agriculture

MDPI

Article

Blueberry Maturity Detection in Natural Orchard Environments Using an Improved YOLOv11n Network

Xinyang Li [1,2], Jinghao Shi [1,2], Yunpeng Li [1,2], Chuang Wang [1,2], Weiqi Sun [1,2], Zonghui Zhuo [1,2], Xin Yue [1,2], Jing Ni [1,2,*] and Kezhu Tan [1,2,*]

[1] Electrical Engineering and Information College, Northeast Agricultural University, Harbin 150030, China; xinyangli@neau.edu.cn (X.L.); s241402027@neau.edu.cn (J.S.); s241402046@neau.edu.cn (Y.L.); s241402038@neau.edu.cn (C.W.); s231401026@neau.edu.cn (W.S.); s231401028@neau.edu.cn (Z.Z.); s241402050@neau.edu.cn (X.Y.)

[2] The National Key Laboratory of Smart Farm Technology and Systems, Northeast Agricultural University, Harbin 150030, China

* Correspondence: nijing@neau.edu.cn (J.N.); kztan@neau.edu.cn (K.T.); Tel.: +86-451-5519-0147 (J.N.); +86-451-5519-0446 or +86-159-0461-9930 (K.T.)

Abstract

To meet the growing demand for automated blueberry harvesting in smart agriculture, this study proposes an improved lightweight detection network, termed M-YOLOv11n, for fast and accurate blueberry maturity detection in complex natural environments. The proposed model enhances feature representation through an improved lightweight multi-scale design, enabling more effective extraction of fruit features under complex orchard conditions. In addition, attention-based feature refinement is incorporated to emphasize discriminative ripeness-related cues while suppressing background interference. These design choices improve robustness to scale variation and occlusion, addressing the limitations of conventional lightweight detectors in detecting small and partially occluded fruits. By incorporating MsBlock and the attention mechanism, M-YOLOv11n achieves improved detection accuracy without significantly increasing computational cost. Experimental results demonstrate that the proposed model attains 97.0% mAP50 on the validation set and maintains robust performance under challenging conditions such as occlusion and varying illumination, achieving 96.5% mAP50. With an inference speed of 176.6 FPS, the model satisfies both accuracy and real-time requirements for blueberry maturity detection. Compared with YOLOv11n, M-YOLOv11n increases the parameter count only marginally from 2.60 M to 2.61 M, while maintaining high inference efficiency. These results indicate that the proposed method is suitable for real-time deployment on embedded vision systems in smart agricultural harvesting robots and supports early yield estimation in complex field environments.

Keywords: blueberry maturity detection; object detection algorithm; depthwise separable lightweight; MsBlock module; Squeeze-and-Excitation module; smart agricultural

Academic Editor: Hyeon Tae Kim

Received: 2 December 2025
Revised: 21 December 2025
Accepted: 25 December 2025
Published: 26 December 2025

1. Introduction

For blueberries, as a fruit with high economic value, the precise detection and classification of their maturity are of great significance for orchard management, automated harvesting, and market grading [1–3]. According to the statistics of the Food and Agriculture Organization of the United Nations (FAO), global blueberry production has increased by nearly 60% over the past decade, with a production of approximately 1.2 million tons

in 2023. China, the United States, and Chile are the main producers [4]. Since the beginning of the 21st century, China has undertaken large-scale cultivation of blueberries and has emerged as a key production region within the Asia-Pacific area. According to recent industry statistics reflected in the cited literature, the total national planting area has reached approximately 1.1 million hectares, with an annual output of up to 500,000 tons. It should be noted that these figures represent general agricultural estimates for the sector as a whole, rather than values tied to specific cultivars or localized markets. While blueberry cultivation carries high agricultural added value, the maturity of the berries significantly affects their taste, nutritional quality, market price, and harvest costs [5,6]. For example, market analyses indicate that, under common commercial grading systems, blueberries classified as maturity level III typically command prices more than 30% higher than those of maturity level I, a trend observed across several major distribution channels [7]. Therefore, the automatic detection and classification of blueberry maturity can not only improve harvesting efficiency but also significantly increase economic benefits [8,9].

The traditional methods for detecting blueberry maturity mostly rely on manual observation and hand-crafted feature extraction [10,11]. However, such approaches are generally labor-intensive, time-consuming, and susceptible to human subjectivity, which makes them difficult to meet the requirements of intelligent and automated agriculture, especially under large-scale production conditions [12]. With the continuous expansion of the blueberry cultivation scale, the limitations of traditional methods have become increasingly prominent. In recent years, the rapid development of computer vision and deep learning technologies has brought about changes in the agricultural field, especially in image-based object detection and classification methods, which have been widely applied [13–16]. However, when deployed in complex natural environments characterized by factors such as occlusion, varying lighting conditions, and significant scale variation among fruits, these models often face challenges in maintaining robust accuracy while meeting the stringent requirements for real-time processing and lightweight deployment essential for practical agricultural applications. To address these specific challenges, this study proposes the M-YOLOv11n network. Building upon the efficient YOLO framework, our work introduces key structural innovations. The novelty of M-YOLOv11n lies in its integrated approach, which strategically incorporates a multi-scale feature extraction module and an attention mechanism within a lightweight architecture. This design aims to enhance feature discrimination for blueberries under challenging conditions without substantially increasing computational cost, thereby advancing the balance between accuracy, robustness, and efficiency for in-field maturity detection. For example, Zhang et al. proposed a fruit detection method based on Faster R-CNN and achieved more than 90% accuracy in the detection task of apples and oranges [13]. However, the high computational complexity of Faster R-CNN makes it perform poorly in real-time detection [17]. In contrast, the YOLO series models have achieved a better balance between real-time performance and accuracy due to their efficient single-stage detection framework, which has been successfully applied in various agricultural detection tasks. For instance, an improved YOLOv8 model has been used for the precise identification and detection of fresh leaves from five different varieties of Yunnan large-leaf tea trees, achieving a mean Average Precision (mAP) of 94.8% [18]. The YOLOv7 framework further demonstrates the series' capability by establishing a new state-of-the-art for real-time general object detectors through its novel trainable optimization strategies [19]. Similarly, a lightweight improved YOLOv4-Tiny network has shown high effectiveness in recognizing blueberry fruits and determining their maturity levels in natural environments, with a detection speed as fast as 5.7 milliseconds per image [20]. Wang et al. proposed a transformer-based grape maturity detection method in Computers and Electronics in Agriculture, but the model complexity limits the practical application

scenarios [14]. At present, certain progress has been made in the research on precise fruit identification and fruit maturity classification, both at home and abroad. In order to detect round-like fruits, Li Ying et al. proposed an improved citrus fruit maturity detection method based on the YOLOv8s model [21]. The mean Average Precision (mAP) mAP(0.5) of the improved YOLOv8s model on the test set is 95.6%. However, the method still has the problem of missed detection due to the occlusion of overlapping fruits in the detection of citrus fruit maturity. Chen Fengjun et al. addressed the issue that *Camellia oleifera* fruits are often obscured in the natural environment [22]. Based on the original YOLOv7 model, they improved and proposed a method for detecting the maturity of *Camellia oleifera* fruits. The mean Average Precision (mAP) mAP of the improved YOLOv7 model under the test set was 94.60%. However, there are still some problems in this method for the detection of camellia fruit maturity, such as missed detection and false detection, and it is not easy to deploy to mobile devices. However, for blueberries as a specific crop, existing detection networks still face significant challenges in complex actual orchard environments: First, blueberry fruits are small in size and densely clustered, making them highly prone to missing detections due to overlapping and occlusion by branches and leaves; Second, the variable lighting conditions in orchards (e.g., backlighting, shadows) severely affect the robustness of color-dependent ripeness discrimination; Third, most improved models increase computational complexity to enhance precision, making it difficult for real-time deployment on edge devices with limited computing resources such as picking robots. Therefore, developing a blueberry detection model with high precision, strong robustness, and lightweight characteristics in complex environments is crucial for realizing automated harvesting [23–27].

To address the above issues, this study proposes a blueberry maturity detection and classification method based on the improved YOLOv11n model. By introducing Multi-Scale Block (MsBlock) with a depth-separable lightweight component into the Backbone and introducing the SE attention mechanism into the feature pyramid, this study aims to enhance the model's feature extraction capability for targets of different scales, thereby improving detection accuracy and robustness. Multi-scale feature extraction has been proven to have significant advantages in object detection tasks. For example, the Feature Pyramid Network (FPN) proposed by He et al. significantly improves the detection ability of small targets by fusing features at different levels [28]. Similarly, the Multi-Scale attention mechanism proposed by Chen et al. further enhances the robustness of the model in complex scenarios by dynamically adjusting the feature weights [29].

Beyond blueberries, imaging-based maturity and quality assessment combined with advanced machine-learning techniques has been increasingly explored across a wide range of agricultural and horticultural crops. Recent studies have demonstrated the integration of imaging and deep learning with quantitative quality and maturity indices in various agricultural and horticultural products, highlighting the potential for cross-crop modeling and transferability of such approaches [30,31]. These cross-crop efforts provide important methodological insights and further support the broader applicability of the proposed framework beyond a single crop species.

Based on various theoretical models and empirical cases mentioned in the previous text, this study makes structural improvements on the basis of the original YOLOv11n backbone network Backbone. Without significantly increasing the memory of the network structure, the Multi-Scale Block (MsBlock) and the SE attention mechanism are introduced. The effect of its application in the detection and recognition of the maturity of blueberry fruit under a natural environment was tested through experiments. This study can provide an important data foundation and recognition basis for subsequent yield estimation, labor

allocation planning, and target locking in automated mechanical harvesting, and exhibits potential for further application in smart agriculture management systems.

In terms of data collection and construction, this study collected blueberry images under natural light conditions on a farm in Florida, USA, and constructed a specialized dataset containing scenes of mild occlusion, severe occlusion, and backlighting. The dataset comprises 876 original images, with 63,728 fruits annotated and categorized into three classes based on maturity. After data augmentation expanded the dataset to 7005 images, it was split into training, validation, and test sets in a ratio of 7:1:2. Regarding experimental design and evaluation, the effectiveness of the MsBlock, SE attention mechanism, and depthwise separable convolution was progressively validated through seven ablation experiments. Comparisons were made with YOLOv8n, SSD-MobileNet, YOLOv11n, and Faster R-CNN across the three test scenarios. All experiments were conducted under consistent software and hardware environments using identical training strategies. Metrics including mAP50, accuracy, recall, F1-score, FPS, and parameter count were employed to comprehensively evaluate model performance.

2. Materials and Methods

2.1. Data Acquisition

The detection of blueberry fruit is based on deep learning algorithms to classify blueberry fruit of different maturity levels in a complex natural environment. Blueberry fruit images in a natural environment are subject to external interferences such as soil, light, and branches and leaves. Blueberries ripen in batches. Usually, each cluster contains 1 to 3 types of mature blueberry, namely ripe fruits, semi-ripe fruits, and unripe fruits. Meanwhile, the unripe fruits are similar in color to the branches and leaves, while the ripe fruits are close in color to the soil.

The collection site of blueberry images in this study was located at Strann Farm in Florida, USA, as shown in Figure 1. The images were collected during the 2022 growing season. The plants belong to a typical early-ripening southern highbush blueberry (*Vaccinium corymbosum*) hybrid, which is widely cultivated in Florida. Detailed records of plant age were not available; however, all sampled plants were mature, commercially productive bushes. Image acquisition was conducted during the main harvest period of the 2022 growing season, covering the typical maturity stages from early to late harvest. All images were collected from a single field site to ensure relatively consistent environmental and management conditions. Under natural illumination conditions, a Canon 200D II DSLR camera (Canon Inc., Tokyo, Japan) with an 18–55 mm lens set to fully automatic mode was used to capture blueberry fruit cluster images at a distance of about 1 m. The camera was operated in fully automatic mode; therefore, exposure and white balance were automatically adjusted by the camera for each image under natural illumination. No color calibration target (e.g., ColorChecker/gray card) was used during data acquisition. A total of 876 original image data were collected, and the images were saved in .jpeg format. The resolution is 3648 × 2736 pixels, which corresponds to the actual scene of about 13 cm × 10 cm, and the compression ratio is 10:1. In recent studies, the image resolution used by Feng et al. for blueberry detection is 1280 × 1562 pixels [27]. In contrast, the resolution of 3648 × 2736 pixels employed in this work belongs to the relatively high-resolution range. The dataset contains blueberry fruit image samples under different degrees of occlusion and illumination, which were categorized based on visual and geometric criteria. Slight occlusion refers to images in which most blueberry fruits are clearly visible, with occlusion affecting only a small portion of the fruit surface (typically less than approximately 30% of the fruit area), often caused by thin branches or leaves. Severe occlusion corresponds to cases where a substantial portion of the fruit is obscured (approximately more than 30% of the visible area), frequently due

to overlapping fruits or dense foliage. Backlight images were defined as samples captured under strong illumination contrast, where the fruit regions appear partially darkened or exhibit reduced color contrast due to direct or near-direct sunlight exposure behind the fruit. Based on these criteria, the dataset was categorized into three groups: images with slight occlusion, images with severe occlusion, and images captured under backlight conditions. The dataset includes 287 slight-occlusion images, each containing 5 to 10 blueberry fruits of different sizes; 391 severe-occlusion images, each containing 10 to 25 blueberry fruits of different sizes; and 198 backlight images, each containing 3 to 15 blueberry fruits of different sizes. The dataset covers various lighting conditions to ensure sufficient diversity.

Strane Farm, Florida A live shot of the farm

Figure 1. The Straughn Farm Experimental Station.

2.2. Data Preprocessing

A total of 876 high-quality blueberry images were obtained through strict screening. Each image contains 5 to 30 blueberry fruits, covering three maturity levels: maturity I (green), maturity II (reddish), and maturity III (purple). The maturity levels adopted in this study were defined based on a visual–physiological correspondence commonly used in blueberry agronomy. Fruit surface color was used as the primary non-destructive indicator of maturity, which has been widely reported to correlate strongly with physicochemical attributes such as soluble solids content, firmness, and anthocyanin accumulation. According to established literature, green berries correspond to early developmental stages with low sugar and anthocyanin levels, reddish berries represent transitional ripening stages, and fully purple berries indicate physiological maturity suitable for harvest. To reduce subjectivity, all blueberry fruits were independently annotated by two annotators with agricultural background. In cases of labeling disagreement, consensus was reached through discussion, and only samples with consistent labels were retained. Although no destructive physicochemical measurements were conducted for each individual fruit, the adopted maturity classification is consistent with commonly accepted standards in blueberry maturity studies. They also cover various actual scenarios such as slight occlusion, severe occlusion, and backlight. The real photos of each scene are shown in Figure 2.

The blueberry fruit in the image were annotated with bounding box using the LabelImg v1.8.6 tool to generate an XML file containing the normalized center (x_c, y_c), width, and height, and then converted into a TXT format annotation file suitable for the YOLO model through a Python script. Training deep neural networks requires a large amount of data. A dataset that is too small can lead to overfitting of the neural network. Therefore, data augmentation needs to be performed on the collected data. In this study, a multi-modal data augmentation strategy was implemented on the training set. Geometric robustness was enhanced through random rotation ($\pm 30^\circ$) and scaling (0.8 to 1.2 times). Brightness, contrast, and saturation adjustments within a range of $\pm 20\%$ were applied to simulate

illumination variations. In addition, to improve the robustness of the model against sensor noise and environmental interference, additive Gaussian noise was introduced during data augmentation. Specifically, after normalizing the image pixel values to the [0, 1] range, zero-mean Gaussian noise with a standard deviation randomly sampled from [0, 0.01] was added identically to all RGB channels of the training images. This noise augmentation was applied with a probability of 0.5 during training. As a result, the size of the dataset was expanded to 7005 images. A total of 63,728 blueberries were annotated in the dataset. Specifically, the dataset comprises 28,966 immature, 15,142 semi-mature, and 19,431 mature fruit instances, as also visualized in Figure 3. The dataset was split into training, validation, and test sets with a ratio of 70%:10%:20%, corresponding to 4904, 696, and 1405 images, respectively. Importantly, the split was performed at the cluster (bush) level, such that all images belonging to the same bush were assigned exclusively to a single subset. This strategy was adopted to avoid spatial correlation and potential data leakage between the training, validation, and test sets, as shown in Table 1.

(**A**) Slight occlusion (**B**) Severe occlusion (**C**) Backlight

Figure 2. Multi-scene real photos of blueberry: (**A**) slight occlusion; (**B**) severe occlusion; (**C**) backlight.

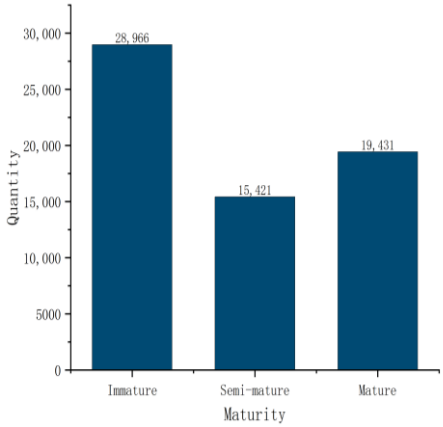


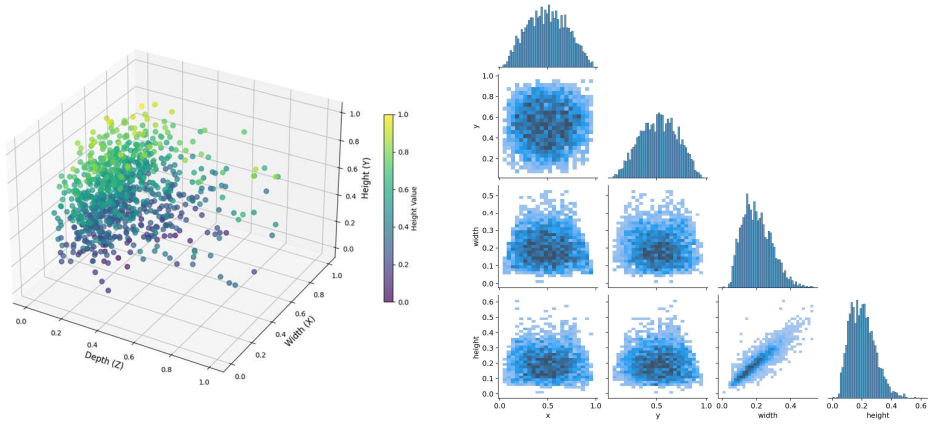
Figure 3. Statistical chart of blueberry quantity at different maturities.

The dataset fully covers realistic variables such as fruit density, light conditions, and background complexity. Its refined annotation and systematic enhancement processing provide a diverse and robust data foundation for model training and validation.

Table 1. Distribution of blueberry datasets after data augmentation.

Dataset	Total Number of Images	Slight Occlusion	Severe Occlusion	Backlight
Training set	4904	1607	2189	1108
Validation set	696	220	312	164
Test set	1405	459	625	321
Total	7005	2286	3126	1593

Figure 4A presents a three-dimensional coordinate system established to describe the spatial distribution and morphological characteristics of blueberry fruits. The X-axis represents height, which is associated with the vertical dimension of the fruits, while the Y-axis and Z-axis both denote distance, reflecting the horizontal spacing between fruits and their distribution along the depth direction, respectively. The construction of this three-dimensional coordinate system provides a quantitative tool for investigating the spatial distribution patterns of blueberry fruits, facilitating the exploration of potential correlations between fruit maturity and spatial positioning. For instance, mature fruits may exhibit a more concentrated vertical distribution owing to increased weight. Although this analysis was not directly utilized in the architectural design of the model in the present study—such as in anchor box size selection or data augmentation range setting—it offers valuable contextual insights into the distribution patterns of blueberry fruits in complex natural environments. Furthermore, it may inform future optimizations in data acquisition and augmentation strategies.



(**A**) (**B**)

Figure 4. (**A**) Three-dimensional coordinate system for the spatial distribution and morphological characteristics of blueberry. (**B**) Blueberry Maturity coordinate analysis diagram.

Figure 4B presents a two-dimensional chart that describes the size characteristics of blueberry fruit and their normalized values, ranging from 0.0 to 1.0, indicating that the data have been normalized to facilitate unified analysis of fruits of different sizes. This graph provides a standardized tool for the study of the morphological characteristics of blueberry fruit through normalized size data, which helps to reveal the quantitative relationship between fruit maturity and size. For example, mature fruits may significantly increase in width and height due to cell expansion. After conducting statistics on the labeled blueberry bounding box data, the results show that the centers (x, y) of most blueberry bounding boxes are distributed throughout the image center and follow a normal distribution pattern. Their aspect ratio is close to 1, indicating that the blueberry is elliptical in shape, and the width and height marked with rectangular boxes are basically the same. This relatively regular annotation box reduces the overfitting of the detection model during the position

prediction process. The combination of the two graphs can provide dual support for spatial and morphological features in the detection of blueberry maturity, enhancing the robustness and generalization ability of the model.

3. Algorithm Design and Experiment

3.1. YOLOv11n Object Detection Network

There are two main categories of object detection methods based on deep learning. The first category is two-stage object detection algorithms based on region proposals, such as R-CNN, Fast R-CNN, and Faster R-CNN. The second category is one-stage object detection methods based on regression, such as YOLO, RetinaNet, and EfficientDet. Since Redmon proposed the first regression-based object detection, YOLOv1 in 2016, it has received extensive attention from researchers [32]. By 2024, the YOLO series network will have been updated to the 11th generation. After verification on standard data sets, YOLOv11 has good performance. However, the detection accuracy in the detection speed and multi-feature environment still cannot meet the real-time requirements, and the network structure occupies a large amount of memory, which makes it difficult to achieve the conditions for deployment on the embedded system carried by the agricultural picking robot.

YOLOv11n is a lightweight object detector with a C3k2 backbone. Compared to C2f, it achieves lightweighting via optimized convolution kernels and group convolution, retaining SiLU activation to balance feature extraction and computational cost [33]. Adopting FPN for multi-scale fusion, it optimizes feature extraction through information flow and cross-scale fusion, focusing on P3/P4/P5 scales. Structural optimization and channel compression reduce computational complexity, meeting real-time multi-scale detection needs in conventional scenarios. In summary, YOLOv11n aims to enhance real-time detection efficiency and reduce computational costs. However, its drawback lies in the fact that it does not introduce advanced feature enhancement modules and attention mechanisms, and thus cannot fully model the importance correlation between feature channels and spatial dimension. Its ability to capture local detail features is relatively weak [34]. When detecting crops, due to factors such as tree branch obstruction and backlight exposure, YOLO11n is prone to problems such as weakened features of small targets and confusion between targets and backgrounds, with obvious missed detection phenomena. There is still considerable room for improvement in detection accuracy in complex environments. Table 2 presents the comparison between YOLOv11n and other existing models.

Table 2. The comparison between YOLOv11n and other existing models.

Model	Model Base	Core Improvements	Target Scenarios	Limitations
MPS-YOLOv7	YOLOv8	Enhanced detail feature + content-aware reassembly	General field conditions	Poor performance under severe occlusion
CES-YOLO	YOLO (optimized)	CES module for edge deployment	Edge device application	Limited multi-scale feature adaptation
YOLO-BLBE	YOLO-BLBE	I-MSRCR image enhancement	Common lighting conditions	Weak anti-interference to backlight
SSC-YOLOv9c	YOLOv9	Improved for dense occlusion	Dense occlusion scenarios	High parameter volume, not lightweight
Original YOLOv11n	YOLOv11n	None (baseline)	Conventional scenarios	Insufficient small-target capture; poor robustness to complex backgrounds
M-YOLOv11n (this study)	YOLOv11n	MsBlock + SE attention + depthwise separable convolution	Slight/severe occlusion + backlight	Slight FPS reduction compared to baseline (negligible for practical use)

In order to further improve the performance and detection accuracy of the object detection network, this study proposes an improved object detection network (M-YOLOv11n) containing a Multi-Scale Block (MsBlock). By introducing Multi-Scale Block (MsBlock) (multi-scale Block, MsBlock) and the SE attention mechanism into the YOLOv11n object detection network, and adopting the hierarchical multi-branch structure and the multi-scale feature extraction strategy based on depthwise separable convolution to capture the feature differences of targets at different scales and integrate local details and global information, the effective transmission of multi-scale features is strengthened, thereby enhancing the extraction of deep information in the network structure.

3.2. Multi-Scale Block

In object detection tasks, the model's ability to capture features of different scales directly determines the detection accuracy of the target in different environments. This feature capture ability is closely related to the receptive field coverage range of the module [35]. In lightweight networks, fixed-scale convolution kernels are often used to control the number of parameters, making it difficult to meet the feature extraction requirements of both the local details of small targets and the global contours of large targets. Especially in natural scenes, detection omissions are prone to occur under different conditions [36].

A Multi-Scale Block (MsBlock) is a feature extraction module composed of multiple parallel convolutional layers, each of which has a different receptive field to capture feature information of different scales [37]. The basic principle of MSBlock can be summarized into the following three core pillars:

In object detection tasks, the model's ability to capture features at different scales is crucial, as it directly affects detection accuracy for targets of varying sizes in complex natural environments. For blueberry fruit detection, the dataset includes both close-range (large targets) and distant (small targets) fruits, often accompanied by occlusion from branches and leaves. To address these multi-scale and partial occlusion challenges, this study designs a Multi-Scale Block (MsBlock), the core of which lies in using a set of parallel convolutional layers to obtain differentiated receptive fields, thereby effectively capturing feature information from local details to global contours.

The structural design of MsBlock directly responds to the characteristics of the blueberry dataset. Among its components, smaller convolution kernels focus on extracting local subtle variations in fruit surface color and texture, which is essential for distinguishing maturity levels; whereas larger convolution kernels help integrate broader contextual information under foliage occlusion to infer the complete contour of the fruit. These multi-scale features are subsequently integrated through weighted fusion, ensuring that the model can simultaneously adapt to blueberry fruits of different sizes and visibility levels in the dataset.

In summary, MsBlock is not a generic multi-scale structure but rather a customized feature extraction scheme designed to address the inherent scale variability and occlusion complexity in field images of blueberries. The selection of the number of branches and the sizes of convolution kernels is based on the specific target size distribution and occlusion situations observed in prior data analysis.

As illustrated in Figure 5, the MS-Block module processes input features by splitting them into multiple parallel branches along the channel dimension. Each branch first performs cross-channel interaction and dimension mapping via a 1×1 convolution, followed by spatial feature extraction using $k \times k$ depthwise convolution. The features are then refined and compressed through another 1×1 convolution before all branches are finally merged via a channel-wise 1×1 convolution to integrate multi-scale information [38]. This

design enhances the model's ability to perceive objects with large-scale variations, making it well-suited for complex detection scenarios.

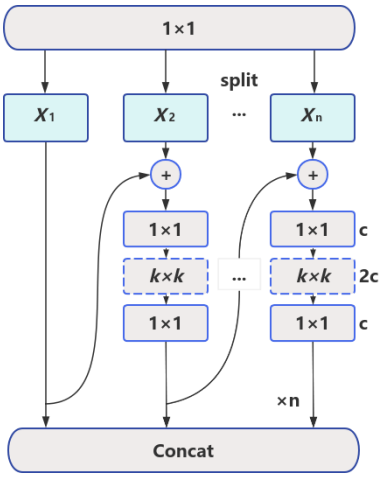


Figure 5. Split–merge multi-branch convolutional module.

In M-YOLOv11n, the MS-Block is integrated into the backbone network (Figure 6), where it extracts both local and global information through multi-scale convolutional kernels. A channel weighting mechanism is applied to adaptively adjust the contribution of features at each scale, thereby improving feature representation. The corresponding calculation is expressed as follows:

$$X' = \sum_{i=1}^{C} \beta_i X_i, \beta_i = \frac{e^{\gamma S_i}}{\sum_{j=1}^{C} e^{\gamma S_j}} \quad (1)$$

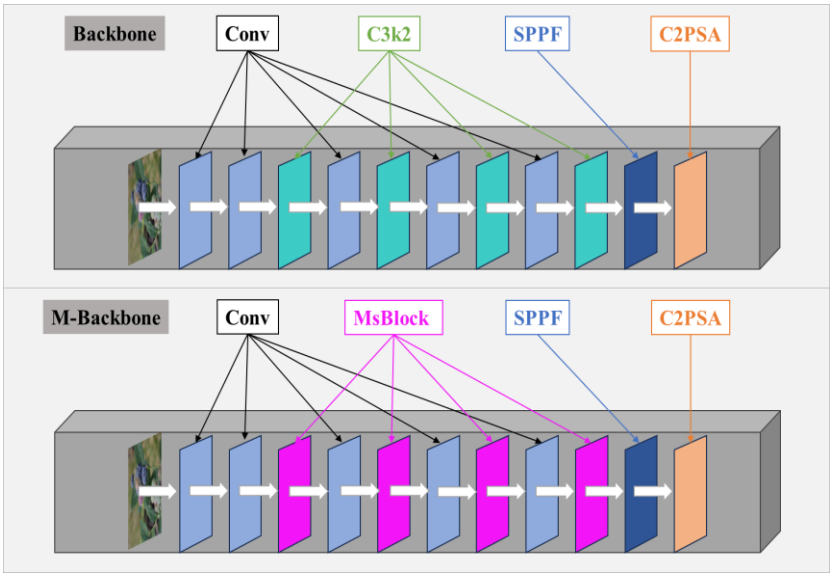


Figure 6. Comparison of Backbone and M-Backbone backbone networks.

Here, X_i is the feature map at the ith scale, β_i is its weighting coefficient, S_i is the channel importance score, and γ is the temperature coefficient, which is used to control the smoothness of the weight distribution. Subsequent experiments have shown that after adding MsBlock to YOLOv11n, the mAP for small object detection has been improved, especially performing better in fine recognition tasks such as fruit ripening detection.

3.3. Depthwise Separable Convolution

Depthwise separable convolution is an efficient convolution structure that is widely used in the design of lightweight neural networks, aiming to significantly reduce the computational complexity and the number of parameters of the model while maintaining a strong feature extraction capability [39]. In this section, the structure principle, computational complexity, and its application in M-YOLOv11n are analyzed in detail.

3.3.1. Structure and Principle

Let H and W denote the spatial dimensions of the input feature map, Cin and Cout denote the input and output channel counts, respectively, and K denotes the convolution kernel size. Computational complexity in the context of convolutional neural networks refers to the number of floating-point operations (FLOPs) required to perform a given layer's computation. The computational complexity of standard convolution ($FLOP_{sstd}$) and depth-separable convolution ($FLOP_{sdsc}$) can be expressed as:

$$FLOPs_{std} = H \times W \times C_{in} \times C_{out} \times K^2 \tag{2}$$

$$FLOPs_{dsc} = FLOPs_{dw} + FLOPs_{pw} = H \times W \times C_{in} \times \left(K^2 + C_{out}\right) \tag{3}$$

Computational Efficiency Comparison: Computational Ratio Between Deep Separable Convolution and Standard Convolution:

$$\frac{FLOPs_{std}}{FLOPs_{std}} = \frac{H \times W \times C_{in} \times \left(K^2 + C_{out}\right)}{H \times W \times C_{in} \times K^2 \times C_{out}} = \frac{1}{C_{out}} + \frac{1}{K^2} \tag{4}$$

Theoretical Engineering Simplification: When the condition $Cout \gg K^2$ is satisfied, the following approximate relationship can be obtained:

$$FLOPs_{dsc} \approx \frac{1}{K^2} \times FLOPs_{std} \tag{5}$$

Taking a typical layer in a backbone network as an example, with parameters C_(in) = 256, C_(out) = 512, and K = 3, standard convolution requires approximately 1.18 GFLOPs, while deep separable convolution requires only about 0.13 GFLOPs—a computational reduction of approximately 88.98%. This significant reduction in computational cost demonstrates that depthwise separable convolution serves as a core component in lightweight network design, enabling substantial computational overhead reduction while maintaining robust feature extraction capabilities.

3.3.2. Application in M-YOLOv11n

To construct a detection model that takes into account both high precision and high efficiency, in this paper, depthwise separable convolution is introduced as the core lightweight component into the M-YOLOv11n network. The core idea of the Multi-Scale Block (MsBlock) proposed in this paper is to enhance the model's performance in processing multi-scale information by improving the size and structure of the convolution kernel and optimizing the feature fusion method, thereby improving the overall object detection accuracy and efficiency [40]. Depthwise separable convolution can effectively achieve hierarchical feature fusion of MsBlock, thereby maintaining the model's strong feature extraction capability while significantly reducing computational complexity and memory usage with depthwise separable convolution.

3.4. Complete Intersection over Union (CIoU) Loss Function

To achieve accurate localization and classification of blueberry fruits in complex natural environments, it is essential to optimize the loss function to balance the training errors associated with bounding box position, confidence, and category. This study adopts the Complete Intersection over Union (CIoU) loss function. Compared to the traditional Intersection over Union (IoU) loss, CIoU mitigates the gradient vanishing problem when the predicted bounding box does not intersect with the ground truth and provides a more comprehensive measure of their overlap [41].

In the context of blueberry detection, where target fruits are often small and frequently exhibit incomplete boundaries due to occlusion by branches and leaves, the CIoU loss offers particular advantages. Beyond merely considering the overlap area, CIoU introduces penalty terms for both the center-point distance and aspect ratio consistency. This design is crucial for bounding box regression under conditions of small targets and partial occlusion. The center-point distance term provides an effective gradient direction even when the overlap between the predicted and ground truth boxes is low, alleviating optimization difficulties caused by small target sizes. Meanwhile, the aspect ratio term constrains the predicted box shape to better conform to the nearly circular appearance of blueberries, thereby enhancing localization stability. Consequently, for blueberry datasets characterized by significant scale variation and frequent occlusion, CIoU contributes to more robust bounding box regression. The formulation of this loss function is given by the following Equation (6).

$$\mathcal{L} = S(\mathrm{E},\mathrm{E}^{\mathrm{gt}}) + D(\mathrm{E},\mathrm{E}^{\mathrm{gt}}) + V(\mathrm{E},\mathrm{E}^{\mathrm{gt}}) \tag{6}$$

In the formula, S, D, and V respectively represent the overlapping area, distance, and aspect ratio, which are two predicted bounding boxes, respectively, $\mathrm{EE}^{\mathrm{gt}}$. However, IoU and GIoU loss only consider the overlapping area, as shown in Equation (7).

$$\mathrm{S} = 1 - \mathrm{IoU} \tag{7}$$

The normalized center point distance is adopted to measure the distance between two predicted bounding boxes, as shown in Equation (8).

$$D = \frac{\rho^2(p, p^{gt})}{c^2} \tag{8}$$

where $\mathrm{p} = [\mathrm{x},\mathrm{y}]^{\mathrm{T}}$ and $\mathrm{p}^{\mathrm{gt}} = [\mathrm{x}^{\mathrm{gt}}, \mathrm{y}^{\mathrm{gt}}]^{\mathrm{T}}$ is the center point of box E and box E^{gt}, and c is the diagonal length of box G. ρ, which is specified as the Euclidean distance.

The consistency of the aspect ratio is achieved, as shown in Equation (9).

$$V = \frac{4}{\pi^2}\left(arctan\frac{w^{gt}}{h^{gt}} - arctan\frac{w}{h}\right)^2 \tag{9}$$

Finally, the loss function CIoU of the complete IoU is obtained, as shown in Equation (10).

$$\mathcal{L}_{CIoU} = 1 - IoU + \frac{\rho^2\left(p, p^{g'}\right)}{c^2} + \alpha V \tag{10}$$

Among them, IoU (Intersection over Union) represents the Intersection over Union (IoU) between the predicted bounding box and the ground truth bounding box, (b, $\mathrm{b_g}$) are the center points of the predicted bounding box and the ground truth bounding box, respectively, $\rho^2(\mathrm{b},\mathrm{b_g})$ is the Euclidean distance between the two center points, c is the diagonal length of the minimum bounding box, υ measures the consistency of the aspect ratio, α is the weight coefficient, which controls the influence of the aspect ratio loss.

Where α is a trade-off parameter, as shown in Equation (11).

$$\alpha = \begin{cases} \frac{V}{(1-\text{IoU})+V} & if\ \text{IoU} > 0.5 \\ 0 & if\ \text{IoU} < 0.5 \\ \frac{V}{(1-\text{IoU})+V} & if\ \text{IoU} = 0.5 \end{cases} \tag{11}$$

The CIoU loss can rapidly shorten the distance between two predicted bounding boxes, so its convergence speed is much faster than that of the GIoU loss. For cases involving two predicted bounding boxes or with extreme aspect ratios, the CIoU loss will make the regression very fast, while the GIoU loss almost degenerates into an IoU loss.

3.5. Adaptive Attention Mechanism

The SE attention mechanism significantly enhances the model's ability to extract key features of blueberry ripeness through dynamic adjustment of channel weights: In terms of color perception, this mechanism can strengthen the color channels related to ripeness (such as purple, red, and green), and suppress background and interfering color channels, thereby improving color robustness under complex lighting conditions; in terms of shape and texture representation, by introducing SE in the shallow P3 layer, it enhances the perception of small target edges and local textures, and realizes adaptive selection of multi-scale features at the cross-scale fusion node to balance details and overall shape. At the same time, SE effectively reduces the interference of surface specular reflection on features by inhibiting the weights of high-brightness and high-contrast channels. Visual analysis further validates the effectiveness of this mechanism, showing that the attention weights can specifically concentrate on the key color channels at different ripeness stages, such as the purple channel of mature blueberries, the red transition channel of semi-ripe fruits, and the green channel of unripe fruits, indicating that the SE mechanism can adaptively focus on the feature information most relevant to ripeness discrimination.

The calculation process of the SE module can be concisely described as follows:

1. Compression stage:

$$Z_c = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} X_c(i,j) \tag{12}$$

where: X is the input feature map, C is the number of channels, H × W is the spatial dimension, Z_C is the global feature descriptor of the CTH channel, and Z is the compressed vector.

2. Motivation stage:

$$a = \delta(W_1 Z + b_1) \tag{13}$$

$$S = \sigma(W_2 a + b_2) \tag{14}$$

where W_1 is the first fully connected layer weight matrix, b_1 is the bias term, δ is the ReLU function, and a is the intermediate feature. W_2 is the weight matrix of the second fully connected layer, b_2 is the bias term, σ is the Sigmoid function, and S is the final channel attention weight vector.

3. Re-weighting stage:

$$\widetilde{X}_C = S_C \cdot X_C \tag{15}$$

where: $\widetilde{X}_C$ is the reweighted feature map of the CTH channel.

The estimation of channel weights is an end-to-end process from global information statistics to nonlinear relationship learning, and then to feature recalibration [42].

3.6. M-YOLOV11n Object Detection Network

While ensuring the real-time performance of the object detection network, it is necessary to meet and improve the accuracy of the object detection network in recognizing blueberry fruit as much as possible. In this study, to enhance the performance of YOLOv11n in multi-scale object detection tasks, especially in terms of accuracy and robustness when dealing with targets of different sizes, it is necessary to optimize its Backbone structure. The C3k2 module in the original YOLOv11n Backbone was replaced by a Multi-Scale Block (MsBlock), and the depthwise separable convolution component was introduced in the MsBlock.

In the proposed M-YOLOv11n, multiple MsBlock modules are sequentially deployed in the backbone. As shown in Table 3, three MsBlock stages output feature maps with resolutions of 1/8, 1/16, and 1/32, which are further forwarded to the neck network for multi-scale feature fusion and detection. The remaining MsBlock operates at a higher-resolution stage to enhance shallow feature representation and is not directly connected to the detection heads. The network structure diagram is shown in Figure 7.

Table 3. Backbone architecture comparison between YOLOv11n and M-YOLOv11n.

Stage	Module (YOLOv11n)	Module (M-YOLOv11n)	Output Size	Stride
Stem	Conv + C3k2	Conv + C3k2	1/4	4
Stage 1	C3k2	MsBlock	1/8	8
Stage 2	C3k2	MsBlock	1/16	16
Stage 3	C3k2	MsBlock	1/32	32
Neck Input	FPN	FPN + SE	P3, P4, P5	–

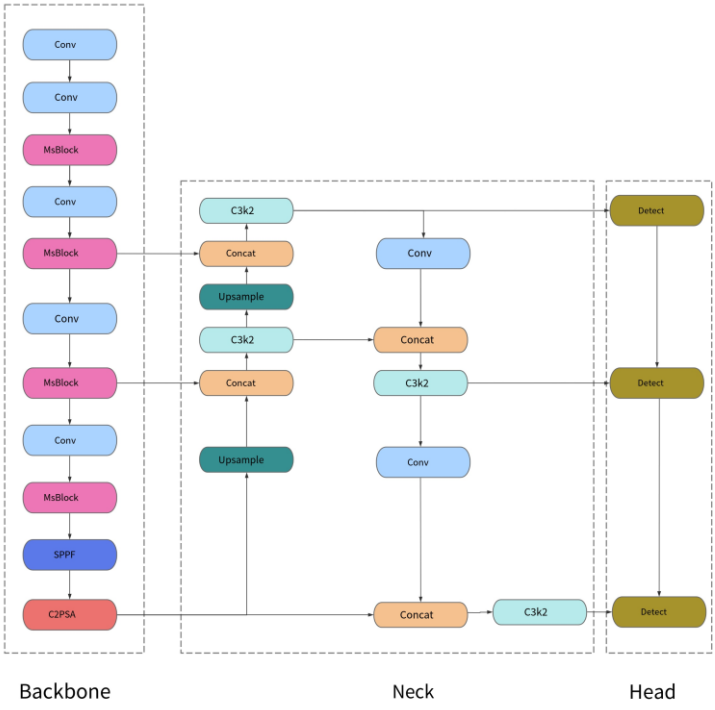


Figure 7. M-YOLOV11n network structure diagram.

The original C3k2 module of YOLOv11n Backbone consists of successive convolutional and pooling layers, where C3 represents a 3 × 3 convolution operation, and K2 represents each convolutional layer followed by a 2 × 2 pooling layer. This structure can effectively extract high-order features of images. However, due to the fixed size of the convolution

kernel, it is prone to poor adaptability to targets with significant scale changes, especially in small object detection and complex backgrounds, where its feature extraction ability is limited [43]. C3k2 is calculated by a standard 3 × 3 convolution as follows.

$$Y = \sigma(W * X + b) \tag{16}$$

where X is the input feature map, W is the 3 × 3 convolution kernel, b is the bias term, which represents the convolution operation, and σ is the activation function. Since all targets use the same scale feature extraction method, this structure may not be able to fully extract key information when dealing with targets with large scale variations or small targets, resulting in a decrease in detection accuracy.

In order to overcome the shortcomings of C3k2, we introduce MsBlock, whose core idea is to use multi-scale convolution kernels to extract feature information of different scales in parallel, and then improve the detection ability through the fusion mechanism. The calculation process of the MsBlock structure is as follows:

$$F_{ms} = \sum_{i=1}^{N} \alpha_i (W_i * X) \tag{17}$$

Among them, W_i represents convolution kernels of different scales, of which α_i are channel weighting factors calculated through the channel attention mechanism, and F_{ms} is the multi-scale feature map after fusion. Specifically, W_i (i = 1, 2, …, N) denotes a set of convolution kernels operating in parallel, with their scales designed to capture multi-granularity information ranging from local details to global context. The channel weighting factor α_i is dynamically generated by a channel attention mechanism (e.g., SE block): the mechanism first learns the importance of each channel through global average pooling and fully-connected layers, then assigns weights via a normalization function, enabling the model to adaptively enhance informative feature channels while suppressing redundant ones. This weighted fusion mechanism allows F_{ms} to effectively integrate multi-scale representations, thereby improving the model's discriminative ability for blueberry maturity detection in complex scenarios.

Convolutional layers of different scales can capture local details, such as clustered small targets like blueberries and global information, and adaptively adjust the contribution of features at each scale through a channel weighting mechanism. Finally, MsBlock generates more diverse feature representations to improve the detection performance of the model in complex scenes.

3.7. Experimental Design and Evaluation

3.7.1. Experimental Platform

The training and testing in this study were run on a computer equipped with Inter® Core™ i5-12400F, NVIDIA RTX 4060 GPU, and 32 GB running memory. With Cuda 12.6.2 parallel computer framework and Cudnn 9.6.0 deep learning acceleration library installed. All experiments were conducted under Windows 11 using Python 3.10.15 and PyTorch 2.5.1. The baseline YOLOv11n model was implemented based on the official open-source YOLOv11 repository, following the default network configuration and training pipeline. To accelerate convergence and improve training stability, all models were initialized with COCO pre-trained weights. No layers were frozen during training, and all parameters were fine-tuned on the blueberry maturity dataset to ensure a fair and consistent comparison among different model variants.

3.7.2. M-YOLOv11n Ablation Experiment

To verify the influence of MsBlock, Squeeze-and-Excitation (SE) module and depthwise separable convolution on model performance, the following seven sets of experiments were designed, as shown in Table 4.

Table 4. Ablation experiment description table.

Experimental Subjects	Annotation
Experiment1: YOLOv11n	Original YOLOv11n model
Experiment2: YOLOv11n + MsBlock	Only replace the C3k2 module with the MsBlock in the Backbone
Experiment3: YOLOv11n + SE	Only add the SE attention mechanism on the basis of Experiment 1
Experiment4: YOLOv11n + depthwise separable convolution	Only add depthwise separable convolution on the basis of Experiment 1
Experiment5: YOLOv11n + MsBlock + depthwise separable convolution	Based on Experiment 2, the MsBlock convolution type was changed to depthwise separable convolution
Experiment6: YOLOv11n + MsBlock + SE	Only add the SE attention mechanism on the basis of Experiment 2
Experiment7: YOLOv11n + MsBlock + depthwise separable convolution +SE	The complete M-YOLOv11n model

Standardized configuration was used in this study to ensure the reproducibility and comparability of results [44]. In terms of data set division, the experiment adopts a fixed training set, validation set and test set division strategy, and all comparison experiments are based on the same data distribution for model training and performance verification, so as to eliminate the influence of data randomness on experimental results [45]. In terms of hyperparameter configuration, the batch size (batch_size) of model training was set as 16, the training period (epochs) was set as 200, the early stop patience was set as 15 epochs, and the minimum improvement threshold (delta) was 0.001. The SGD optimizer combined with cosine annealing learning rate scheduling strategy was used. The initial learning rate was set to 0.005, the momentum coefficient was set to 0.937, and the weight decay coefficient was 0.0005. In order to comprehensively evaluate the performance of the model, multi-dimensional evaluation indicators are selected as follows: In terms of detection accuracy, the mean average accuracy (mAP50) and recall rate (Recall) are adopted as the main indicators. Real-time performance is evaluated by the frame rate per second (FPS) to measure the inference speed, and model complexity is quantified by the number of trainable parameters (Params). The FPS test is performed on a single NVIDIA RTX 4060 GPU with 640×640 input resolution.

All reported metrics are obtained by averaging the results of three independent runs with different random seeds. This experimental protocol, which is commonly adopted in object detection and YOLO-based studies, is used to mitigate the randomness introduced by stochastic optimization and to evaluate the consistency of performance trends. The training, validation, and test splits are kept identical across all runs to ensure experimental consistency and fair comparison.

3.7.3. Comparison Test

To comprehensively evaluate the proposed M-YOLOv11n, a representative set of baseline models was selected for comparison. The selection covers diverse architectural paradigms and design priorities to ensure a robust assessment. Specifically, YOLOv11n serves as the direct baseline to isolate the contribution of the proposed MsBlock and SE modules. YOLOv8n represents the widely adopted previous-generation state-of-the-art

in lightweight YOLO detectors, providing an evolutionary benchmark. SSD-MobileNet is included as a classical, efficiency-optimized one-stage detector, establishing a standard for mobile and embedded performance. Finally, Faster R-CNN provides a high-accuracy two-stage detector reference, representing an accuracy upper bound to contextualize the speed–accuracy trade-off of the proposed lightweight model. This multifaceted comparison across architecture types, generations, and design goals offers a comprehensive evaluation context for M-YOLOv11n.

All comparison models were trained and tested on the blueberry dataset constructed in this study. A uniform experimental configuration was adopted to ensure fairness. The input image resolution was 640 × 640 pixels, the batch size was 16, and the training period was 200. The optimizer adopted SDG and was combined with the cosine annealing learning rate scheduling strategy. The initial learning rate was 0.005. The initial learning rate is 0.005, the momentum coefficient is 0.937, and the weight decay coefficient is 0.0005.

3.7.4. Experimental Index

For the recognition of blueberry targets in natural and complex environments, the accuracy and real-time performance of the detection network need to be taken into consideration.

In this study, Mean Average Precision (mAP, %) is adopted as the evaluation index of the model's detection accuracy. mAP is related to recall and accuracy rate, and its calculation is shown in Equations (18) to (21) [46].

Recall: It reflects the completeness of the model's detection of positive samples

$$Recall = \frac{TP}{TP + FN} \tag{18}$$

Precision: The true reliability of a positive sample in the response model's detection

$$Precision = \frac{TP}{TP + FP} \tag{19}$$

Average Precision: It reflects the comprehensive continuity of the detection accuracy of a single category and is the integral representation of the precision rate—recall curve

$$AP = \int_0^1 P(R)dR \tag{20}$$

Mean Average Precision: It reflects the overall average level of detection accuracy of the model for all categories and is the summary of the mean values of each single-category AP

$$\text{mAP} = \frac{1}{M}\sum_{k=1}^{M} \text{AP}(k) \tag{21}$$

In the above equation, TP is the number of samples correctly classified as positive, FP is the number of samples incorrectly classified as positive, FN is the number of samples incorrectly classified as negative, M is the total number of categories, and AP(k) is the AP value of the KTH class.

The F1 score is a metric used to measure the accuracy of a binary classification model, and is often used as an experimental metric for comparison. The F1 score can be regarded as a weighted average of the model's accuracy and recall, with a maximum value of 1 and a minimum value of 0, as shown in Equation (26).

$$F1 = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall} \tag{22}$$

In object detection models, FPS usually refers to Frames Per Second, which is used as an experimental metric to measure the speed of the model in real-time images, specifically how many detections the model can complete per second.

4. Results

4.1. Analysis of MsBlock Ablation Results

To ensure controlled analysis and reduce confounding factors, the initial ablation experiments were conducted under the slight occlusion scenario. In addition, to comprehensively evaluate the robustness contribution of each module, supplementary ablation results under severe occlusion and backlight scenarios, as well as a global evaluation over all scenarios, are further reported and discussed, as shown in Table 5.

Table 5. Analysis table of ablation experiment results.

Scenario	Experimental Group	mAP50	Precision	Recall	FPS	F1 Score	Parameter
slight occlusion	Experiment 1: YOLOv11n	0.930	0.936	0.917	186.237	0.926	2.60 M
	Experiment 2: YOLOv11n + MsBlock	0.947	0.940	0.924	178.077	0.932	2.81 M
	Experiment 3: YOLOv11n + SE	0.939	0.945	0.898	182.044	0.921	2.64 M
	Experiment 4: YOLOv11n + depthwise separable convolution	0.919	0.941	0.916	197.973	0.928	2.36 M
	Experiment 5: YOLOv11n + MsBlock + depthwise separable convolution	0.945	0.956	0.932	189.104	0.944	2.57 M
	Experiment 6: YOLOv11n + MsBlock + SE	0.962	0.963	0.957	174.979	0.960	2.85 M
	Experiment 7: YOLOv11n + MsBlock + SE + depthwise separable convolution	0.970	0.974	0.971	177.620	0.973	2.61 M
severe occlusion	Experiment 1: YOLOv11n	0.915	0.910	0.909	182.654	0.908	2.60 M
	Experiment 2: YOLOv11n + MsBlock	0.928	0.922	0.920	176.664	0.921	2.81 M
	Experiment 3: YOLOv11n + SE	0.922	0.933	0.905	183.452	0.919	2.64 M
	Experiment 4: YOLOv11n + depthwise separable convolution	0.912	0.912	0.905	197.067	0.908	2.36 M
	Experiment 5: YOLOv11n + MsBlock+ depthwise separable convolution	0.936	0.940	0.930	190.008	0.935	2.57 M
	Experiment 6: YOLOv11n + MsBlock + SE	0.948	0.952	0.945	172.336	0.948	2.85 M
	Experiment 7: YOLOv11n + MsBlock + SE + depthwise separable convolution	0.961	0.959	0.955	175.427	0.957	2.61 M
backlight	Experiment 1: YOLOv11n	0.917	0.912	0.914	183.991	0.913	2.60 M
	Experiment 2: YOLOv11n + MsBlock	0.930	0.929	0.933	179.332	0.931	2.81 M
	Experiment 3: YOLOv11n + SE	0.938	0.942	0.923	181.187	0.932	2.64 M
	Experiment 4: YOLOv11n + depthwise separable convolution	0.915	0.914	0.910	195.479	0.912	2.36 M
	Experiment 5: YOLOv11n + MsBlock+ depthwise separable convolution	0.938	0.940	0.935	188.791	0.938	2.57 M
	Experiment 6: YOLOv11n + MsBlock + SE	0.955	0.958	0.939	173.759	0.948	2.85 M
	Experiment 7: YOLOv11n + MsBlock + SE + depthwise separable convolution	0.963	0.958	0.943	176.827	0.950	2.61 M

All indicators are averaged through three independent experiments to reduce random errors. All results reported in Tables 2 and 3 are obtained by averaging three independent runs with identical experimental settings. Although the standard deviation is not explicitly reported, the proposed M-YOLOv11n consistently demonstrates performance improvements across all runs, with stable performance trends and model rankings. Given the limited number of repeated experiments (n = 3), reporting standard deviation or hypothesis-testing-based statistical measures may be unstable and potentially misleading. Therefore, this study emphasizes consistency across independent runs as an indicator of the robustness and reliability of the proposed improvements.

As shown in Table 5, the introduction of the Multi-Scale Block (MsBlock) (MsBlock) alone can achieve a 1.83% improvement in mAP50, a 0.76% improvement in Recall, and a 0.65% improvement in F1-score, verifying its effectiveness in enhancing the feature representation ability of the model. However, the number of parameters increases by 0.21 M and the FPS slightly decreases. The introduction of the Channel Attention Module (SE) alone increased mAP50 and Precision by 0.97% and 0.96%, respectively, and the number of parameters (+0.04 M) was almost the same. Although Recall decreased by 2.07% and F1-score slightly decreased, it highlighted the advantage of its refined feature selection. The strategy of using depthwise separable convolution alone performs most prominently in terms of efficiency optimization. While significantly reducing the number of parameters (−0.24 M) and greatly increasing the frame rate (+11.74 FPS), it still maintains a high Precision and F1-score, demonstrating its key value in lightweight design.

The combination of MsBlock and depthwise separable convolution achieved a 1.61% improvement in mAP50 under the conditions that the number of parameters was nearly the same as the baseline (−0.03 M) and the FPS remained at a high level. Meanwhile, the Precision and F1-score increased significantly by 2.14% and 1.94%, respectively. This demonstrates the good synergy between the two in balancing accuracy and efficiency. The combination of MsBlock and the Squeeze-and-Excitation module has produced a more powerful performance gain. mAP50, Precision, Recall and F1-score have increased significantly by 3.44%, 2.88%, 4.36%, and 3.67%, respectively, and the number of parameters has increased slightly by 0.25 M. Due to the MsBlock and attention mechanism adding additional computational paths and memory access, this will reduce the inference speed on the GPU, thereby resulting in a decrease in FPS. The overall performance also strongly verifies that the combination of multi-scale features and channel attention can bring about a substantial leap in detection accuracy. Overall, the comparison results are shown in Figure 8.

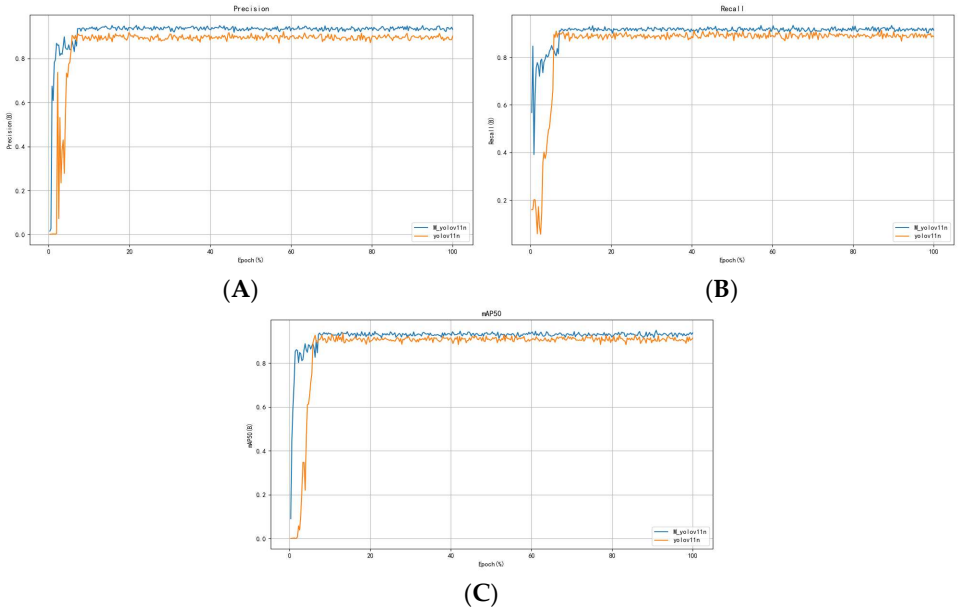


Figure 8. Comparison of results between M-YOLOv11n and YOLOv11n: (**A**) Comparison of accuracy between M-YOLOv11n and YOLOv11n; (**B**) Comparison of recall between M-YOLOv11n and YOLOv11n; (**C**) mAP of M-YOLOv11n and YOLOv11n.

In severe occlusion scenarios, the SE module alone increases accuracy by 2.22% but slightly reduces recall (−0.44%), indicating enhanced classification confidence at the cost of missing heavily occluded regions due to its focus on locally salient features. By contrast, the

multi-scale module provides larger gains in recall and F1 score, highlighting its advantage in detecting occluded targets. Under backlight conditions, where low contrast increases false detections, the SE module improves accuracy (+3.29%) much more than recall (+1.0%), confirming its effectiveness in enhancing illumination-insensitive features and suppressing light–shadow-induced false positives. Depthwise separable convolution substantially improves inference speed (≈195–197 FPS) with negligible impact on accuracy and recall, demonstrating its lightweight efficiency. Overall, the integrated model achieves a well-balanced improvement in accuracy and recall under both severe occlusion (0.959/0.955) and backlight conditions (0.958/0.943) without a notable increase in parameters, validating the effectiveness and general applicability of the proposed multi-scale, attention-enhanced, and lightweight design strategy in complex natural scenes.

When MsBlock, SE attention, and depthwise separable convolution are jointly integrated, the overall best performance is achieved. Compared with the baseline model, the final model achieves practically meaningful improvements in all core accuracy indicators of mAP50 (+4.30%), Precision (+4.06%), Recall (+5.89%), and F1-score (+5.08%). Most importantly, this optimal result was achieved under the condition of a slight increase in the number of parameters (+0.01 M) while the FPS remained in a highly efficient state. These results consistently indicate that the improved model proposed in this study effectively combines the lightweight advantages of multi-scale feature extraction, channel attention recalibration, and depthwise separable convolution, achieving a better balance among model accuracy, efficiency, and complexity. This fully verified the effectiveness and advancement of the integrated strategy of MsBlock, the SE attention mechanism, and depthwise separable convolution.

4.2. Analysis of Comparative Experimental Results

As shown in Table 6, in this study, five object detection networks, namely M-YOLOv11n, Faster RCNN, YOLOv11n, YOLOv8n, and SSD-MobileNet, were trained on the self-built blueberry maturity dataset. And recognition experiments were conducted on blueberry fruits of different maturity under three natural environments: slight occlusion, severe occlusion, and backlight.

Table 6. Analysis table of comparative test results.

Scenario	Model	mAP50	mAP50:95	Precision	Recall	FPS	F1 Score	Parameter
slight occlusion	SSD-MobileNet	0.905	0.543	0.917	0.876	134.600	0.897	2.44 M
	YOLOv8n	0.928	0.571	0.924	0.921	206.719	0.923	3.20 M
	YOLOv11n	0.930	0.620	0.936	0.917	186.237	0.926	2.60 M
	Faster R-CNN	0.921	0.623	0.936	0.907	39.308	0.921	41.76 M
	M-YOLOv11n	0.970	0.681	0.974	0.971	177.620	0.973	2.61 M
severe occlusion	SSD-MobileNet	0.833	0.497	0.845	0.821	139.665	0.833	2.44 M
	YOLOv8n	0.909	0.577	0.904	0.902	201.416	0.903	3.20 M
	YOLOv11n	0.915	0.586	0.910	0.909	182.654	0.908	2.60 M
	Faster R-CNN	0.911	0.583	0.903	0.895	38.640	0.899	41.76 M
	M-YOLOv11n	0.961	0.643	0.959	0.955	175.427	0.957	2.61 M
backlight	SSD-MobileNet	0.833	0.478	0.829	0.804	117.647	0.817	2.44 M
	YOLOv8n	0.909	0.533	0.902	0.911	202.324	0.907	3.20 M
	YOLOv11n	0.917	0.575	0.912	0.914	183.991	0.913	2.60 M
	Faster R-CNN	0.910	0.569	0.917	0.891	37.258	0.904	41.76 M
	M-YOLOv11n	0.963	0.622	0.958	0.943	176.827	0.950	2.61 M

The comparison results of the five object detection networks are shown in Table 6. Under the condition of slight occlusion of fruits, the average accuracy of the five object

detection networks all reached over 90%. Among them, the average accuracy, accuracy rate, recall, and F1 score of the M-YOLOv11n network structure were all higher than those of the other four network structures. The results are 97%, 97.4%, 97.1%, and 97.3%, respectively, which proves that it has a significant improvement in detection accuracy compared with the other four object detection networks. In the case of severe occlusion, the average accuracy and F1 score of the M-YOLOv11n object detection network proposed in this study are 5 and 5.8 percentage points higher than those of Faster R-CNN, respectively, and 5.2 and 5.4 percentage points higher than those of YOLOv8n, respectively. Compared with YOLOv11n, which has a similar inference speed, the average accuracy and F1 score have also increased by 4.6 and 4.9 percentage points, respectively. Under backlight conditions, the average accuracy of the four network structures, namely M-YOLOv11n, Faster RCNN, YOLOv11n, and YOLOv8n, all reach more than 90%. However, the detection time of the Faster R-CNN object detection network is much longer than that of the object detection network M-YOLOv11n in this study. In terms of mAP50:95, M-YOLOv11n shows consistently higher scores across the three scenarios. The inference speed of the M-YOLOv11n object detection network is 176.827 frames·s^{-1}. By comparison, it can be seen that the inference speed of M-YOLOv11n is nearly five times higher than that of Faster R-CNN. Although M-YOLOv11n is approximately 7 frames·s^{-1} slower than YOLOv11n in terms of inference speed, M-YOLOv11n improves its average accuracy and F1 score by 4.6 and 3.7 percentage points, respectively, compared to YOLOv11n. In terms of parameter count, although SSD-MobileNet possesses the smallest number of parameters, it exhibits the lowest accuracy and poorest overall detection performance among the five models. In comparison, the proposed M-YOLOv11n marginally increases the parameter count relative to the baseline YOLOv11n but achieves substantial improvement in detection performance. Across the three scenarios evaluated in the table, the M-YOLOv11n object detection network proposed in this study attains average precision and F1 scores of 96.5% and 96%, respectively, with virtually no increase in parameter volume.

The recognition effects of different object detection networks on blueberry fruit in different natural environments are shown in Figure 9. Through comprehensive comparison, it can be seen that the M-YOLOv11n object detection algorithm can accurately frame unripe, semi-ripe, and ripe blueberry fruit under different quantities, sizes, lighting conditions, and degrees of occlusion by branches, leaves, and fruits. It has a very high recognition accuracy rate, while the other four object detection networks have experienced false detections and missed detections. Therefore, the algorithm in this study has strong robustness and can adapt to different situations in the natural environment.

To further explain the mechanism behind the performance improvement, Figure 10 presents qualitative activation map comparisons under three representative scenarios: clustered fruits with occlusion, mixed maturity stages, and backlit conditions. Each group includes the original image, the activation map of YOLOv11n, and that of the proposed M-YOLOv11n.

In the clustered and partially occluded scenario (Figure 10A), the baseline YOLOv11n shows relatively dispersed activations that extend to surrounding leaves, indicating interference from occluding structures. In contrast, M-YOLOv11n produces more compact and fruit-aligned responses, especially on partially visible berries. This suggests that the multi-scale aggregation in MsBlock enables complementary modeling of local exposed regions (via small receptive fields) and cluster-level context (via larger receptive fields), while the SE module suppresses channels dominated by leaf textures, improving target–background separation.

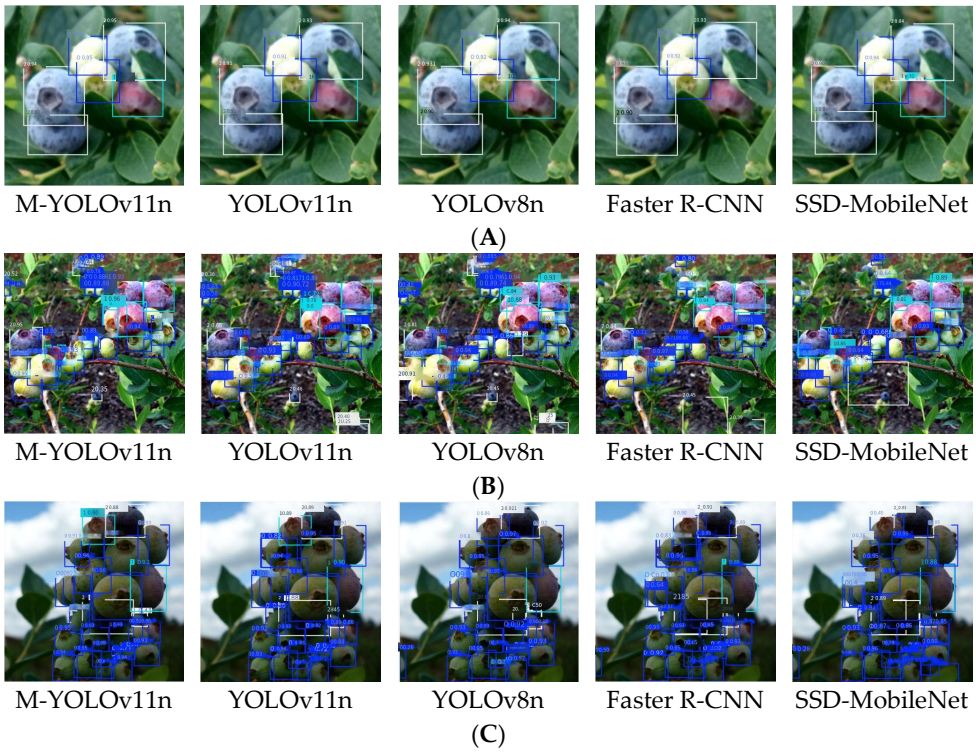


Figure 9. Shows the detection effects of different object detection networks on blueberry fruit in three scenarios: (**A**) slight occlusion; (**B**) severe occlusion; (**C**) backlight.

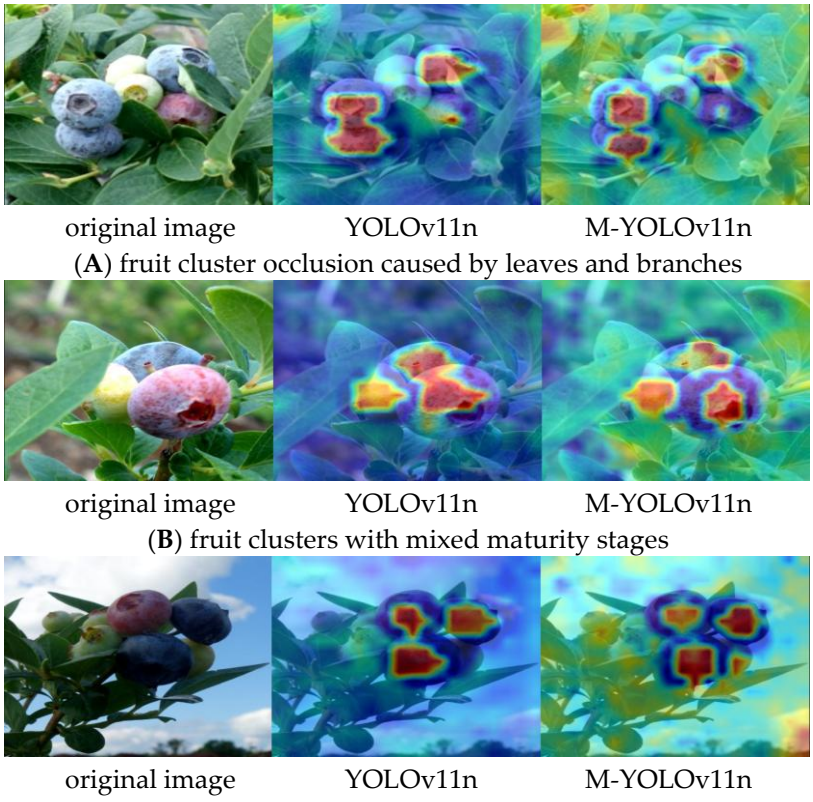


Figure 10. Heatmaps illustrating the network performance before and after improvement in three scenarios: (**A**) fruit cluster occlusion caused by leaves and branches; (**B**) fruit clusters with mixed maturity stages; (**C**) strong backlight interference.

In the mixed maturity scenario (Figure 10B), YOLOv11n exhibits similar response intensities across berries of different maturity levels, implying limited discrimination of subtle chromatic differences. However, M-YOLOv11n highlights berries with stronger red–purple color responses more distinctly, while reducing attention on immature green berries. This indicates that MsBlock captures color cues at multiple spatial granularities, and the SE module further amplifies maturity-related color channels, enhancing sensitivity to color gradients that are critical for blueberry maturity recognition.

Under backlit conditions (Figure 10C), the baseline model shows weakened and fragmented activations due to reduced contrast and specular reflections. By comparison, M-YOLOv11n maintains stable and continuous responses over fruit regions, despite illumination variation. This demonstrates that multi-scale features help integrate local high-frequency cues (e.g., specular highlights and edges) with global shape information, while SE adaptively reweights illumination-robust channels, resulting in improved robustness under challenging lighting.

Overall, these activation maps indicate that the proposed MsBlock and SE design enhances feature representation by jointly addressing scale variation, occlusion, and illumination changes, thereby encouraging the detector to place greater emphasis on maturity-discriminative visual cues while reducing the influence of background-related responses.

To complement the success cases presented earlier, Figure 11 offers a targeted qualitative analysis of the M-YOLOv11n model under two highly demanding field conditions: severe multi-layer occlusion and strong backlighting. Under extreme occlusion (Figure 11A,B), the model fails to detect blueberries that are visually occluded by more than one layer of foliage or other fruits (red circles in Figure 11B). This suggests that while the model can handle partial or single-layer obstructions, its current architecture reaches a limit when object contours and color cues are nearly absent from the input. In intense backlight scenarios (Figure 11C,D), high dynamic range and deep shadows lead to false positives, where background patches with specular highlights or shadowed textures are misinterpreted as fruit surfaces (red circles in Figure 11D). These observed failure modes highlight specific environmental thresholds—namely, near-total visual occlusion and extreme luminance contrast—beyond which the model's reliability decreases. Such insights are critical for understanding the operational boundaries of the proposed system in real agricultural settings.

The P-R curve is a curve with accuracy as the vertical axis and recall as the horizontal axis, which can reflect the overall performance of the object detection network. The P-R curves of the five object detection networks constructed under the blueberry test set are shown in Figure 12. The curve of the M-YOLOv11n object detection network is outside the curves of the other four object detection networks. Moreover, the position of the equilibrium point (the value when the accuracy is equal to the recall rate) is closer to the coordinate (1,1), which proves that the detection accuracy of the M-YOLOv11n proposed in this study is higher than that of the other four object detection networks. The variation curve of the loss value of the M-YOLOv11n object detection network during training after the number of training rounds is shown in Figure 13. It can be seen from the figure that when the number of training rounds exceeds 200, the loss value basically tends to be stable, and then the network structure converts. The main limitations of this study lie in its reliance on data from a single farm and cultivar, as well as the lack of deployment validation on real agricultural edge devices. The proposed lightweight multi-scale architecture demonstrates potential for migration to other blueberry varieties and similar small-fruit detection tasks. This work provides a technical example for achieving high-precision real-time visual detection in resource-constrained environments, contributing to the practical application and deployment of lightweight AI models in smart agriculture.

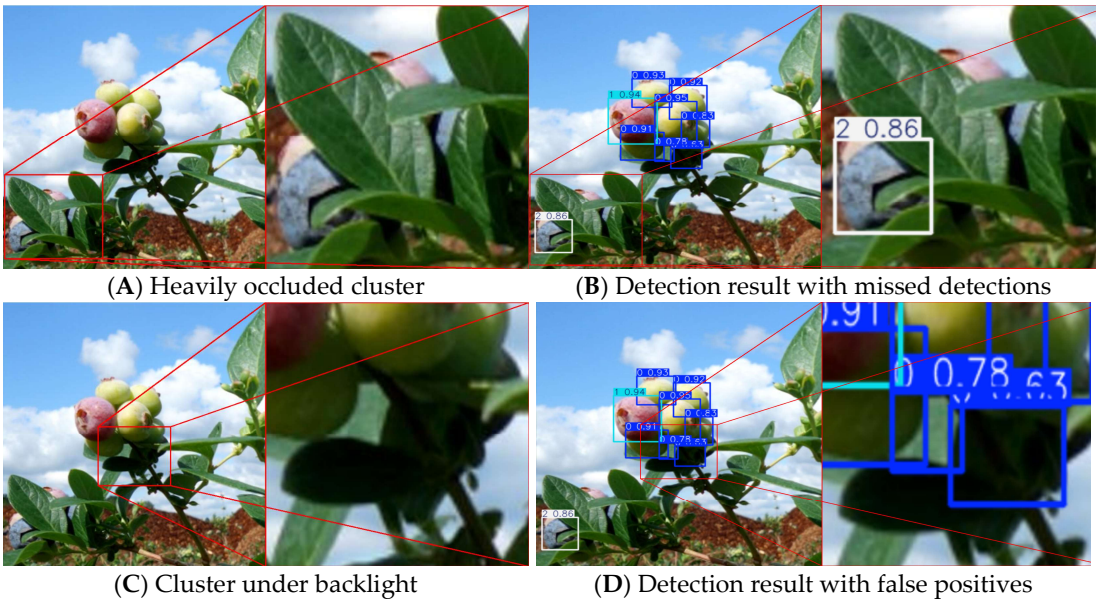


Figure 11. Analysis of model performance under challenging conditions. (**A**) Heavily occluded cluster; (**B**) Detection result with missed detections; (**C**) Cluster under backlight; (**D**) Detection result with false positives.

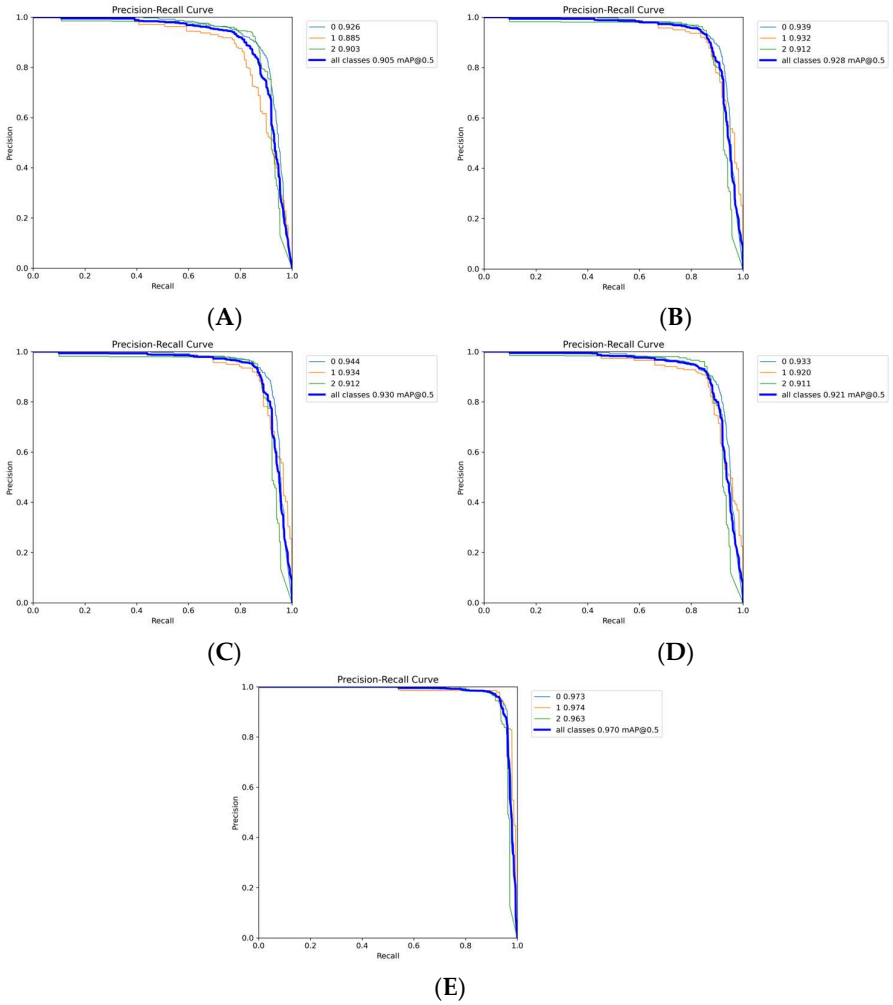


Figure 12. Comparison of PR curves of each model under slight occlusion conditions: (**A**) PR curve of the SSD-MobileNet model; (**B**) PR curve of the YOLOv8n model; (**C**) PR curve of the YOLOv11n model; (**D**) PR curve of the Faster R-CNN model; (**E**) PR curve of the M-YOLOv11n model.

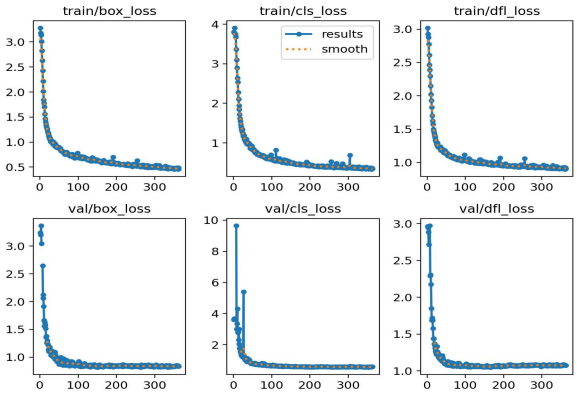


Figure 13. Convergence curve of model training and validation loss.

5. Discussion

This study proposes an improved YOLOv11n multi-scale module-based object detection network (M-YOLOv11n) for the recognition and detection of blueberry fruit of different ripeness. On the YOLOv11n object detection network, fusion introduces the Multi-Scale Block (MsBlock) (multi-scale Block, MsBlock) of depthwise separable convolution and the adaptive attention module (Squeeze-and-Excitation, SE). While significantly improving mAP, Precision, Recall, and F1-score, it only brings about a small increase in the number of parameters and memory usage. Overall, it still maintains a lightweight structure, which is conducive to deployment on agricultural embedded mobile devices and provides a reliable detection basis for picking robots and early yield estimation of crops.

According to the different scenes in the natural environment, blueberry image datasets in three scenarios, namely slight occlusion, severe occlusion, and backlight, were created. Comparative experiments were conducted using the YOLOv11n object detection network before and after improvement with the SSD-MobileNet, YOLOv8n, and Faster R-CNN object detection networks. The results show that the average accuracy and F1 score of the improved object detection network (M-YOLOv11n) reach 96.5% and 96%, respectively. For the detection of three different maturity blueberries, the M-YOLOv11n object detection network performs better and can provide higher recognition accuracy on the basis of achieving real-time performance.

In the current study, the evaluation of model performance mainly focuses on the characteristics of target detection and multi-classification tasks. Therefore, metrics such as mean average precision (mAP), precision, recall rate, and F1 score are used for measurement. These metrics directly reflect the model's ability to locate blueberry fruits in complex natural scenes and accurately classify their maturity categories. However, it should be noted that indicators such as coefficient of determination (R^2), mean absolute error (MAE), mean squared error (MSE), and root mean squared error (RMSE) are typically applicable to regression analysis of continuous variables, such as predicting specific physiological parameters like fruit size, sugar content, or hardness. The lightweight detection framework established in this study has laid a stable foundation for further implementation of continuous quantitative prediction of blueberry maturity. In future work, if relevant regression analysis is conducted, the above error and goodness-of-fit indicators will have significant evaluation value.

The M-YOLOv11n model proposed in this study demonstrated superior detection accuracy and real-time performance in natural environments, but still has some notable limitations. Firstly, at the commercialization and practical deployment level, although the model has the characteristic of being lightweight, integrating it into actual agricultural production processes (such as robot harvesting or post-harvest sorting lines) still requires

addressing numerous engineering issues. As a robot-mounted perception system, its performance needs to be systematically verified on mobile platforms, vibration disturbances, and different lighting periods, and it must meet the extremely low latency requirements for harvesting action planning. In the application of sorting lines, the camera installation position, production line speed, and the software and hardware coordination with existing sensors (such as weighing and spectrometers) need to be considered. Additionally, the long-term maintenance of the model after deployment is also crucial, including: dealing with the possible regular re-training or domain adaptation required when different production areas have different lighting conditions, variations in blueberry varieties (fruit size, fruit frost, color depth), and when considering the need for regular re-training or domain adaptation; establishing calibration procedures for different camera models and installation positions; and evaluating the model's robustness to common on-site disturbances such as lens stains and dust. Secondly, in terms of the generalization and transferability of the model, the dataset used in this study comes from a fixed camera configuration on a single farm in Florida, USA, which covers complex scenarios such as occlusion and backlighting, but does not systematically cover the diverse varieties and cultivation patterns of major global production areas. The model's adaptability to different blueberry varieties (especially those with significant differences in fruit skin luster or fruit frost characteristics) needs further verification. However, the core contribution of this study—that is, enhancing the model's feature representation and discrimination ability in complex scenarios through multi-scale modules (MsBlock) and channel attention (SE)—has a universal design concept. Therefore, the proposed improved architecture is expected to be directly transferable to other small fruit detection tasks (such as grapes, cherries, strawberries) and non-fruit agricultural target detection tasks (such as pest and disease identification, flower counting) with similar detection challenges, but data collection and fine-tuning for specific target crops are required. These limitations indicate the direction for future research, which is to promote the model from laboratory performance verification to field engineering application and cross-crop generalization.

Future research should focus on advancing the model from laboratory performance validation toward field-ready engineering applications and cross-crop generalization. Key directions include conducting hardware-in-loop validation on typical edge computing platforms (e.g., the Jetson series) to optimize throughput, power consumption, and stability in real-world deployments, as well as constructing dynamic datasets that encompass factors such as platform vibration, multi-period lighting variations, and seasonal changes to enhance system robustness. Concurrently, it is essential to collaboratively build multi-regional benchmark datasets spanning different cultivars and growing seasons, and to develop lightweight domain adaptation methods suitable for embedded devices, thereby reducing the cost of adapting the model to new environments and crops. A full lifecycle management system for the model should also be established, incorporating online performance monitoring, automated calibration, and interactive update protocols. Furthermore, the core architecture proposed in this study should be migrated to other intensive agricultural vision tasks to systematically verify its potential as a general lightweight detection framework. Through these efforts, a practical, maintainable, and scalable agricultural vision solution can be formed, providing solid technical support for the perception layer of smart agriculture.

Author Contributions: Conceptualization, X.L. and J.S.; methodology, X.L.; software, X.L. and Y.L.; validation, X.L., J.S. and C.W.; formal analysis, X.L., C.W. and W.S.; investigation, X.L.; resources, Z.Z. and X.Y.; data curation, X.L., J.S. and Y.L.; writing—original draft preparation, X.L.; writing—review and editing, X.L. and W.S.; visualization, X.L. and Y.L.; supervision, K.T., W.S. and J.N.; project

administration, K.T. and J.N.; funding acquisition, K.T. and J.N. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The datasets presented in this article are not readily available because some follow-up experiments and projects still need to use the data from this study. Requests to access the datasets should be directed to the corresponding authors.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Aguilera, C.A.; Figueroa-Flores, C.; Aguilera, C.; Navarrete, C. Comprehensive Analysis of Model Errors in Blueberry Detection and Maturity Classification: Identifying Limitations and Proposing Future Improvements in Agricultural Monitoring. *Agriculture* **2023**, *14*, 18. [CrossRef]
2. Rigid-Flexible Coupling Simulation Analysis and Test of Portable Blueberry Harvester. Available online: http://www.linyekexue.net/EN/10.11707/j.1001-7488.LYKX20240478 (accessed on 2 December 2025).
3. Tian, Y.; Qin, S.; Yan, Y.; Wang, J.; Jiang, F. Detecting Blueberry Maturity Under Complex Field Conditions Using Improved YOLOv8. *Trans. Chin. Soc. Agric. Eng.* **2024**, *40*, 153–162. [CrossRef]
4. 2022 Report. Int. Blueberry Organ. Available online: https://www.fao.org/faostat/zh/#data/QCL (accessed on 2 December 2025).
5. Cvetković, M.; Kočić, M.; Zagorac, D.D.; Ćirić, I.; Natić, M.; Hajder, Đ.; Životić, A.; Akšić, M.F. When Is the Right Moment to Pick Blueberries? Variation in Agronomic and Chemical Properties of Blueberry (*Vaccinium Corymbosum*) Cultivars at Different Harvest Times. *Metabolites* **2022**, *12*, 798. [CrossRef] [PubMed]
6. Hwang, H.; Kim, Y.-J.; Shin, Y. Assessment of Physicochemical Quality, Antioxidant Content and Activity, and Inhibition of Cholinesterase between Unripe and Ripe Blueberry Fruit. *Foods* **2020**, *9*, 690. [CrossRef]
7. AEB0075 | College of Agricultural Sciences. Available online: https://appliedecon.oregonstate.edu/enterprise-budgets/aeb0075?utm_source (accessed on 30 November 2025).
8. Yang, W.; Ma, X.; An, H. Blueberry Ripeness Detection Model Based on Enhanced Detail Feature and Content-Aware Reassembly. *Agronomy* **2023**, *13*, 1613. [CrossRef]
9. Yuan, J.; Fan, J.; Sun, Z.; Liu, H.; Yan, W.; Li, D.; Liu, H.; Wang, J.; Huang, D. Deployment of CES-YOLO: An Optimized YOLO-Based Model for Blueberry Ripeness Detection on Edge Devices. *Agronomy* **2025**, *15*, 1948. [CrossRef]
10. Ma, J.; Li, M.; Fan, W.; Liu, J. State-of-the-Art Techniques for Fruit Maturity Detection. *Agronomy* **2024**, *14*, 2783. [CrossRef]
11. Liu, Y.; Zheng, H.; Zhang, Y.; Zhang, Q.; Chen, H.; Xu, X.; Wang, G. "Is This Blueberry Ripe?": A Blueberry Ripeness Detection Algorithm for Use on Picking Robots. *Front. Plant Sci.* **2023**, *14*, 1198650. [CrossRef]
12. Bhargava, A.; Bansal, A. Fruits and Vegetables Quality Evaluation Using Computer Vision: A Review. *J. King Saud Univ. Comput. Inf. Sci.* **2021**, *33*, 243–257. [CrossRef]
13. Yang, Y.; Han, Y.; Li, S.; Yang, Y.; Zhang, M.; Li, H. Vision Based Fruit Recognition and Positioning Technology for Harvesting Robots. *Comput. Electron. Agric.* **2023**, *213*, 108258. [CrossRef]
14. Koirala, A.; Walsh, K.B.; Wang, Z.; McCarthy, C. Deep Learning—Method Overview and Review of Use for Fruit Detection and Yield Estimation. *Comput. Electron. Agric.* **2019**, *162*, 219–234. [CrossRef]
15. Xiao, X.; Jiang, Y.; Wang, Y. Key Technologies for Machine Vision for Picking Robots: Review and Benchmarking. *Mach. Intell. Res.* **2025**, *22*, 2–16. [CrossRef]
16. Hou, G.; Chen, H.; Jiang, M.; Niu, R.; Hou, G.; Chen, H.; Jiang, M.; Niu, R. An Overview of the Application of Machine Vision in Recognition and Localization of Fruit and Vegetable Harvesting Robots. *Agriculture* **2023**, *13*, 1814. [CrossRef]
17. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1137–1149. [CrossRef] [PubMed]
18. Wang, C.; Li, H.; Deng, X.; Liu, Y.; Wu, T.; Liu, W.; Xiao, R.; Wang, Z.; Wang, B.; Wang, C.; et al. Improved You Only Look Once v.8 Model Based on Deep Learning: Precision Detection and Recognition of Fresh Leaves from Yunnan Large-Leaf Tea Tree. *Agriculture* **2024**, *14*, 2324. [CrossRef]
19. CVPR 2023 Open Access Repository. Available online: https://openaccess.thecvf.com/content/CVPR2023/html/Wang_YOLOv7_Trainable_Bag-of-Freebies_Sets_New_State-of-the-Art_for_Real-Time_Object_Detectors_CVPR_2023_paper.html?utm_source=chatgpt.com (accessed on 30 November 2025).
20. Wang, L.; Qin, M.; Lei, J.; Wang, X.; Tan, K. Blueberry Maturity Recognition Method Based on Improved YOLOv4-Tiny. *Trans. Chin. Soc. Agric. Eng.* **2021**, *37*, 170–178. [CrossRef]

21. Li, Y.; Liu, M.; He, Z.; Lou, Y. Transactions of the Chinese Society of Agricultural Engineering (Transactions of the CSAE). Available online: http://www.tcsae.org/en/search (accessed on 2 December 2025).
22. Ma, C.; Zhang, H.; Ma, X.; Wang, J.; Zhang, Y.; Zhang, X. Method for the Lightweight Detection of Wheat Disease Using Improved YOLOv8. *Trans. Chin. Soc. Agric. Eng.* **2024**, *44*, 187–195. [CrossRef]
23. Ji, S.-J.; Ling, Q.-H.; Han, F. An Improved Algorithm for Small Object Detection Based on YOLO v4 and Multi-Scale Contextual Information. *Comput. Electr. Eng.* **2023**, *105*, 108490. [CrossRef]
24. Liu, Q.; Zhang, J.; Zhang, Z.; Bu, X.; Hanajima, N. A Lightweight YOLO Object Detection Algorithm Based on Bidirectional Multi-Scale Feature Enhancement. *Adv. Theory Simul.* **2024**, *7*, 2301025. [CrossRef]
25. Wang, H.; Liu, J.; Zhao, J.; Zhang, J.; Zhao, D. Precision and Speed: LSOD-YOLO for Lightweight Small Object Detection. *Expert Syst. Appl.* **2025**, *269*, 126440. [CrossRef]
26. Wang, C.; Han, Q.; Li, J.; Li, C.; Zou, X. YOLO-BLBE: A Novel Model for Identifying Blueberry Fruits with Different Maturities Using the I-MSRCR Method. *Agronomy* **2024**, *14*, 658. [CrossRef]
27. Feng, W.; Liu, M.; Sun, Y.; Wang, S.; Wang, J. The Use of a Blueberry Ripeness Detection Model in Dense Occlusion Scenarios Based on the Improved YOLOv9. *Agronomy* **2024**, *14*, 1860. [CrossRef]
28. Lin, T.-Y.; Dollar, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature Pyramid Networks for Object Detection. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 2117–2125.
29. Chen, Y.; Yuan, X.; Wang, J.; Wu, R.; Li, X.; Hou, Q.; Cheng, M.-M. YOLO-MS: Rethinking Multi-Scale Representation Learning for Real-Time Object Detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2025**, *47*, 4240–4252. [CrossRef] [PubMed]
30. Makraki, T.; Tsaniklidis, G.; Papadimitriou, D.M.; Taheri-Garavand, A.; Fanourakis, D. Non-Destructive Monitoring of Postharvest Hydration in Cucumber Fruit Using Visible-Light Color Analysis and Machine-Learning Models. *Horticulturae* **2025**, *11*, 1283. [CrossRef]
31. Tsaniklidis, G.; Makraki, T.; Papadimitriou, D.; Nikoloudakis, N.; Taheri-Garavand, A.; Fanourakis, D. Non-Destructive Estimation of Area and Greenness in Leaf and Seedling Scales: A Case Study in Cucumber. *Agronomy* **2025**, *15*, 2294. [CrossRef]
32. Kotthapalli, M.; Ravipati, D.; Bhatia, R. YOLOv1 to YOLOv11: A Comprehensive Survey of Real-Time Object Detection Innovations and Challenges. *arXiv* **2025**, arXiv:2508.02067. [CrossRef]
33. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. In Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.
34. Rasheed, A.F.; Zarkoosh, M. YOLOv11 Optimization for Efficient Resource Utilization. *arXiv* **2024**, arXiv:2412.14790. [CrossRef]
35. Gallagher, J.E.; Oughton, E.J. Surveying You Only Look Once (YOLO) Multispectral Object Detection Advancements, Applications, and Challenges. *IEEE Access* **2025**, *13*, 7366–7395. Available online: https://ieeexplore.ieee.org/document/10829595 (accessed on 30 November 2025). [CrossRef]
36. Li, S.; Yuan, Z.; Peng, R.; Leybourne, D.; Xue, Q.; Li, Y.; Yang, P. An Effective Farmer-Centred Mobile Intelligence Solution Using Lightweight Deep Learning for Integrated Wheat Pest Management. *J. Ind. Inf. Integr.* **2024**, *42*, 100705. [CrossRef]
37. Yuan, H.; Zhang, B.; Wang, Y. MSEB: Plug and Play Multi-Scale Image Embedding Block for Vision Backbone. *Neurocomputing* **2025**, *617*, 129040. [CrossRef]
38. Sitaula, C.; Aryal, J.; Bhattacharya, A. A Novel Multiscale Attention Feature Extraction Block for Aerial Remote Sensing Image Classification. *IEEE Geosci. Remote Sens. Lett.* **2023**, *20*, 1–5. [CrossRef]
39. Baharani, M.; Sunil, U.; Manohar, K.; Furgurson, S.; Tabkhi, H. DeepDive: An Integrative Algorithm/Architecture Co-Design for Deep Separable Convolutional Neural Networks. *arXiv* **2020**, arXiv:2007.09490.
40. A Multi-Scale Hierarchical Node Graph Neural Network for Few-Shot Learning. Multimedia Tools and Applications. Available online: https://link.springer.com/article/10.1007/s11042-023-17059-1 (accessed on 30 November 2025).
41. Rezatofighi, H.; Tsoi, N.; Gwak, J.; Sadeghian, A.; Reid, I.; Savarese, S. Generalized Intersection Over Union: A Metric and a Loss for Bounding Box Regression. In Proceedings of the 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 15–20 June 2019; pp. 658–666.
42. Hu, J.; Shen, L.; Sun, G. Squeeze-and-Excitation Networks. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141.
43. Sun, S.; Zhang, R.; Gu, Z.; Fang, X.; Ma, T.; Li, H.; Lin, Z.; Yi, H.; Zhang, X.; Wu, L.; et al. Interfacial Stabilization Mechanism of Zirconium-Based Halide and Li–In Alloy Anode in All-Solid-State Lithium Batteries. *Adv. Energy Mater.* **2025**, *15*, 18167–18175. [CrossRef]
44. Enhancing Image–Text Matching Through Multi-Level Semantic Consistency Alignment. The Visual Computer. Available online: https://link.springer.com/article/10.1007/s00371-025-03981-y (accessed on 20 December 2025).

45. Yuan, J.; Zhou, L.; He, M.; Luo, C.; Zhang, J. A Lightweight Dual Path Kolmogorov-Arnold Convolution Network for Medical Optical Image Segmentation. *Neurocomputing* **2026**, *659*, 131776. [CrossRef]
46. KTMN: Knowledge-Driven Two-Stage Modulation Network for Visual Question Answering. Multimedia Systems. Available online: https://link.springer.com/article/10.1007/s00530-024-01568-6 (accessed on 20 December 2025).