidential.

Conflicts of Interest

The authors declare no conflicts of interest.

Reference

[1] Jiao, L., Zhang, F., Liu, F., Yang, S., Li, L., Feng, Z., and Qu, R. 2019. "A Survey of Deep Learning-Based Object Detection." *IEEE Access* 7: 128837-68. https://doi.org/10.1109/ACCESS.2019.2939201.

[2] Hu, H., Kaizu, Y., Zhang, H., Xu, Y., Imou, K., Li, M., Huang, J., and Dai, S. 2022. "Recognition and Localization of Strawberries from 3D Binocular Cameras for a Strawberry Picking Robot Using Coupled YOLO/Mask R-CNN." *Int. J. Agric. Biol. Eng*. 15 (6): 175-9. https://doi.org/ 10.25165/j.ijabe.20221506.7306.

[3] Yang, S., Wang, R., Gao, S., and Shao, M. 2023. "YOLOv5-Based Lightweight Network Model for Strawberry Detection." *For. Electron. Measur. Technol*. 42: 86-95. doi: 10.19652/j. cnki.femt.2304668.

[4] Chen, G., Hu, J., Li, D., et al. 2012. "Image Recognition of Maize Diseases Based on Fuzzy Clustering and Support Vector Machine Algorithm." *Sensor Letters* 10 (1-2): 433-8.

[5] Bachhal, P., Kukreja, V., Ahuja, S., et al. 2024. "Maize Leaf Disease Recognition Using PRF-SVM Integration: A Breakthrough Technique." *Scientific Reports* 14 (1): 10219.

[6] Mustafa, N. B. A., Arumugam, K., Ahmed, S. K., et al. 2011. "Classification of Fruits Using Probabilistic Neural Networks-Improvement Using Color Features." In *Proceedings of the TENCON 2011-2011 IEEE Region 10 Conference*, pp. 264-9.

[7] Li, M., Wang, Q., and Zhu, J. 2012. "Automatic Recognition of Grapes' Size Level Based on Machine Vision." *Journal of Food Agriculture & Environment* 10 (3-4): 78-80.

[8] Qiaohua, W., Yihua, T., and Zhuang, X. 2017. "Grape Size Detection and Online Gradation Based on Machine Vision." *International Journal of Agricultural and Biological Engineering* 10 (1): 226-33.

[9] Reis, M. J. C. S., Morais, R., Pereira, C., et al. 2011. "A Low-Cost System to Detect Bunches of Grapes in Natural Environment from Color Images." In *Proceedings of the Advanced Concepts for Intelligent Vision Systems: 13th International Conference, ACIVS 2011*, Ghent, Belgium, August 22-25, 2011, pp. 92-102.

[10] Arefi, A., Motlagh, A. M., Mollazade, K., et al. 2011. "Recognition and Localization of Ripen Tomato Based on Machine Vision." *Australian Journal of Crop Science* 5 (10): 1144-9.

[11] Goyal, K., Kumar, P., and Verma, K. 2023. "AI-Based Fruit Identification and Quality Detection System." *Multimedia Tools and Applications* 82 (16): 24573-604.

[12] Ma, L., Guo, X., Zhao, S., et al. 2021. "Algorithm of Strawberry Disease Recognition Based on Deep Convolutional Neural Network." *Complexity* 2021 (1): 6683255.

[13] Javanmardi, S., Ashtiani, S. H. M., Verbeek, F. J., et al. 2021. "Computer-Vision Classification of Corn Seed Varieties Using Deep Convolutional Neural Network." *Journal of Stored Products Research* 92: 101800.

[14] Ashtiani, S. H. M., Javanmardi, S., Jahanbanifard, M., et al. 2021. "Detection of Mulberry Ripeness Stages Using Deep Learning Models." *IEEE Access* 9: 100380-94.

[15] Jeong, H., Moon, H., Jeong, Y., et al. 2024. "Automated Technology for Strawberry Size Measurement and Weight Prediction Using AI." *IEEE Access* 11: 1.

[16] Chen, S., Liao, Y., Lin, F., et al. 2023. "An Improved Lightweight YOLOv5 Algorithm for Detecting Strawberry Diseases." *IEEE Access* 11: 54080-92.

[17] Ridho, M. F. 2021. "Strawberry Fruit Quality Assessment for Harvesting Robot Using SSD Convolutional Neural Network." In *Proceedings of the 2021 8th International Conference on Electrical Engineering, Computer Science and Informatics (EECSI)*, pp. 157-62.

[18] Kim, S. J., Jeong, S., Kim, H., et al. 2022. “Detecting Ripeness of Strawberry and Coordinates of Strawberry Stalk Using Deep Learning.” In *Proceedings of the 2022 Thirteenth International Conference on Ubiquitous and Future Networks (ICUFN)*, pp. 454-8.

[19] Wang, G., Zheng, H., and Li, X. 2023. “ResNeXt-SVM: A Novel Strawberry Appearance Quality Identification Method Based on ResNeXt Network and Support Vector Machine.” *Journal of Food Measurement and Characterization* 17 (5): 434556.

[20] Yang, S., Wang, W., Gao, S., and Deng, Z. 2023. “Strawberry Ripeness Detection Based on YOLOv8 Algorithm Fused with LW-Swin Transformer.” *Computers and Electronics in Agriculture* 215: 108360.

[21] Chen, J., Ma, A., Huang, L., Li, H., Zhang, H., Huang, Y., and Zhu, T. 2024. “Efficient and Lightweight Grape and Picking Point Synchronous Detection Model Based on Key Point Detection.” *Computers and Electronics in Agriculture* 217: 108612.

[22] Bochkovskiy, A., Wang, C.-Y., and Liao, H.-Y.-M. 2020. “Yolov4: Optimal Speed and Accuracy of Object Detection.” arXiv preprint arXiv:2004.10934. https://doi.org/10.4855 0/arXiv.2004.10934.

[23] Redmon, J., and Farhadi, A. 2018. “Yolov3: An Incremental Improvement.” arXiv preprint arXiv: 1804.02767. https://doi.org/10.48550/arXiv.1804.02767.

[24] Redmon, J., and Farhadi, A. 2017. “YOLO9000: Better, Faster, Stronger.” In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7263-71. https://doi.org/10.48550/arXiv.1612.08242.

[25] Wang, C.-Y., Bochkovskiy, A., and Liao, H.-Y.-M. 2023. “YOLOv7: Trainable Bag-of-Freebies Sets New State-of-the-Art for Real-Time Object Detectors.” In *Proceedings of the IEEE/ CVF Conference on Computer Vision and Pattern Recognition*, pp. 7464-75. https://doi.org/10.48550/ arXiv.2207.02696.

[26] Feng, C., Zhong, Y., Gao, Y., Scott, M. R., and Huang, W. 2021. “Tood: Task-Aligned One-Stage Object Detection.” In *2021 IEEE/CVF International Conference on Computer Vision (ICCV). IEEE Computer Society*, pp. 3490-9. https://doi.org/10.48550/ arXiv.2108.07755.

[27] Chen, J., Kao, S.-H., He, H., Zhuo, W., Wen, S., Lee, C.-H., and Chan, S.-H.-G. 2023. “Run, Don’t Walk: Chasing Higher FLOPS for Faster Neural Networks.” In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1202112031. https://doi.org/10.48550/arXiv.2303.03667.

[28] Ouyang, D., He, S., Zhang, G., Luo, M., Guo, H., Zhan, J., and Huang, Z. 2023. “Efficient Multi-scale Attention Module with Cross-Spatial Learning.” In *Proceedings of the ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1-5. https://10.1109/ICASSP49357.2023.10096516.

[29] Hou, Q., Zhou, D., and Feng, J. 2021. “Coordinate Attention for Efficient Mobile Network Design.” In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13713-22. https://doi.org/10.48550/arXiv.2103.02907.

[30] Han, K., Wang, Y., Tian, Q., Guo, J., Xu, C., and Xu, C. 2020. “Ghostnet: More Features from Cheap Operations.” In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1580-9.

[31] Zhang, Y. F., Ren, W., Zhang, Z., Jia, Z., Wang, L., and Tan, T. 2021. “Focal and Efficient IOU Loss for Accurate Bounding Box Regression.” *Neurocomputing* 506: 146-57. https://doi.org/ 10.1016/j.neucom.2022.07.042.

[32] Tong, Z., Chen, Y., Xu, Z., and Yu, R., 2023. “Wise-IoU: Bounding Box Regression Loss with Dynamic Focusing Mechanism.” arXiv preprint. https://doi.org/10.48550/ arXiv.2301.10051arXiv:2301.10051.