RESEARCH ON A CHERRY MATURITY DETECTION MODEL BASED ON IMPROVED YOLOV11N
/
基于改进 YOLOv11n 的樱桃成熟度检测模型研究

Zhixiang FENG[1)], Xuanyu CAO[1)], Hao JI[1)], Jiarui ZHANG[1)], Jianyu CHEN[1)], Shuo LIU[1)], Lijun CHENG*[1)]
[1)] College of Software, Shanxi Agricultural University, Shanxi / China;
Correspondent authors: Lijun CHENG: Tel: 13835441585; E-mail: cljzyb@sxau.edu.cn;
DOI: https://doi.org/10.35633/inmateh-77-51

Keywords: *Cherry, YOLOv11, Attention Mechanism, Deformable Convolution, Maturity Detection*

ABSTRACT

Currently, research on cherry detection and recognition is relatively limited, and existing methods for agricultural product inspection often suffer from slow speed and low classification accuracy. To address these issues, this paper introduces an improved YOLOv11n-based model for detecting cherry ripeness, designed to enhance both the accuracy and efficiency of identifying cherries at different maturity stages. First, improvements were made to the backbone network of the YOLOv11n model by replacing the original backbone with ConvNeXtv2. This replacement achieved a broader global receptive field and enhanced multi-scale learning, which helped reduce computational costs and significantly improve efficiency while maintaining high performance. Second, a DCNv4 convolution module—an advanced convolutional layer with adaptive receptive fields—was added to the neck of the model. The neck is an intermediate stage that combines features from different layers, and the DCNv4 adapts the receptive field to help accurately locate occluded cherries of any shape and scale. This improves detection performance for small cherries without increasing computational complexity. Finally, the convolutional attention module CBAM was introduced. CBAM adaptively focuses on important image features while suppressing irrelevant background by using both channel and spatial attention mechanisms. Together, these additions significantly improve cherry detection accuracy and robustness. Our experimental results show that the improved M-YOLOv11n algorithm achieved a 4.84% increase in mAP@50 compared to the original YOLOv11n model. Precision and recall also improved by 1.25% and 0.4%, respectively. Overall, the enhanced model outperformed not only its base version but also the YOLOv5n and YOLOv8n models. Compared to multi-stage models, the proposed model demonstrates superior accuracy, speed, and reduced computational requirements. This improvement enables more efficient and precise identification of cherry ripeness, thereby enhancing the efficiency of cherry harvesting and facilitating optimal harvest timing. These advancements support the optimization of storage and transportation conditions for cherries and provide robust technical support for intelligent orchard management and the advancement of automated fruit sorting systems.

摘要

针对当前樱桃检测与识别研究较少，农产品检测与识别速度慢、分类精度低等问题，本文提出了一种基于改进 YOLOv11n 的樱桃成熟度检测模型，旨在提高不同成熟度的樱桃检测的准确性和效率。首先，针对 YOLOv11n 模型的主干网络进行了改进，将原有的主干网络替换为 ConvNeXtv2，通过替换主干网络 CSPDarknet11 实现全局的感受野和多尺度学习，有助于降低计算成本，在保持高性能的同时，显著提高了计算效率。其次，在模型的颈部添加了 DCNv4 卷积模块，通过自适应地调整膨胀卷积的感受野，精准定位任意形状、任意尺度被遮挡的樱桃，在不增加额外计算量的同时改善小目标的检测效果。最后，引入卷积注意力模块 CBAM，通过协同利用通道与空间注意力机制，自适应地聚焦关键特征并抑制背景干扰，从而显著提升模型对樱桃的检测精度与鲁棒性。实验结果表明，改进后的算法 M-YOLOv11n 相比原 YOLOv11n 模型 mAP@50 提高了 4.84 个百分点，精确率和召回率分别提高了 1.25 个百分点和 0.4 个百分点，均优于 YOLOv5n、YOLOv8n 和 YOLOv11n 模型。此外,与多阶段模型相比,该模型在平均精度、效率和计算负载方面均表现优越。由此可见,改进后的模型能够更加高效、精准地进行樱桃成熟度识别，这不仅提高了樱桃采摘效率，还精确控制了采摘时间，进而优化了果实的储存和运输条件，为果园智能化管理及水果自动分拣装备的开发提供了有效的技术支撑。

INTRODUCTION

Anthocyanins, vitamin C, potassium, and dietary fiber are all abundant in cherries, sometimes referred to as bird cherries, sweet cherries, or cherries. They are praised as the "diamond of fruits" due to their great nutritional content. The demand for cherries on the market is still rising as a result of improvements in consumption and the growth of international trade. China is now one of the world's biggest markets for cherry consumption, and the Food and Agriculture Organization of the United Nations (FAO) reports that worldwide cherry output has grown by about 35% in the last ten years *(Wang et al., 2025)*. Large-scale cherry cultivation enterprises have developed concurrently in areas like Shandong, Liaoning, Shaanxi, Gansu, and Sichuan, progressively moving toward precision and intelligent farming methods. As non-climacteric fruits, cherries' flavor and quality are mostly established when they are ripening on the tree. Therefore, it is crucial to make an accurate assessment of the optimal time to harvest. Cherry skins have a brief ripening time and are fragile and easily damaged. Conventional hand inspection techniques are expensive and ineffective, which makes them inappropriate for the demands of commercial harvesting on a wide scale. The assessment of ripeness is further complicated by the fact that variables, including variety, climate, and production methods, affect maturity indications like sugar content, hardness, and pigmentation. Therefore, improving harvesting efficiency, cutting losses, and guaranteeing fruit quality all depend on the development of intelligent cherry maturity detecting technology.

Many automated fruit and vegetable ripeness detection studies have surfaced worldwide as a result of the quick development of deep learning and computer vision technology. Although their ablation tests produced less than ideal results, Albarrak Khalied et al. used transfer learning based on the MobileNetV2 architecture to classify eight distinct data types with 99% accuracy *(Albarrak et al., 2022)*. Farjana Sultana Mim et al. created an automated classification system employing the HIS model by performing global threshold segmentation on mango photos using digital image processing techniques. However, poor model resilience and recognition mistakes resulted from the tiny dataset size *(Mim et al., 2018)*. Deep transfer learning was used by DANH et al. to classify tomatoes. According to experimental results, the VGG19 model detected the maturity of cherry tomatoes with an accuracy of 94.14% *(Danh et al., 2021)*. Chen et al. proposed a method that detects the ripeness of citrus fruits by combining visual saliency with convolutional neural networks to identify three levels of maturity *(Chen et al., 2022)*. Wu et al. proposed a DeepLabV3-based method to achieve rapid segmentation and recognition of cherries in complex orchard environments, including front light, backlight, rainy weather, single fruits, multiple fruits, fruit overlap, and branch/leaf shading *(Wu et al., 2024)*. Gai and colleagues developed an improved version of the YOLO-V4 deep learning algorithm that can effectively detect small cherry fruits in images *(Gai et al., 2023)*.

The YOLO (You Only Look Once) series of algorithms has demonstrated remarkable performance in real-time fruit and vegetable ripeness detection, owing to their high efficiency and accuracy. For instance, Wang Lishu et al. developed an improved YOLOv4-Tiny network, which achieved an average precision of 96.24% in classifying unripe, underripe, and ripe blueberries under challenging conditions such as occlusion and uneven lighting, with an average detection time of only 5.723 ms. This result satisfies both accuracy and speed requirements for practical blueberry recognition *(Wang et al., 2021)*. In another study, MACEACHERN et al. applied YOLOv4 to blueberry ripeness detection and reported high accuracy; However, the substantial computational cost of YOLOv4 led to a significant decrease in inference speed when deployed on resource-constrained embedded devices *(Maceachern et al., 2023)*. To address similar challenges in other fruit detection tasks, Liang Ao et al. introduced YOLOv5s-SCS, a real-time strawberry ripeness detection algorithm based on YOLOv5s. This method enhances detection performance in the presence of high fruit density, small target size, occlusion, overlap, and crowding by mitigating false positives and false negatives, thereby improving both detection accuracy and speed *(Liang et al., 2024)*. Similarly, Li Ying et al. proposed an improved YOLOv8s-based approach for citrus ripeness detection. By integrating a Hybrid Attention Transformer (HAT) module and an Adaptive Spatial Feature Fusion (FASFF) detection head, the model's ability to discern citrus ripeness was significantly strengthened *(Li et al., 2024)*. Furthermore, Tian Ronghui et al. developed an enhanced YOLOv7-ST-ASFF model for apple ripeness detection in complex orchard environments, where it exhibited outstanding performance, particularly in scenes containing multiple ripe apples and backlit unripe fruits *(Tian et al., 2022)*.

Traditional deep learning models are often plagued by high computational complexity, excessive parameters, and slow inference speeds, making them impractical for real-time detection and mobile deployment in agricultural environments. Consequently, developing lightweight versions of YOLO-series algorithms is of considerable practical significance. Recent years have witnessed progressive improvements in the YOLO family.

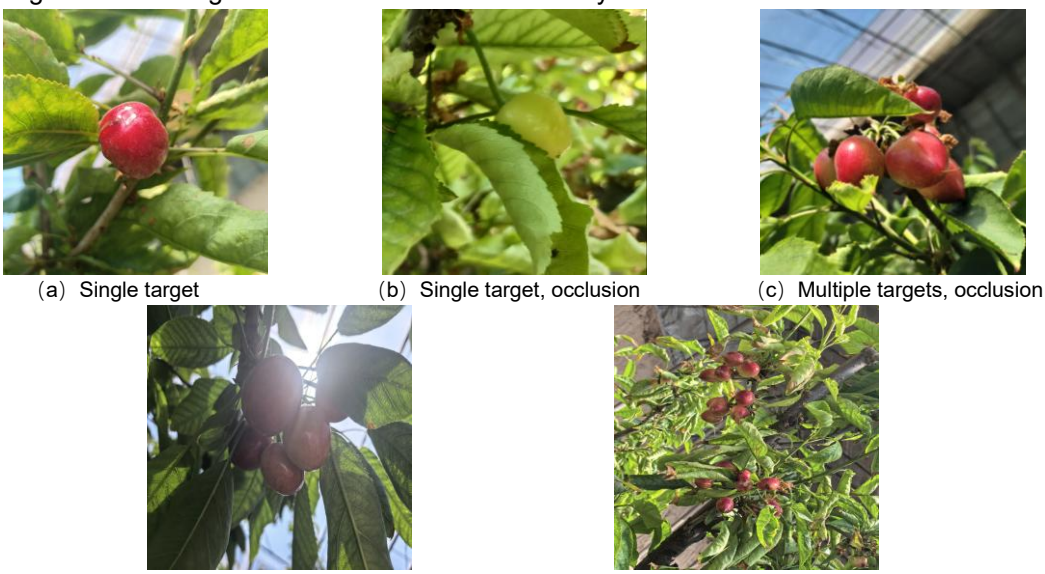
In 2024, YOLOv9 was introduced to enhance detection accuracy in complex scenarios through strengthened multi-scale feature fusion. It incorporates novel efficient convolutional layers that substantially reduce latency and hardware dependency, while a dynamic parameter adjustment mechanism enables real-time optimization, improving efficiency without sacrificing precision *(Li et al., 2024)*. Building on this, YOLOv10 employs adaptive feature enhancement to selectively accentuate critical regions, accelerates inference via a multi-branch parallel architecture, and minimizes redundancy through lightweight module design. These innovations allow it to achieve a more favorable accuracy-speed trade-off, particularly in mobile and real-time settings *(Gao et al., 2025)*. YOLOv11 further leverages the use of separable convolutions and enhances multi-scale representation learning, thereby reducing computational overhead while strengthening robustness to occluded and challenging targets. It also introduces a hierarchical feature fusion network that integrates both fine-grained details and rich semantic information. By combining dynamic kernel adaptation with learnable feature selection, YOLOv11 actively adjusts to varying contexts, resulting in synergistic gains in detection performance, inference speed, and generalization ability *(Wang et al., 2025)*.

Despite these advances, mainstream detection networks still face challenges in applications such as cherry detection, where targets are small, densely distributed, and frequently overlapping. Their high computational and parametric costs also limit deployment in resource-constrained scenarios. To address these issues, this study proposes a lightweight cherry ripeness detection model that integrates a ConvNeXtv2 backbone *(Xu et al., 2024)*, a DCNv4 convolution module *(Han et al., 2025)*, and a CBAM attention mechanism *(Hi et al., 2023)*. The proposed system not only delivers strong recognition and localization performance but also enables accurate maturity classification under challenging conditions such as leafy occlusion and fruit overlap—all while maintaining high inference efficiency. This research is expected to offer valuable technical support for automated cherry harvesting, yield prediction, and intelligent quality grading, thereby contributing to the transition toward smart and precision-oriented cherry cultivation.

MATERIALS AND METHODS

Image Data Acquisition

A custom cherry image dataset was constructed by collecting photographs at Juxin Cherry Farm in Taigu, Jinzhong, Shanxi Province, China. In collaboration with horticultural specialists, healthy 'Sam' cultivar cherries were selected for imaging. A Huawei Mate 60 smartphone was used to photograph fruit of various sizes and shapes. The dataset comprises 1,085 JPG images, each with a resolution of 4,096 × 3,072 pixels. To capture ecological and morphological diversity, images were taken under a range of conditions, including isolated and grouped cherries, direct sunlight, backlighting, and partial occlusion by branches or leaves. Figure 1 presents examples that illustrate the variety of conditions represented in the dataset. This variation facilitates robust training of visual recognition models for accurate cherry detection in diverse real-world environments.



(a) Single target (b) Single target, occlusion (c) Multiple targets, occlusion

(d) Multiple objects, occlusion, backlighting (e) Multiple (small) targets, occlusion

Fig. 1 - Images of cherries in different scenarios

Image Preprocessing

The imgaug library was used to perform random combinations of cropping, rotation, flipping, scaling, and translation to increase the diversity of the training samples, ultimately generating 6,579 enhanced images.

3,255 of these images were randomly selected and divided into ten equal parts, with a ratio of 8:2 between the training and validation sets. Cherry fruit maturity was categorized into four categories: unripe (green fruit), semi-ripe (green with red), ripe (red or purple), and overripe (shriveled and gray). The data were annotated using I, S, M, and L to represent the four stages of cherry: unripe, semi-ripe, ripe, and lesion, respectively, as shown in Figure 2.

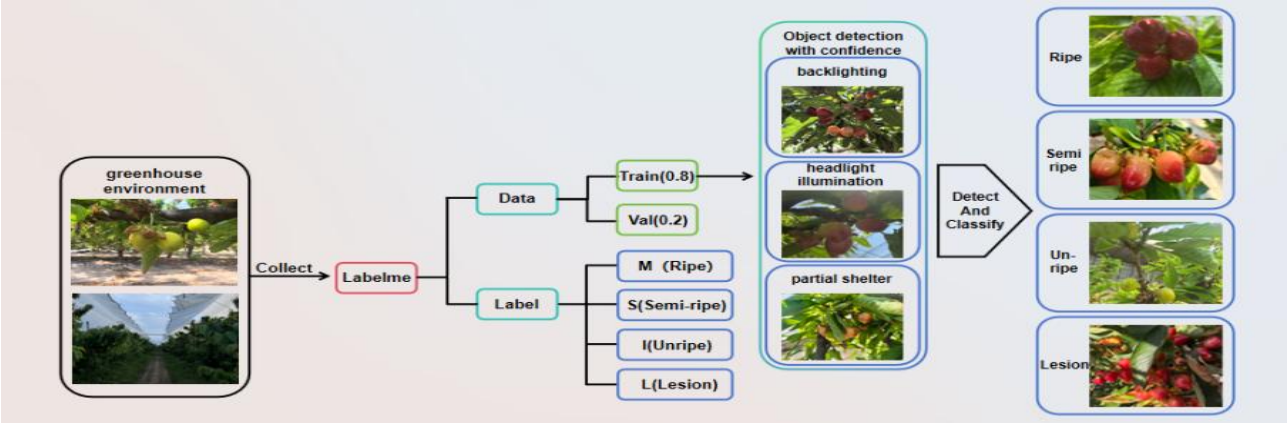


Fig. 2 - Image Processing Process

YOLOv11n model architecture

YOLOv11n is a next-generation object detection model based on the YOLO family of architectures, developed by a cutting-edge research team. YOLOv11n introduces depth-wise separable convolutions. The interaction Mechanism with Multi-Scale Features, Hierarchical Feature Fusion Network, and a joint optimization strategy of dynamic convolution and adaptive feature selection significantly improve the model's detection robustness and efficiency in complex scenarios *(Chen et al., 2020)*. Considering the accuracy and speed requirements for cherry detection in real-world scenarios, YOLOv8n, a derivative of YOLOv11n, was selected as the baseline model for improvement.

Improvements to the YOLOv11n Model

First, the backbone network of the YOLOv11n model was optimized by replacing the original backbone structure with ConvNeXtv2. ConvNeXtv2 cleverly combines the local feature extraction capabilities of convolutional neural networks (CNNs) with the global context modeling advantages of transformers, significantly improving the model's feature representation capabilities and computational efficiency, thereby enhancing cherry detection accuracy. Second, the DCNv4 convolutional module was incorporated into the model backbone. By dynamically adjusting the receptive field, it accurately captures the features of fruits of varying shapes, occlusions, and scales, significantly improving detection accuracy and robustness in complex environments. To enhance the performance of the cherry maturity detection model in complex agricultural environments, this study introduced a convolutional block attention module (CBAM), further enhancing the detection accuracy, as shown in Figure 3.

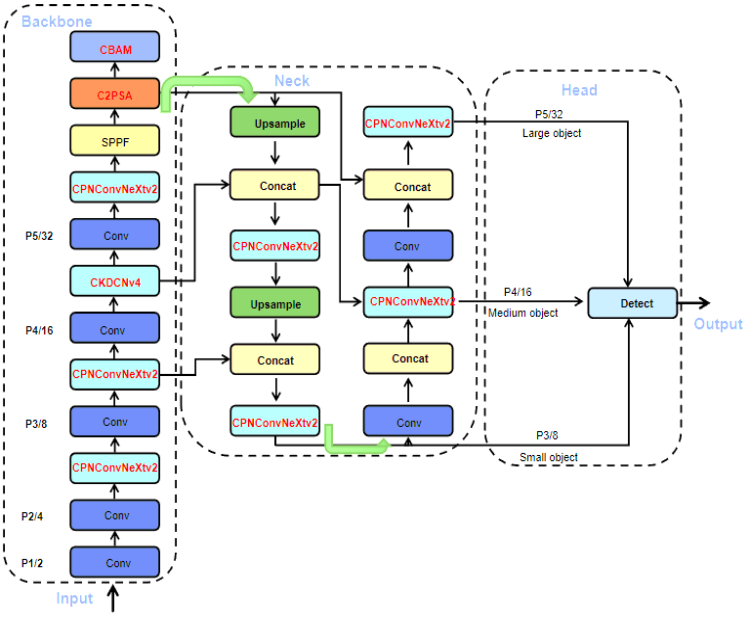


Fig. 3 - M-YOLOv11n Model Structure

ConvNeXtv2

ConvNeXtv2 comprises a unified macro-architecture with a fully convolutional block as its core module. It achieves feature recalibration through the synergy of global response normalization (GRN) and channel-wise FFN. Compared to traditional convolutional neural networks (CNNs), ConvNeXtv2 incorporates a GRN mechanism into the feature feedforward process, which not only enhances the feature selectivity and representation stability of the model but also promotes a diversified representation of features across different channels *(Fu et al., 2025)*. This enhances the model's generalization and performance. To further tap the model's potential, ConvNeXtv2 embeds the GRN into the identity branch of the FFN, effectively fusing and enhancing the convolutional features with global statistical information *(Ma et al., 2025)*. Finally, through convolutional layers and skip connections, the features are transformed and transferred, resulting in a highly expressive and robust feature representation, as shown in Figure 4.

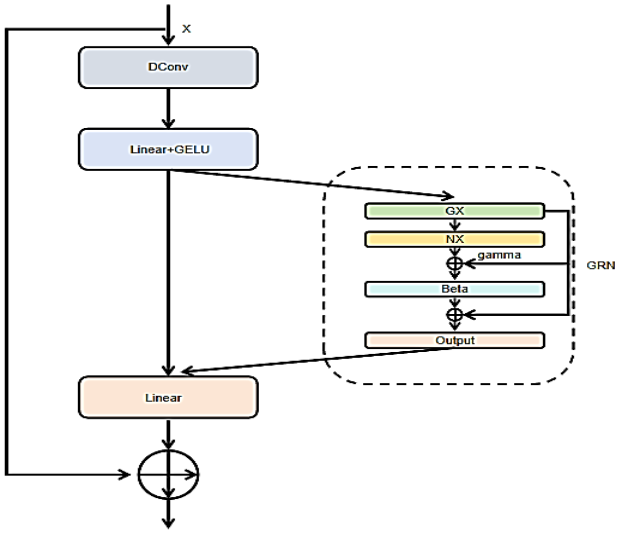


Fig. 4 - ConvNeXtv2 Network Structure Diagram

After the cherry feature map X enters the ConvNeXtV2 module, it undergoes a 7×7 depthwise convolution to extract deep semantic information from the images. Grouped convolution is also used to ensure that each input channel uses only its own convolution kernel, thereby reducing computational complexity and parameter requirements. The extracted feature map is then fed into a Layer Normalization (LN) layer for normalization and a nonlinear transformation using the GELU activation function. To further enhance the model's feature representation capabilities, a Global Response Normalization (GRN) layer was introduced. The GRN layer effectively enhances competition between feature channels through global feature aggregation (GX), feature normalization (NX), and feature calibration operations, thereby improving the model's expressiveness. The execution process of GX is illustrated in Equation (1), which indicates that each channel feature is spatially aggregated in the height H and width W dimensions using the L2 norm.

$$G(X)_i = ||X_i||_2 = \sqrt{\sum_{h=0}^{H-1}\sum_{w=0}^{W-1} X_i(h,w)^2} \quad (1)$$

Finally, the feature map processed by the GRN layer was input into the linear layer and added to the input feature map X to enhance the robustness of the YOLOv11n model. The calculation formula is given by Equation (2).

$$N(G(X)_i) = \frac{||X_i||_2}{\sum_{j=1}^{C} ||X_j||_2} \quad (2)$$

YOLOv11n's C3k2 module was replaced with the ConvNeXtV2 module, which significantly enhanced the model's ability to extract cherry features while reducing the model complexity. The introduction of this module enables the model to demonstrate excellent performance in complex environments, accurately and efficiently identifying cherries of varying ripeness, and reducing the environmental requirements for system deployment.

DCNv4

DCNv4, a next-generation deformable convolution operator, is built on a core architecture that combines dynamic sparse sampling and hardware-aware optimization. Compared with traditional static convolution kernels, DCNv4 achieves adaptive receptive field adjustment through predicted offsets.

This not only enables the accurate capture of irregular objects and long-range dependent features but also significantly improves computational efficiency through kernel-level rewriting and memory optimization. This significantly reduces computational latency and memory usage while enhancing the geometric modeling capabilities of the model *(Liu et al., 2025)*. To fully exploit the potential of hardware computing power, DCNv4 employs a tiling implementation and gradient-aware weight distribution mechanism to achieve efficient alignment and fusion of sample points and feature maps. Ultimately, through sub-thread parallelization and CUDA graph optimization techniques, it achieves a near-linear speedup on modern GPU architectures, becoming a next-generation fundamental computing unit to replace standard convolution and self-attention, as shown in Figure 5.

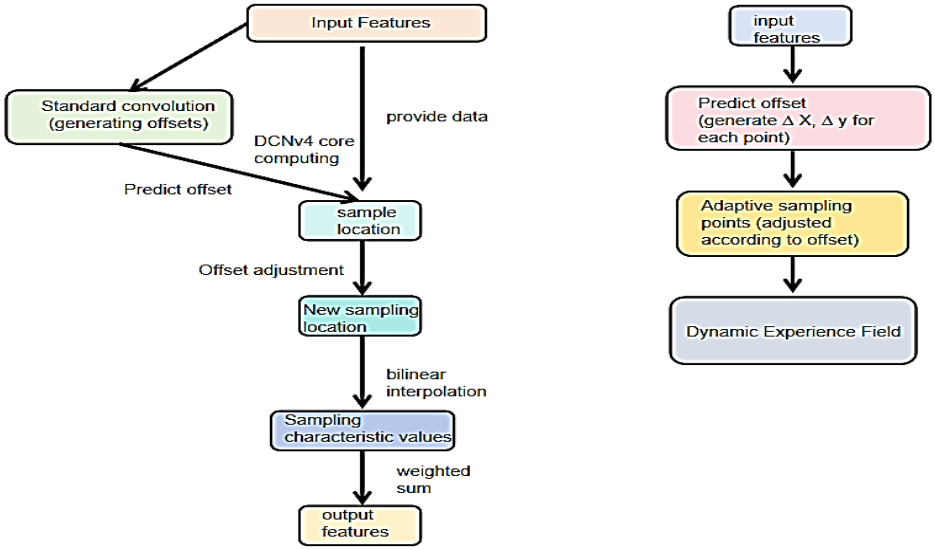


Fig. 5 - DCNv4 Network Structure Diagram

The core operator of DCNv4 is an advanced deformable convolution unit. For any position p on the output feature map, its output value is transformed and calculated using the following formula (3):

$$y(p) = \sum_{k=1}^{k} w_k \cdot x(p + p_k + \Delta p_k) \quad (3)$$

where the notations are defined as below:

$y(p) \in R^{C_{out}}$ represents the output eigenvector at position p; $x(\ ) \in R^{C_{in}}$ is the input feature map, which samples the input features at a specified coordinate (possibly non integer); K is the total number of sampling points; pk is a predefined fixed offset used to determine the sampling grid of the regular convolution kernel; $\Delta p_k \in R^2$ is a dynamic offset predicted by a lightweight quantum network (usually a lightweight convolutional layer) based on input features, allowing the model to adaptively adjust the position of each sampling point according to the input content, thereby focusing the receptive field on areas with richer information; $w_k \in R^{C_{out} \times R^{C_{in}}}$ is the learnable weight matrix corresponding to the k-th sampling position, which is shared among different spatial positions and consistent with standard convolution.

CBAM Attention Mechanism

The attention mechanism selectively ignores invalid information in the image, focusing on valid information and reducing resource consumption in invalid areas. This improves network utilization and enhances object detection capabilities. Therefore, the CBAM attention mechanism was integrated into the feature extraction network, combining the channel and spatial attention mechanisms to form a simple yet effective attention module, as shown in Figure 6.

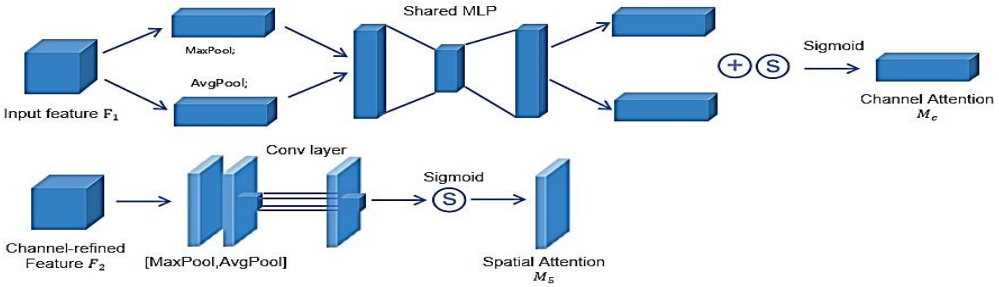


Fig. 6 - CBAM Attention Mechanism Structure Diagram

In the channel attention module, global average pooling and max pooling are applied to the same input feature space to extract the spatial information from the feature map. The obtained feature space information is then input into the next-layer multi-layer perception mechanism module for dimensionality reduction and increase. The weights of the two shared convolutional layers in the multilayer perception network were shared. The features output by the perception network were then added and processed using the sigmoid activation function to obtain channel attention. The calculation formula is given by Equation (4).

$$\mathrm{Mc}(\mathrm{F}) = \varepsilon\left[\mathrm{MLP}\left(\mathrm{F}_{\mathrm{avg}}^{\mathrm{c}}\right) + \mathrm{MLP}\left(\mathrm{F}_{\mathrm{max}}^{\mathrm{c}}\right)\right] \quad (4)$$

where Mc is the channel attention module calculation factor, ε is the sigmoid activation function, MLP is the multilayer perceptron, and F is the feature vector. Spatial attention features complement the channel attention and reflect the importance of the input value in the spatial dimension. The calculation formula is given by Equation (5). First, global average pooling and global maximum pooling are performed on the channel dimension of the feature map. The two features were then concatenated. Finally, the sigmoid function was used to reduce the dimension to 1 × 7 × 7 convolution. A spatial attention feature map was generated after processing the channels. The calculation formula is given by Equation (5).

$$\mathrm{Ms}(\mathrm{F}) = \varepsilon\left\{\mathrm{conv}_{7\times7}\left[\mathrm{unit}\left(\mathrm{F}_{\mathrm{avg}}^{\mathrm{s}},\ \mathrm{F}_{\mathrm{max}}^{\mathrm{s}}\right)\right]\right\} \quad (5)$$

where: Ms is the spatial attention module calculation factor, ε is the sigmoid activation function, MLP is the multi-layer perceptron, F is the feature vector, unit is the channel combination, and conv represents the convolution operation.

To facilitate the use of pre-trained models in the experiment, the CBAM was not embedded in all convolutional residual blocks. It only takes effect after the different convolutional layers.

RESULTS

Parameter Configuration and Evaluation Indicators

The experimental hardware configuration consisted of a GeForce RTX 4060 D GPU, an Intel(R) Xeon(R) Platinum 8270 CPU @ 2.70 GHz (2 processors), and 128GB of RAM. The software configuration comprised Windows 11 Professional Workstation Edition, Python 3.11.9, and CUDA 12.0. YOLOv11 and its improved versions are run on the PyTorch deep learning framework.

The model training parameters were as follows: input image resolution (image_size) of 640 × 640 pixels, initial learning rate (learning_rate) set to 0.01, batch size (batch_size) set to 64, and number of epochs (epochs) set to 500.

This study evaluated the model's performance using accuracy (Precision, P), mean average precision (mAP), model parameter size (parameters), and model computational effort (GFLOPS).

The calculation formula for each metric is as follows.

$$\begin{aligned} \mathrm{P} &= \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}} \\ \mathrm{R} &= \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \\ AP &= \int_0^1 \mathrm{P}(\mathrm{R})\mathrm{dr} \\ \mathrm{mAP} &= \frac{1}{\mathrm{n}}\sum_{\mathrm{i}=1}^{\mathrm{n}} \mathrm{AP}_{\mathrm{i}} \end{aligned} \quad (6)$$

where TP is the number of correctly detected cherries, FP is the number of background impurities that are incorrectly detected as cherries, FN is the number of cherries identified as impurities, R is the recall value at the current accuracy, which refers to the proportion of cherries detected by the model to all actual cherries, and the area under the PR curve drawn with precision P and recall R represents the average precision AP value of the category. The higher the AP value, the better the detection performance of the algorithm.

Backbone Network Comparison Test

To evaluate the advantages of using ConvNeXtv2 as a backbone alternative, three lightweight models (C2f_RepLKBlock, C2f_QARep, C2f_LSK, and ConvNeXtv2) were selected as the backbone networks of the original YOLOv11n model for comparative experiments. The experimental results are presented in Table 1.

Table 1

Comparison Test of Different Backbone Networks

Model	Parameters/k	R/%	P/%	mAP50/%
YOLOv11n	2590620	76.44	87.05	83.96
YOLOv11n+C2f_RepLKBlock	2644044	75.70	84.80	82.80
YOLOv11n+C2f_QARep	2629964	74.20	88.60	82.80
YOLOv11n+C2f_LSK	2699940	72.20	82.30	77.40
YOLOv11n+ConvNeXtv2	2539340	81.70	88.90	88.89

The experimental data in Table 1 indicate that replacing the backbone network of YOLOv11n has a significant impact on the model's performance. Among the many improved solutions, ConvNeXtv2 performed the best, significantly improving the overall detection performance of the model while maintaining a minimum number of parameters. Compared to the YOLOv11n baseline model, ConvNeXtv2 achieved a significant improvement of 4.93 pp in mAP50 (from 83.96% to 88.89%), while reducing the number of parameters by 2.02%. Furthermore, the recall (R) increased significantly by 5.26 percentage points, and the precision (P) increased by 1.85 percentage points, surpassing all other compared solutions. In summary, ConvNeXtv2's unique architectural design achieves dual optimization of parameter count and precision, maintaining high accuracy while reducing model complexity, making it ideal for deployment on mobile and edge computing devices.

Comparison Experiment of Different Convolutional Networks

Analysis of the experimental data showed that introducing different convolutional operations to the ConvNeXtv2 backbone network had a significant impact on the model performance. Among the many improved solutions, the ConvNeXtv2 + DCNv4 combination performed the best, achieving a comprehensive improvement in the detection performance with only a 0.89% increase in the parameters. Compared to the baseline ConvNeXtv2 model, the DCNv4 version improved mAP50 by 0.16 percentage points (reaching 89.05%), while also increasing recall (R) by 0.67 percentage points and precision (P) by a significant 2.95 percentage points, demonstrating an optimal precision-efficiency balance. Notably, this solution achieved a precision of 90.80% while maintaining a high recall, demonstrating its exceptional ability to reduce false detections. In summary, DCNv4, with its dynamic receptive field and hardware optimization features, complements ConvNeXtv2 well, achieving overall performance improvements with a slight increase in the number of parameters, thereby making it the most effective performance enhancement for ConvNeXtv2. Although ShiftConv performed well for some metrics, its overall stability was inferior to that of the DCNv4 solution. The remaining convolutional schemes failed to surpass the overall performance of DCNv4, either because of limited improvement or compatibility issues. The experimental results are presented in Table 2.

Table 2

Comparison Test of Different Convolutional Networks

Model	Parameters/k	R/%	P/%	mAP50/%
YOLOv11n+ConvNeXtv2+ARconv	2554893	81.20	87.85	87.53
YOLOv11n+ConvNeXtv2+DCNv4	2578076	81.87	90.80	89.05
YOLOv11n+ConvNeXtv2+shiftconv	2552940	83.92	92.10	88.37
YOLOv11n+ConvNeXtv2+pinwheelconv	2542716	79.90	85.50	86.50
YOLOv11n+ConvNeXtv2+scconv	4370636	77.60	87.30	85.70

Comparison Experiment of Attention Mechanisms

To further enhance the robustness of the model, the attention modules CBAM, EMA, SimAM, SA, and SK were introduced into the model. As shown in Table 3, the SK attention module significantly increases the model parameter size to 13.67 MB, a 430% increase, which severely impacts the model's parameter efficiency. The SimAM attention module maintained a parameter size similar to the baseline but failed to improve performance. The parameter size increase for the remaining modules was maintained within 2%, demonstrating good parameter efficiency. The comprehensive mAP50 metric showed that the CBAM attention module led with a score of 88.80%, followed closely by the SA attention module with a score of 88.51%.

Both modules achieved a good balance between performance improvement and parameter efficiency. Although the EMA and SK attention modules excel in certain individual metrics, their overall performance does not significantly surpass the baseline.

The CBAM attention mechanism effectively addresses the primary challenges encountered in cherry maturity detection within complex agricultural environments through its dual-dimensional feature optimization, illumination adaptability, and nuanced feature resolution capabilities, thereby providing an efficient and reliable solution for agricultural visual inspection tasks. The experimental results are presented in Table 3.

Table 3

Comparison Test of Different Attention Mechanisms

Model	Parameters/k	R/%	P/%	mAP50/%
YOLOv11n+ConvNeXtv2+DCNv4+CBAM	2622178	82.10	88.30	88.80
YOLOv11n+ConvNeXtv2+DCNv4+EMA	2736636	90.51	79.25	87.89
YOLOv11n+ConvNeXtv2+DCNv4+SimAM	2578076	81.19	87.75	87.04
YOLOv11n+ConvNeXtv2+DCNv4+SA	2572476	80.93	89.94	88.51
YOLOv11n+ConvNeXtv2+DCNv4+SK	13672508	79.31	91.93	87.38

Comparison of Different Models

To further verify the superiority of the proposed M-YOLOv11 model (i.e., the combined model of YOLOv11n, ConvNeXt, DCNv4, and CBAM) for detecting cherries of varying ripeness, the improved model was compared with YOLOv5n, YOLOv8n, YOLOv11n, and Faster R-CNN models. The final comparison results are presented in Table 4. The experimental results show that the accuracy, recall, and mAP50 values for each model were relatively close. The improved model, M-YOLOv11, achieved an accuracy of 88.30% while maintaining a reasonable number of parameters. Compared to the original model, the improved model, M-YOLOv11n, increases precision by 1.25 percentage points, recall by 5.66 percentage points, and mAP by 4.84 percentage points, while also incurring a 1.21% increase in parameters. Compared with YOLOv5n, YOLOv8n, YOLOv11n, and Faster R-CNN, M-YOLOv11n achieved the best combined precision, recall, and mAP, with a minimal increase in the parameters. The experimental results are presented in Table 4.

Table 4

Comparison Test of Different Models

Model	Parameters/k	R/%	P/%	mAP50/%
YOLOv5n	2182444	77.50	88.10	84.44
YOLOv8n	2685148	75.60	87.50	83.40
YOLOv11n	2590620	76.44	87.05	83.96
Faster R-CNN	138357544	87.13	62.38	85.25
M-YOLOv11n	2622178	82.10	88.30	88.80

Cherry Ripeness Detection System

In this study, a cherry maturity detection system was developed based on M-YOLOv11n. The main functional interface of the system is the cherry maturity detection interface.

Cherry Ripeness Detection

The purpose of this interface is to provide a convenient tool for effectively detecting cherry ripeness in preparation for the construction of a cherry-picking robot. Users can use the interface to select locally stored cherry images or video files and upload them to the system, or take photos on site and upload them for automatic recognition. The system utilizes deep learning algorithms to automatically identify and label cherries in images, displaying the labeled images in real-time on the interface. Users can also choose to save labeled images locally as shown in Figure 7.

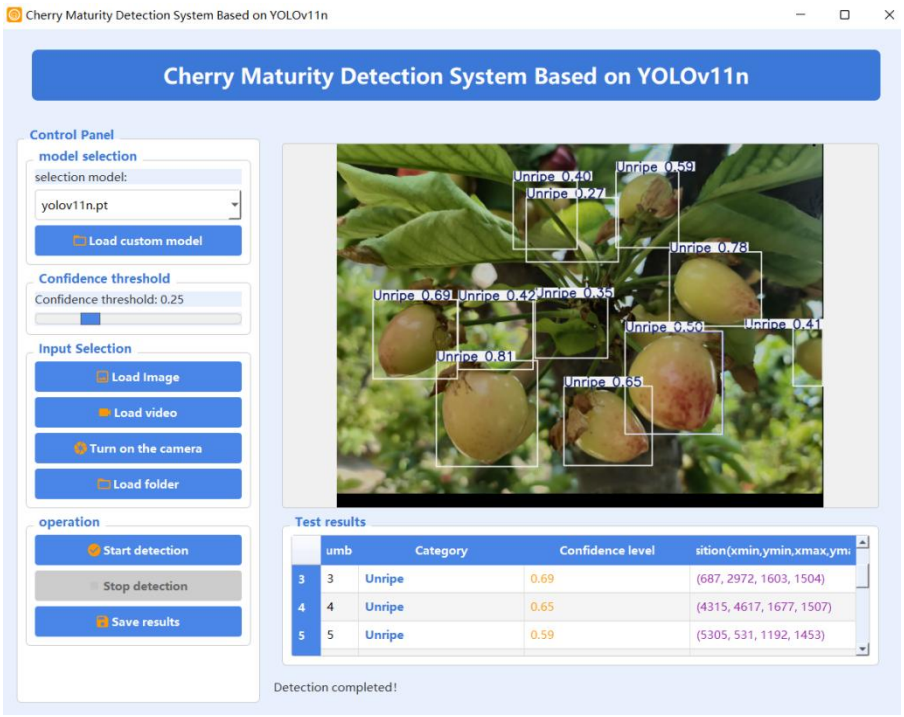


Fig. 7 - System Detection Interface

CONCLUSIONS

This study proposes a lightweight ConvNeXtv2 network to replace the backbone network, introduces DCNv4, and incorporates an attention mechanism. Comparative experiments verified that the improved model, M-YOLOv11n, achieved a 4.84 percentage point increase in mAP@50, 1.25 percentage points in precision, and 0.4 percentage points in recall, compared with the original YOLOv11n model, while maintaining high detection accuracy. The improved model had an mAP50 value of 88.80%, a P value of 88.30%, and an R value of 82.10%. Compared with the mainstream target detection networks, the M-YOLOv11n model proposed in this study has certain advantages in terms of detection accuracy and model lightweight in complex environments.

This research enables the non-contact, high-precision, real-time detection of cherry maturity, a key step in the full mechanization of post-harvest agricultural product processing. Automated detection systems can effectively replace repetitive manual labor, overcome the limitations of manual sorting, such as low efficiency, inconsistent standards, and susceptibility to fatigue, and provide core technical support for building smart orchard production management systems.

Future work will focus on exploring more advanced lightweight techniques and optimization strategies to further improve the model's detection performance and practicality, thereby promoting the continued development and application of intelligent technologies in agriculture.

REFERENCES

[1] Albarrak, K., Gulzar, Y., Hamid, Y., Mehmood, A., & Soomro, A. B. (2022). A Deep Learning-Based Model for Date Fruit Classification. *Sustainability*, Vol. 14, pp. 6339, Saudi Arabia.

[2] Chen S, Xiong J, Jiao J (2022). Citrus fruits maturity detection in natural environments based on convolutional neural networks and visual saliency map (基于卷积神经网络和视觉显著性图谱的自然环境中柑橘类水果成熟度检测). *Precision Agriculture*, Vol. 23, pp. 1515-1531. Germany.

[3] Chen Y, Dai X, Liu M (2020). Dynamic Convolution: Attention over Convolution Kernels (动态卷积：对卷积核的关注). *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp.11030-11039. America.

[4] Danh Phuoc Huynh, My Van Vo, Nghin Van Dang (2021). Classifying maturity of cherry tomatoes using deep transfer learning techiques. *Materials Science and Engineering, IOP Conference Series:* Vol. 1109, pp. 12058. Manila/ Philippines.

[5] Fu Jiarui, Li Zhaofei, Zhou Hao & Huang Wei. (2025). Detection of Camouflage Targets Using Convnextv2 and Texture Edge Guidance (基于 Convnextv2 与纹理边缘引导的伪装目标检测). *Journal of Zhejiang University (Engineering Edition)*, vol. 59, pp. 1718-1726. Zhejiang/China.

[6] Gai, R., Chen, N. & Yuan, H. (2023). A detection algorithm for cherry fruits based on the improved YOLO-v4 model (一种基于改进 YOLO-v4 模型的樱桃果实检测算法). *Neural Comput & Applic*, Vol. 35, pp. 13895-13906. Britain.

[7] Gao Lipeng, Zhou Mengran, Hu Feng (2024). A REIW-YOLOv10n-based small object detection algorithm for underground safety helmets (基于 REIW-YOLOv10n 的井下安全帽小目标检测算法). *Coal Science and Technology*, pp. 1-13. BeiJing/China.

[8] H. Xin and L. Li (2023). Arbitrary Style Transfer with Fused Convolutional Block Attention Modules (使用融合卷积块注意力模块的任意风格迁移). *IEEE Access*, vol. 11, pp. 44977-44988. America.

[9] Han Ruisong, Wang Bin (2025). Filter bi-level routing attention recurrent vision transformers for object detection with event cameras (基于双级路由注意力循环视觉变换器的事件相机目标检测滤波器). *Journal of Electronic Imaging*, Vol. 34. America.

[10] Li Lin, Jin Zhixin, Yu Xiaolei (2024). Optimal YOLOv9 road vehicle and pedestrian detection via Haar wavelet subsampling (Haar 小波下采样优化 YOLOv9 的道路车辆和行人检测). *Computer Engineering and Applications*, Vol. 60, pp. 207-214. BeiJing/China.

[11] Li Ying, Liu Menglian, He Zifen (2024). Citrus fruit maturity detection based on improved YOLOv8s (基于改进 YOLOv8s 的柑橘果实成熟度检测). *Transactions of the Chinese Society of Agricultural Engineering*, Vol. 40, pp. 157-164. BeiJing/China.

[12] Liang Ao, Dai Dongnan, Niu Siqi (2024). Real-time detection algorithm of strawberry maturity based on improved YOLOv5s (基于改进 YOLOv5s 的草莓成熟度实时检测算法). *Shandong Agricultural Science*, Vol. 56, pp. 156-163. Shandong/China.

[13] Liu Hongzhi, Ma Yue, Qiu Bin (2025). Research on Improving the Application of YOLOv11n for Detecting Major Defects in Transmission Line Insulators (改进 YOLOv11n 在输电线路绝缘子主要缺陷检测中的应用研究). *High Voltage Apparatus*, vol. 61, pp. 149-158. ShanXi/China.

[14] Ma Xubang, Wu Xuan Yu, Hu Bingtao (2025). Real-time lightweight defect detection model for aviation carbon fiber components based on multi-dimensional collaborative attention mechanism (基于多维协同注意力机制的航空碳纤维构件缺陷轻量化实时检测模型). *Computer Integrated Manufacturing Systems*, pp. 1-20. BeiJing/China.

[15] Maceachern C B, Esau T J, Schumann A W (2023). Detection of fruit maturity stage and yield estimation in wild blueberry using deep learning convolutional neural networks. *Smart Agricultural Technology*, Vol. 3, pp. 100099. Holland.

[16] Mim F S，Galib S M，Hasan M F. (2018). Automatic detection of mango ripening stages: an application of information technology to botany. *Scientia Horticulturae*, Vol. 237, pp. 156-163, Bangladesh.

[17] Tian Youwen, Qin Shangsheng, Yan Yubo (2024). Blueberry maturity detection in a complex field environment based on improved YOLOv8 (基于改进的 YOLOv8 的复杂野外环境下蓝莓成熟度检测). *Transactions of the Chinese Society of Agricultural Engineering*, Vol. 15, pp. 153-162. BeiJing/China.

[18] Wang F (2025). Improving YOLOv11 for marine water quality monitoring and pollution source identification (改进 YOLOv11 用于海洋水质监测与污染源识别). *Scientific Reports*, Vol. 15, pp. 21367-21367. Britain.

[19] Wang Jianhui, Liu Zhihan, Liu Dayu (2025). Research on China's cherry production and development strategies (中国樱桃生产与发展战略研究). *Sichuan Agricultural Science and Technology*, Vol. 05, pp. 122-124+132 Sichuan/China.

[20] Wang Lishu, Qin Mingxia, Lei Jieya (2021). Blueberry maturity recognition method based on improved YOLOv4 Tiny (基于改进 YOLOv4-Tiny 的蓝莓成熟度识别方法). *Transactions of the Chinese Society of Agricultural Engineering*, Vol. 37, pp. 170-178. BeiJing / China.

[21] Wu Jinlong, Miao Ronghui (2024). Cherry segmentation and identification based on DEEPLABV3 in complex orchard environment (基于 DEEPLABV3 的复杂果园环境下樱桃分割与识别). *INMATEH - Agricultural Engineering*, Vol. 72, pp. 689-698. Romania.

[22] Xu, Y., Li, J., Zhang, L., Liu, H., & Zhang, F. (2024). CNTCB-YOLOv7: An Effective Forest Fire Detection Model Based on ConvNeXtV2 and CBAM (CNTCB-YOLOv7：一种基于 ConvNeXtV2 与 CBAM 的高效森林火灾检测模型). *Fire*, Vol. 7, pp. 54. Switzerland.