Journal of Agricultural Engineering
https://www.agroengineering.org/

YOLO-Banana-Seg: a lightweight and efficient model for rapid segmentation of banana bunches and stalks in complex orchards

Runbo Fu,[1,2] Fuqin Deng,[2] Zhijie Wu,[2] Lanhui Fu,[2] Lei Zhou[2,3]

[1]College of Artificial Intelligence and Low-Altitude Technology, South China Agricultural University, Guangzhou
[2]School of Electronics and Information Engineering, Wuyi University, Jiangmen
[3]Guangdong Laboratory of Artificial Intelligence and Digital Economy (SZ), Shenzhen, China

Corresponding authors: *Lanhui Fu*, School of Electronics and Information Engineering, Wuyi University, Jiangmen 529020, Guangdong, China. E-mail: j002886@wyu.edu.cn
Lei Zhou, School of Electronics and Information Engineering, Wuyi University, Jiangmen 529020, Guangdong, China. E-mail zhoulei@gml.ac.cn

Please cite this article as doi: 10.4081/jae.2026.2095

Submitted: 10 January 2026
Accepted: 27 August 2026

YOLO-Banana-Seg: a lightweight and efficient model for rapid segmentation of banana bunches and stalks in complex orchards

Runbo Fu,[1,2] Fuqin Deng,[2] Zhijie Wu,[2] Lanhui Fu,[2] Lei Zhou[2,3]

[1]College of Artificial Intelligence and Low-Altitude Technology, South China Agricultural University, Guangzhou
[2]School of Electronics and Information Engineering, Wuyi University, Jiangmen
[3]Guangdong Laboratory of Artificial Intelligence and Digital Economy (SZ), Shenzhen, China

Corresponding author: Lanhui Fu, Lei Zhou, School of Electronics and Information Engineering, Wuyi University, Jiangmen 529020, Guangdong, China. E-mail: j002886@wyu.edu.cn, zhoulei@gml.ac.cn

Contributions: Runbo Fu, conceptualization, methodology, investigation; **Fuqin Deng**, investigation; **Zhijie Wu**, resources; **Lanhui Fu**, supervision; **Lei Zhou**, investigation, writing – review & editing. All authors read and approved the final version of the manuscript and agreed to be accountable for all aspects of the work.

Conflict of interest: the authors declare no conflict of interest.

Availability of data and materials: All data generated or analyzed during this study are included in this published article.

Funding: This research was funded by the Research on Key Technologies of Laser Processing Control System (HX25084), the Ph.D. Research Start-Up Fund of Wuyi University (BSQD2222), the National Natural Science Foundation of China Project (62073274), the Wuyi University, Hong Kong and Macau Joint Funding Scheme (2022WGALH17).

Abstract

In banana orchards, accurate segmentation of bunches and stalks is vital for automated harvesting and yield estimation. However, color similarity between green bananas and leaves, irregular shapes, and occlusion makes instance segmentation highly challenging. This study proposes an improved YOLOv11-based segmentation model, YOLO-Banana-

Seg, for banana bunches and stalks. We compiled a multi-condition banana dataset encompassing varying illumination intensities and occlusion levels, while the proposed model integrates multi-scale feature fusion and adaptive attention mechanisms to significantly improve occluded target segmentation accuracy. The results show that YOLO-Banana-Seg achieves high segmentation accuracy while reducing parameters and computational complexity, ensuring its applicability in resource-limited scenarios. Experiments demonstrate 97.4% accuracy and 82.9% recall with 16.7% fewer parameters than YOLOv11n, balancing precision and efficiency. The model's high precision and efficiency directly could support the development of harvesting robots and yield estimation systems for agricultural applications.

Key words: banana segmentation; orchard; stalk segmentation; YOLOv11n-seg.

Introduction

Banana is one of the most significant crops in the global economy, cultivated in over 130 countries; in 2022, China alone produced 11.7 million tons on 326.85 thousand hectares. However, the agro-food sector is facing critical labor shortages, making automated harvesting an urgent priority (Ryan, 2023). In complex orchard environments, fast and precise identification and segmentation of banana bunches and stalks are crucial for tasks like maturity assessment, yield prediction, and obstacle navigation. Yet, the color similarity between green bananas and leaves, irregular shapes, occlusion, and uneven lighting make instance segmentation particularly difficult. Deep learning has driven major advances in smart agriculture; two-stage models like R-CNN (Girshick *et al.*, 2014), Fast R-CNN, and Faster R-CNN (Ren *et al.*, 2017) have shown excellent detection performance. For example, Wang *et al.* (2021) proposed an attention-enhanced Faster R-CNN that achieved 98.5% mAP for young tomato detection under near-color backgrounds, but its 84ms per image inference speed underscores the real-time limitations of such models.

To address the need for faster and more efficient detection, single-stage models such as YOLO (Redmon *et al.*, 2016) and EfficientDet (Tan *et al.*, 2020) have become prominent, with EfficientDet combining EfficientNet and BiFPN for an optimal accuracy-efficiency balance. For banana, YOLO-Banana (Fu *et al.*, 2022b) achieved accurate bunch and stalk detection, and subsequent work explored the effect of input resolution (Fu *et al.*, 2022a) and leveraged depth data for 3D stalk positioning and yield estimation (Zhou *et al.*, 2024). Similar YOLO-based improvements have also been reported for other fruits, such as mango (Koirala *et al.*, 2019; Li H. *et al.*, 2024), litchi (Peng *et al.*, 2025), and citrus (Deng *et al.*, 2025), with further enhancement achieved via attention mechanisms (Liu Q. *et al.*, 2024).

Instance segmentation, which combines object detection and semantic segmentation, provides pixel-level boundaries of individual objects, making it valuable for precision agriculture. In banana-related studies, Wu *et al.* (2023) combined DeepLabv3+ with traditional image processing to finely segment bunches and count them during debudding and harvest, achieving high counting accuracy. Chen *et al.* (2021) developed a lightweight multi-feature fusion network for banana stalk segmentation, improving accuracy and efficiency, while He *et al.* (2023) improved DeepLabv3+ for banana crown segmentation, significantly reducing parameters and improving speed. For other crops, instance segmentation has also been extensively investigated. Yu *et al.* (2019) used Mask R-CNN with ResNet50 for strawberry detection, and Lu *et al.* (2023) proposed Mask Positioner for green fruit. Cheng *et al.* (2024) introduced a U-Net with EMA attention for peach orchard mapping, while Li *et al.* (2025) designed DS-UNet for kiwifruit and its stem segmentation. Multi-task and YOLO-based models have been applied to tomato, including MTA-YOLACT (Li *et al.*, 2023) for cluster and peduncle segmentation, TPS-YOLO (Liang *et al.*, 2025) for pruning point positioning, Y-HRNet (Liu M. *et al.*, 2024) for ripeness segmentation, and YOLOv8s-Seg (Yue *et al.*, 2023) for real-time fruit and surface feature segmentation. Wang *et al.* (2023) similarly used YOLOv8-Seg for litchi fruit and branch segmentation to construct an obstacle-avoidance picking system. Accurate detection of picking points and fruits is a prerequisite for orchard robots to correctly plan routes and avoid obstacles (Chen *et al.*, 2024; Li Z. *et al.*, 2024; Tang *et al.*, 2024; Ye *et al.*, 2023). However, in green banana harvesting, the strong color resemblance with leaves, combined with the thin and irregular stalks, still challenges existing networks in balancing accuracy and speed.

Bananas are harvested while still green, resulting in strong color similarity among stalks, bunches, and leaves, a challenge further compounded by complex backgrounds and irregular shapes. These factors make real-time segmentation extremely difficult, and existing networks still struggle to achieve both high accuracy and fast speed. Although YOLOv11-seg provides an efficient baseline for real-time instance segmentation, current methods still fall short in meeting the stringent accuracy–speed balance required in practice. To this end, the main contributions of this work are as follows:

(1) A selective multi-path fusion neck network was proposed, incorporating a CFM(Cross-scale Fusion Model) to achieve selective feature screening and efficient fusion of deep semantic information with shallow localization details, while reducing the number of parameters.

(2) An efficient segment network, YOLO-Banana-Seg, was proposed for banana precision detection in the orchard. This network improved the MLCA(Mixed Local Channel Attention) backbone, SMFN(Selective Multi-path Fusion Network) neck, and CAA(Context Anchor Attention) to enhance feature extraction and segmentation accuracy based on YOLOv11-seg model.

(3) Through model comparison experiments, the optimal backbone and neck network were selected, and the CAA attention mechanism model was determined. Ablation experiments validated the effectiveness of the proposed method, comparing the segmentation accuracy of banana bunches and stalks between the improved network and other segmentation networks.

Materials and Methods

Gathering data and processing

The dataset consists of images taken from a banana plantation located at the Guangdong Academy of Agricultural Sciences in Guangdong Province, China. A Canon sx610hs digital color camera, with a resolution of 2048 × 1536 pixels and automatic exposure settings, was utilized to capture the images. The photographs were taken at a distance ranging from approximately 800 to 1200 mm and stored in JPG format. To boost the model's ability to generalize and remain robust, data was gathered under a range of lighting conditions, from sunny to cloudy, at various times throughout the day. The annotation process was carried out using Labelme, with a focus on identifying banana bunches and stalks. To ensure annotation precision, particularly for regions with ambiguous boundaries, cross-validation and multiple annotators were employed. The labeled data, initially saved in JSON format, were subsequently converted into txt files compatible with YOLOv11-seg. A total of 1520 images were collected, consisting of 816 images captured under sunny conditions and 704 under cloudy conditions. Using a 7:2:1 ratio, the images were partitioned into training (1064 images: 574 sunny and 490 cloudy), validation (304 images), and test sets (152 images). The dataset construction workflow is illustrated in Figure 1.

Banana instance segmentation based on improved YOLOv11-seg

YOLOv11-seg adopts innovative C3k2 and C2PSA models, with its backbone network utilizing small convolutional kernels to enhance fine-grained feature extraction. The C2PSA model leverages a spatial attention mechanism to address occlusion challenges, while the neck network integrates multi-scale information. The head network combines CBS and NMS to achieve efficient and accurate predictions. Building on the strengths of YOLOv11-seg, we propose YOLO-Banana-Seg, an enhanced network specifically designed for banana bunches and stalks segmentation in complex agricultural environments. As showed in Figure 2, the backbone network replaces the original C3k2 with the C3k2-MLCA model to strengthen feature extraction capabilities. The neck network is upgraded to the SMFN model, which more effectively integrates low-level and high-level features through adaptive weighting, significantly improving the model's ability to handle irregular shapes and dense distributions of banana clusters. Furthermore, the head network incorporates a CAA mechanism to refine segmentation precision. The improved YOLO-Banana-Seg achieves higher segmentation accuracy while notably reducing parameter count and computational complexity. Experimental validation demonstrates its superior performance in occlusion handling and edge detail preservation compared to existing methods, making it particularly suitable for agricultural automation scenarios with limited computational resources.

Figure 3 summarizes the overall framework adopted in this work and highlights the key modules of the proposed approach. In the remainder of this paper, we systematically detail the refinements made to each component and discuss how these modifications contribute to improved performance.

Backbone improvement

The C3k2 model is the main component of YOLOv11's backbone network, which is used for feature extraction. In this study, we introduce a novel feature extraction model, C3k2-MLCA (Figure), to enhance the network's performance. The C3k2-MLCA model was developed based on the C3k2 model, and it is capable of integrating both channel and spatial information, as well as local and global details, to enhance the network's representational capacity. The MLCA improves the network's capability to extract and prioritize relevant spatial and channel-specific features. The MLCA principle, as described by (Wan *et al.*, 2023), involves two key steps of pooling and fusion to achieve mixed attention.

(1) Local and Global Feature Extraction.

The input feature vectors undergo two stages of pooling. The first stage extracts local spatial information by transforming the input into a vector of size $L \times L \times C$, where C denotes the number of channels and L represents the kernel size. This is followed by two branches: one capturing global information and the other preserving local spatial details. These branches are then processed through one-dimensional convolution and anti-pooling to restore the original resolution, enabling effective fusion of global and local features.

(2) Channel Interaction and Attention.

The MLCA mechanism adopts a one-dimensional convolution with a kernel size L that scales with the channel dimension C, as expressed in the formula:

$$L = \phi(C) = \left| \frac{log_2(C)}{\gamma} + \frac{b}{\gamma} \right|_{obb} \tag{1}$$

where, the hyperparameters γ and b are assigned a default value of two, and L is constrained to odd values. This ensures that the network captures cross-channel interactions while maintaining computational efficiency.

By combining global and local spatial information, MLCA improves the network's ability to distinguish between overlapping banana bunches and banana stalks. The adaptive pooling strategies allow the network to handle the irregular and dense growth patterns of banana plants effectively, as shown in Figure . In our experiments, the MLCA mechanism demonstrated superior performance in segmenting banana bunches and stalks.

Neck improvement

The neck network of YOLO-Banana-Seg employs the Selective Multi-path Fusion Network (SMFN), which is an improved version of the HSFPN proposed in MFDS-DETR by (Chen Y. *et al.*, 2024). As shown in Figure 2, SMFM incorporates the CAA attention mechanism to selectively screen features from the input feature maps. Through multi-path fusion, SMFN boosts the network's capacity to recognize multi-scale objects, addressing the size variation between banana bunches and stalks and adapting more effectively to segmentation needs.

SMFN, through selective feature screening and multi-scale feature fusion, greatly enhances the model's segmentation effectiveness in intricate orchard conditions. In SMFN, the nn.Transpose2d model is used to achieve bottom-up deep feature transmission, enabling fusion with shallow features. Meanwhile, the nn.Conv2d model performs downsampling on deep features and concatenates them with high-resolution shallow features to progressively extract high-level semantic information while compressing the feature maps, thereby improving computational efficiency.

In the process of feature extraction, deep-layer features encompass rich semantic information but exhibit coarse localization, whereas shallow-layer features provide precise target localization. Common solutions merely stack feature layers without performing feature selection. To address this, we propose a Cross-scale Fusion Model (CFM). The CFM transforms significant features into attention weights through the CAA model, effectively filtering out low-level features and establishing an efficient cross-scale information transmission mechanism. This ensures comprehensive integration of deep semantic information and shallow localization details, thereby enhancing the accuracy of instance segmentation. Given an input low-level feature $f_{low} \in R^{C\times H\times W}$ and an input high-level feature $f_{high} \in R^{C\times H_1\times W_1}$. the CAA model assigns weights to the feature maps. To balance the dimensions of high-level and low-level features, the low-level feature undergoes downsampling using Conv2d, producing $f_{\widehat{low}} \in R^{C\times H_2\times W_2}$. Finally, the low-level and high-level features are fused to enhance segmentation precision, to enhance the model's feature representation and obtain $f_{out} \in R^{C\times H\times W}$. Equations (2) and (3) illustrate the fusion process for feature selection:

$$f_{\widehat{low}} = Conv2d\big(CAA(f_{low})\big) \tag{2}$$

$$f_{out} = f_{\widehat{low}} * \Big(CAA\big(f_{high}\big)\Big) + f_{high} \tag{3}$$

The CFM model enhances feature maps with attention information, integrating features of different sizes to achieve deep-shallow feature fusion. This approach leverages shallow details for precise localization of banana bunches and stalks while utilizing deep-level information to ensure boundary completeness. Compared to the original network, employing SMFN as the neck network in YOLO-Banana-Seg significantly improves accuracy while slightly reducing the number of parameters.

Attention model

Based on the original channel attention mechanism of SMFN, the CAA attention mechanism is incorporated. By utilizing global average pooling and fully connected layers to generate channel weights, the model optimizes feature selection through dynamic weight allocation. The feature selection mechanism effectively focuses on key instance regions, providing advantages in banana bunch segmentation under leaf occlusion. CAA is capable of capturing contextual dependencies between distant pixels while enhancing central features. Its structure is illustrated in Figure 4. First, global pooling is used on the input feature map to capture global contextual information., followed by a 1×1 convolution to obtain local region features. Unlike the fully connected layers (FC) in the SE model, CAA adopts depthwise separable convolution (H-

Conv and V-Conv) to reduce the number of parameters. They consists of depthwise convolution, which extracts features within each channel while preserving channel independence, and pointwise convolution, which fuses information across different channels to improve inter-channel correlation. Finally, a 1×1 convolution further integrates features, enhancing information interaction. The Sigmoid activation function generates channel weights, which are applied to the final feature map to weight each channel accordingly, thereby improving the model's perception of banana bunch and stalk regions and enhancing segmentation accuracy. CAA can better adapt to complex orchard environments with variations in lighting and leaf occlusion, increasing the model's robustness and attention to small targets such as banana stalks.

Results

Experiment environment

The banana dataset was used to assess the performance of each model, followed by a sequence of experiments. All tests were run on a computing platform with 32GB of RAM, an Intel i9 processor, an NVIDIA GeForce RTX 3090 GPU, and Ubuntu 16.04. PyTorch 1.8.1, combined with CUDA 11.1, was utilized for deep learning tasks. Consistent training datasets were utilized across all models to ensure fairness in comparison. Additionally, the training parameters for each model were aligned with those used in the YOLO-Banana-Seg framework to maintain uniformity in the experimental setup.

Model evaluation

In this research, the enhanced YOLOv11n-seg model was evaluated based on several key metrics, including precision, recall, segmentation mean average precision at IoU thresholds of 50% (mAP@50) and 50-95% (mAP@50-95), as well as computational efficiency measured in giga floating point operations per one second (GFLOPs). The localization accuracy of banana bunches and stalks was determined using precision and recall, while the quality of segmentation outcomes was analyzed through segment mAP scores. The calculations for precision, recall, average precision (AP), segment mAP, and

GFLOPs were performed using Equations (4) to (8).

$$P = \frac{TP}{(TP + FP)} \times 100\% \tag{4}$$

$$R = \frac{TP}{(TP + FN)} \times 100\% \tag{5}$$

$$AP = \int_0^1 P(R)dR \tag{6}$$

$$mAP = \sum_{i=1}^{n} \frac{AP(i)}{n} \tag{7}$$

$$GFLOPs = O\left(\sum_{i=1}^{n} K_i^2 \times C_{i-1}^2 \times C_i^2 + \sum_{i=1}^{n} M^2 \times C_i\right) \tag{8}$$

where, TP denotes to a correctly predicted positive sample, while FP represents a negative sample incorrectly predicted as positive, and FN denotes a positive sample incorrectly predicted as negative. AP quantifies the segmentation accuracy, with higher AP scores indicating better segmentation performance. The variable n corresponds to the number of categories in the segmentation task. O is a constant term representing the computational order. K stands for the size of the convolutional kernel. C indicates the number of channels. i refers to the iteration count. M represents the dimensions of the input image.

Experimental validation

The segmentation performance of banana bunches and banana stalks in the test set was evaluated by comparing the YOLO-Banana-Seg model with the YOLOv11n-seg model. The comparison results are presented in Table 1. From this table, it is evident that the YOLOv11n-seg model achieves AP values of 82% for banana bunches and 92.1% for stalks, which are 0.7% and 1.2% lower, respectively, than those achieved by YOLO-Banana-Seg. Additionally, the mAP_M@50 score of YOLOv11n-seg is 1.0% lower compared to YOLO-Banana-Seg. Furthermore, the mAP_M@50-95 score of YOLOv11n-seg is 1.2% lower than that of YOLO-Banana-Seg, which attains a score of 64.7%.

Additionally, The AP value for banana stalks segmented by two models is greatly higher than that for banana bunches. This is primarily because the area of bunches is generally larger than that of stalks. Larger objects are more susceptible to local errors in IoU computation, which can lead to a reduction in AP. In contrast, the smaller area of stalks makes them less affected by local errors, resulting in better performance. Meanwhile, the internal texture of bunches is often more complex, introducing noisy features that increase the difficulty for models to accurately delineate their boundaries. Comparing the number of parameters of the two models, the parameters of the YOLO-Banana-Seg network decreased significantly from 2,834,958 to 2,362,466, resulting in a reduction of 472,492. This corresponds to a 16.67% decrease compared to the original network.

The performance of YOLOv11n-seg and YOLO-Banana-Seg under various conditions. Under sunny conditions, as shown in

Figure 5, YOLOv11n-Seg misjudges leaves as part of the banana bunch when the stalk and petiole are close together. This misjudgment occurs because the color and shape of the banana petiole closely resemble those of the stalk. In contrast, under cloudy conditions, as depicted in Figure 6, both YOLOv11n-seg and YOLO-Banana-Seg successfully segment banana bunches and banana stalks. However, YOLOv11n-seg erroneously classifies a nearby petiole as a stalk, leading to inaccuracies. When leaves partially obscure the target, the precision of distinguishing target edges becomes a critical factor in evaluating segmentation performance. Figure 7 highlights that YOLOv11n-seg exhibits larger misjudged areas compared to YOLO-Banana-Seg.

The data in the Table 1 further supports these observations. YOLO-Banana-Seg achieves higher accuracy in both banana and stalk detection, with a mAP@50 of 88.0% compared to YOLOv11n-seg 87.02%. Additionally, YOLO-Banana-Seg shows better performance in mAP@50-95, indicating superior boundary localization. Overall, YOLO-Banana-Seg outperforms in both accurate target identification and precise boundary localization.

Statistical analysis

All statistical analyses were conducted in Python 3.8 using the SciPy and statsmodels libraries. To evaluate the reliability of observed performance differences, bootstrap resampling (1,000 iterations) was performed on the test set predictions, generating distributions of metrics for each model. For pairwise comparisons between two models (Table 1), the Wilcoxon signed-rank test was applied to the paired bootstrap metric differences, with the resulting p-values reported. When comparing multiple models, backbones, necks, or attention mechanisms (Tables 2 to 6), the Friedman test was first used to detect overall differences; if significant (p < 0.05), pairwise Nemenyi post-hoc tests were conducted. To control the false discovery rate arising from multiple pairwise comparisons, the Benjamini-Hochberg procedure was applied, and adjusted p-values (q-values) were reported where appropriate. A p-value <0.05 (or q <0.05 after correction) was considered statistically significant. Throughout all tables, different lowercase letters in the same column indicate statistically significant differences based on the tests described above. Performance metrics are presented as single-run evaluation values, with 95% bootstrap confidence intervals shown in brackets where applicable. Parameters and GFLOPs are deterministic hardware-independent quantities and are not subject to statistical testing.

Discussion

Comparative analysis of different models

In order to impartially assess the performance of the improved model on our dataset, a series of comparative experiments were carried out using various segmentation models from the YOLO series. The models evaluated in this study encompass YOLOv5n-seg, YOLOv5s-seg, YOLOv8n-seg, YOLOv8s-seg, YOLOv11s-seg, YOLOv11n-seg, and the proposed improved network, YOLO-Banana-Seg. Each model was trained and tested on the banana dataset under identical conditions. The evaluation criteria included precision (P), recall (R), mask mean average precision (mAP_M), parameter count, and computational complexity measured in GFLOPs. The detailed comparison of the results across these models is presented in Table 2.

It can be observed that YOLOv11n-seg has the lowest number of parameters and GFLOPs while maintaining a balanced accuracy. The lightweight nature and balanced performance of YOLOv11n-seg make it an ideal foundation for improving the network, enabling enhanced performance while maintaining efficiency. Compared to other networks, the improved YOLO-Banana-Seg network demonstrates significant advantages in segmentation accuracy. Among networks of the n-level, the YOLO-Banana-Seg network achieves P, R, mAP_M@50 and mAP_M@50-95 values of 97.4%, 82.9%, 88%, and 64.7%, respectively, outperforming other networks of the same level. The YOLO-Banana-Seg network improves accuracy while reducing the number of parameters and computational complexity. Inheriting the low GFLOPs of the original YOLOv11n-seg network, the GFLOPs of the YOLO-Banana-Seg network only increased by 0.3M. Compared to other networks of the n-level, YOLO-Banana-Seg has a lower number of parameters and GFLOPs. Additionally, its higher segmentation accuracy of 97.44% is a standout feature when compared to other networks, even when compared to s-level networks. The number of parameters and GFLOPs are significantly lower than those of s-level networks. For example, YOLOv8s-seg has 11,790,854 parameters and 42.7M GFLOPs. Although the mAP is close to that of s-level networks, the YOLO-Banana-Seg network has clear advantages in segmentation accuracy, number of parameters, and GFLOPs.

Comparative analysis of backbone architectures

In this study, the backbone network of YOLOv11-seg was replaced with several alternative architectures, including the CMUNeXt block, Context Guided block, DySnakeConv, and MCLA networks, to conduct a series of comparative experiments. The original YOLOv11-seg network served as the baseline for these comparisons.

The YOLOv11n + MLCA network achieved the best recognition accuracy, as evidenced by Table 3, with mAP_B@50 and mAP_M@50 values of 87.5% and 87.1%, respectively. Notably, its computational cost, measured in GFLOPs, remained comparable to the original model at 10.2M. In contrast, the YOLOv11n + CMUNeXt block, YOLOv11n + Context Guided block, and YOLOv11n + DySnakeConv configurations exhibited the highest computational complexity, with GFLOPs reaching 11.6. However, these modifications resulted in a reduction in accuracy, as reflected by lower mAP_B@50 and mAP_M@50 scores compared to the baseline YOLOv11-seg model.

Comparative analysis of neck architectures

To conduct comparative experiments, we replace the neck network of YOLOv11n-seg with BiFPN, ASF, HSFPN, and SMFN, using YOLOv11n-seg as the benchmark. According to Table 4, YOLOv11n + HSFPN shows the most substantial parameter reduction, decreasing by about 30% compared to YOLOv11n-seg. Building on this, the refined SMFN improves accuracy while increasing the parameter count by approximately 7%, with GFLOPs rising by only 0.3.

Comparative analysis of attention mechanisms

In this study, the YOLOv11n-seg network was utilized as a baseline to evaluate the impact of integrating various attention mechanisms, including CAFM, SEAM, MSAA, and CAA, positioned before the head network. The objective was to identify the attention mechanism that delivers the highest recognition accuracy.

As demonstrated in Table 5, the incorporation of different attention mechanisms yielded varying results. Specifically, the SEAM and MSAA mechanisms led to a slight reduction in accuracy compared to the original network. In contrast, the CAFM mechanism provided a marginal enhancement in mAP_B@50. The most notable improvement was observed with the CAA attention mechanism, where mAP_B@50 increased significantly from 87.2% to 87.6%. Furthermore, mAP_M@50 saw an increment of 0.2, with only a minimal rise in computational cost, as GFLOPs increased by just 0.2M.

Ablation studies on various enhanced model configurations

A range of ablation experiments were carried out on the upgraded YOLOv11n-seg model to evaluate its segmentation performance on the banana dataset. The results of these experiments are detailed in Table 6. By substituting the neck network of YOLOv11n-seg with the SMFN network, the computational cost, measured in GFLOPs, decreased by 0.6 M, and the parameter count was reduced by 17% compared to the original model.

This improvement can be attributed to the SMFN network's utilization of the CFM model, which optimizes efficiency by sharing weights across certain convolutional layers in both the top-down and bottom-up pathways. Additionally, the attention model within the SMFN network possesses the capability of dynamic feature filtering, with the filtering networks at different levels sharing part of the attention generators. Replacing the original attention mechanism of the SMFN network with the CAA model enables more precise localization of key feature channels and noise suppression through depthwise separable convolution and dynamic channel weighting. This replacement significantly improves segmentation accuracy while increasing the number of parameters by only 9%, which remains lower than the baseline model.

However, the segmentation accuracy reaches 97.4%, demonstrating a substantial improvement. Replacing the C3k2 model in YOLOv11n-Seg with C3k2-MLCA enhances the backbone network. The adaptive pooling strategy, which combines global and local spatial information, enhances the feature map channels critical to the current task. A notable accuracy improvement is observed when integrating the SMFN neck network, the CAA attention mechanism, and the enhanced C3k2-MLCA model into the YOLOv11n-seg network. Specifically, mAP_B@50, mAP_B@50-95, P_M, mAP_M@50, and mAP_M@50-95 achieve values of 87.5%, 65.4%, 97.4%, 88%, and 64.7%, respectively. These results represent improvements of 0.3%, 1.6%, 1.7%, 1.0%, and 1.2% compared to the YOLOv11n-seg model. Additionally, the total number of network parameters is reduced by 17%, while the computational cost, measured in GFLOPs, increases by only 0.3M. The experimental data analysis confirms the effectiveness of the proposed improved model on the banana dataset used in this study.

Conclusions

The primary goal of this research was to achieve precise real-time instance segmentation of banana bunches and stalks in complex orchard environments. To this end, an enhanced YOLO-Banana-Seg network was developed based on the YOLOv11n-seg framework, incorporating a Mixed Local Channel Attention (MLCA) backbone, a Selective Multi-path Fusion Network (SMFN) neck, and a Context Anchor Attention (CAA) mechanism. Compared to the baseline, the proposed network achieves a 1.6% increase in mAP_B@50-95 and a 1.2% increase in mAP_M@50-95, while reducing the total parameter count by 17%, with only a marginal increase of 0.3 GFLOPs in computational cost. It attains a frame rate of 35.5 fps, meeting real-time requirements, together with a segmentation precision of 97.4% and a recall of 82.9%.

Beyond these performance gains, this study offers several theoretical insights: first, selective multi-path feature fusion effectively reduces redundant cross-scale connections while preserving fine localization details for slender stalks; second, mixed local-channel attention enhances discrimination between green bananas and leaves under near-color backgrounds; third, context anchor attention strengthens boundary perception for irregular, thin structures. These design principles can inform similar segmentation tasks for other fruit crops. From a practical perspective, the lightweight architecture enables deployment on edge devices for in-field, real-time harvesting guidance. Accurate segmentation of bunches and stalks allows robots to precisely locate picking points, plan obstacle-avoidance paths, and support yield estimation and maturity assessment. This directly benefits producers by reducing dependence on manual labor, lowering harvesting damage, and improving orchard management efficiency.

In future work, the segmentation model will be extended to extract real-world 3D coordinates of banana bunches and stalks, enabling automated spatial localization and further contributing to intelligent banana orchard management.

References

Chen M, Chen Z, Luo L, Tang Y, Cheng J, Wei H, et al., 2024. Dynamic visual servo control methods for continuous operation of a fruit harvesting robot working throughout an orchard. Comput Electron Agric 219:108774.

Chen T, Zhang R, Zhu L, Zhang S, Li X, 2021. A method of fast segmentation for banana stalk exploited lightweight multi-feature fusion deep neural network. Machines 9:66.

Chen Y, Zhang C, Chen B, Huang Y, Sun Y, Wang C, et al., 2024. Accurate leukocyte detection based on deformable-DETR and multi-level feature fusion for aiding diagnosis of blood diseases. Comput Biol Med 170:107917.

Cheng J, Zhu Y, Zhao Y, Li T, Chen M, Sun Q, et al., 2024. Application of an improved U-net with image-to-image translation and transfer learning in peach orchard segmentation. Int J Appl Earth Obs Geoinf 130:103871.

Deng F, He Z, Fu L, Chen J, Li N, Chen W, et al., 2025. A new maturity recognition algorithm for Xinhui citrus based on improved YOLOv8. Front Plant Sci 16:1472230.

Fu L, Wu F, Zou X, Jiang Y, Lin J, Yang Z, et al., 2022a. Fast detection of banana bunches and stalks in the natural environment based on deep learning. Comput Electron Agric 194:106800.

Fu L, Yang Z, Wu F, Zou X, Lin J, Cao Y, et al., 2022b. YOLO-banana: a lightweight neural network for rapid detection of banana bunches and stalks in the natural environment. Agronomy 12:391.

Girshick R, Donahue J, Darrell T, Malik J, 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), Columbus; pp. 580-587.

He J, Duan J, Yang Z, Ou J, Ou X, Yu S, et al., 2023. Method for segmentation of banana crown based on improved DeepLabv3+. Agronomy 13:1838.

Koirala A, Walsh KB, Wang Z, McCarthy C, 2019. Deep learning for real-time fruit detection and orchard fruit load estimation: benchmarking of 'MangoYOLO'. Precis Agric 20:1107-35.

Li H, Huang J, Gu Z, He D, Huang J, Wang C, 2024. Positioning of mango picking point using an improved YOLOv8 architecture with object detection and instance segmentation. Biosyst Eng 247:202-20.

Li L, Li K, He Z, Li H, Cui Y, 2025. Kiwifruit segmentation and identification of picking point on its stem in orchards. Comput Electron Agric 229:109748.

Li Y, Feng Q, Liu C, Xiong Z, Sun Y, Xie F, et al., 2023. MTA-YOLACT: multitask-aware network on fruit bunch identification for cherry tomato robotic harvesting. Eur J Agron 146:126812.

Li Z, Wang D, Zhu T, Tao Y, Ni C, 2024. Review of deep learning-based methods for non-destructive evaluation of agricultural products. Biosyst Eng 245:56-83.

Liang X, Wei Z, Chen K, 2025. A method for segmentation and localization of tomato lateral pruning points in complex environments based on improved YOLOv5. Comput Electron Agric 229:109731.

Liu M, Chen W, Cheng J, Wang Y, Zhao C, 2024. Y-HRNet: research on multi-category cherry tomato instance segmentation model based on improved YOLOv7 and HRNet fusion. Comput Electron Agric 227:109531.

Liu Q, Lv J, Zhang C, 2024. MAE-YOLOv8-based small object detection of green crisp plum in real complex orchard environments. Comput Electron Agric 226:109458.

Lu Y, Ji Z, Yang L, Jia W, 2023. Mask positioner: an effective segmentation algorithm for green fruit

Peng H, Li Z, Zou X, Wang H, Xiong J, 2025. Research on litchi image detection in orchard using UAV based on improved YOLOv5. Expert Syst Appl 263:125828.

Redmon J, Divvala S, Girshick R, Farhadi A, 2016. You only look once: unified, real-time object detection. Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), Las Vegas; pp. 779-788.

Ren S, He K, Girshick R, Sun J, 2017. Faster R-CNN: towards real-time object detection with region proposal networks. IEEE Trans Pattern Anal Mach Intell 39:1137-49.

in complex environment. J King Saud Univ Comput Inf Sci 35:101598.

Ryan M, 2023. Labour and Skills Shortages in the Agro-Food Sector. OECD Food, Agriculture and Fisheries Papers, No. 189. Paris, OECD Publishing.

Tan M, Pang R, Le QV, 2020. EfficientDet: scalable and efficient object detection. In: Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). Seattle, WA, USA, IEEE; pp. 10778-87.

Tang Y, Qi S, Zhu L, Zhuo X, Zhang Y, Meng F, 2024. Obstacle avoidance motion in mobile robotics. J Syst Simul 36:1.

Wan D, Lu R, Shen S, Xu T, Lang X, Ren Z, 2023. Mixed local channel attention for object detection. Eng Appl Artif Intell 123:106442.

Wang C, Li C, Han Q, Wu F, Zou X, 2023. A performance analysis of a litchi picking robot system for actively removing obstructions, using an artificial intelligence algorithm. Agronomy 13:2795.

Wang P, Niu T, He D, 2021. Tomato young fruits detection method under near color background based on improved Faster R-CNN with attention mechanism. Agriculture 11:1059.

Wu F, Yang Z, Mo X, Wu Z, Tang W, Duan J, et al., 2023. Detection and counting of banana bunches by integrating deep learning and classic image-processing algorithms. Comput Electron Agric 209:107827.

Ye L, Wu F, Zou X, Li J, 2023. Path planning for mobile robots in unstructured orchard environments: an improved kinematically constrained bi-directional RRT approach. Comput Electron Agric 215:108453.

Yu Y, Zhang K, Yang L, Zhang D, 2019. Fruit detection for strawberry harvesting robot in non-structural environment based on Mask R-CNN. Comput Electron Agric 163:104846.

Yue X, Qi K, Na X, Zhang Y, Liu Y, Liu C, 2023. Improved YOLOv8-seg network for instance segmentation of healthy and diseased tomato plants in the growth stage. Agriculture 13:1643.

Zhou L, Yang Z, Deng F, Zhang J, Xiao Q, Fu L, et al., 2024. Banana bunch weight estimation and stalk central point localization in banana orchards based on RGB-D images. Agronomy 14:1123.

Figure 1. Dataset acquisition in banana orchards.

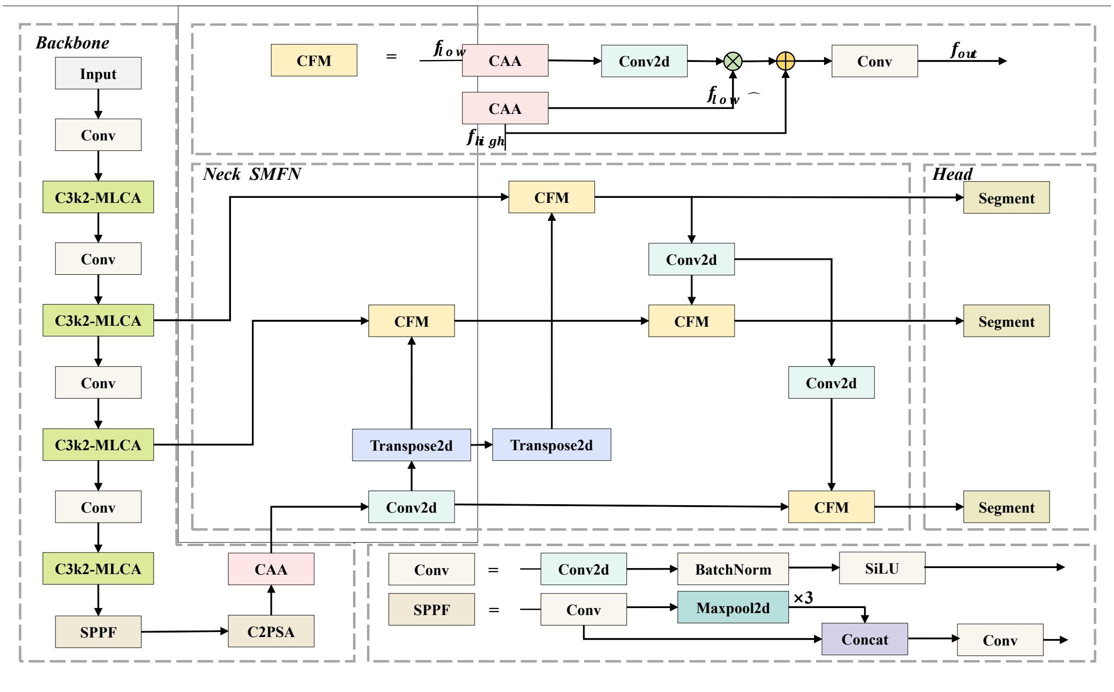


Figure 2. YOLO-Banana-Seg-based segmentation process of banana bunches and stalk.

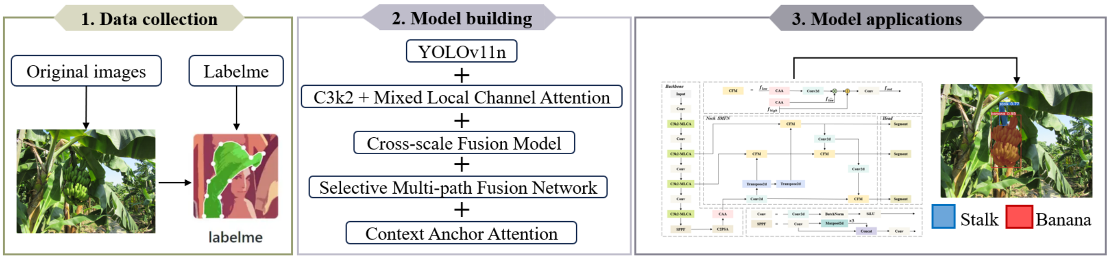


Figure 3. Overview of the study workflow.

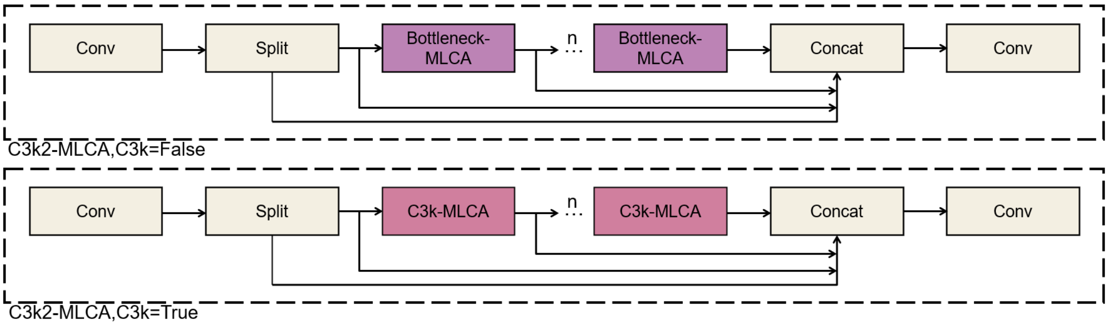


Figure 4. C3k2-MLCA structure diagram.

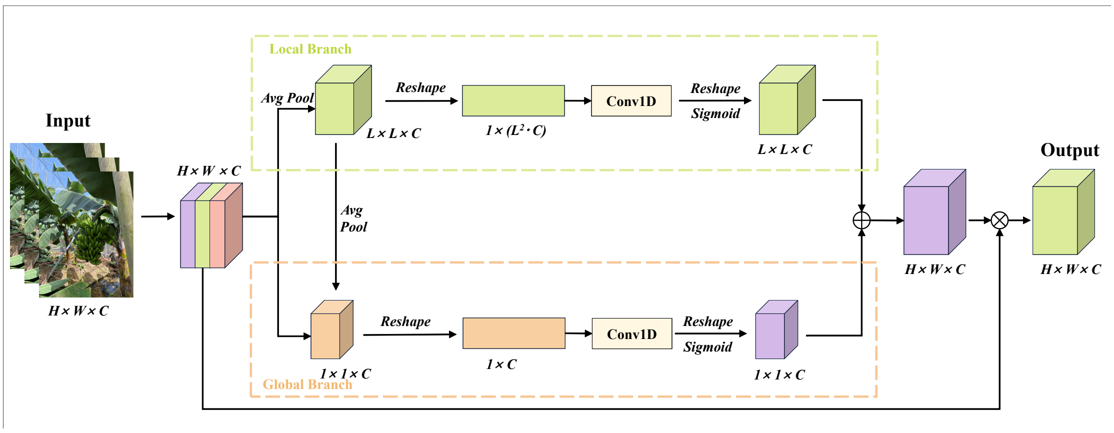


Figure 5. Mixed Local Channel Attention (MLCA).

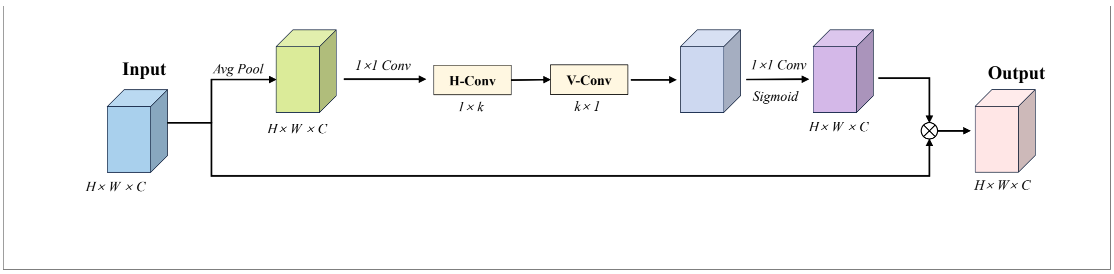


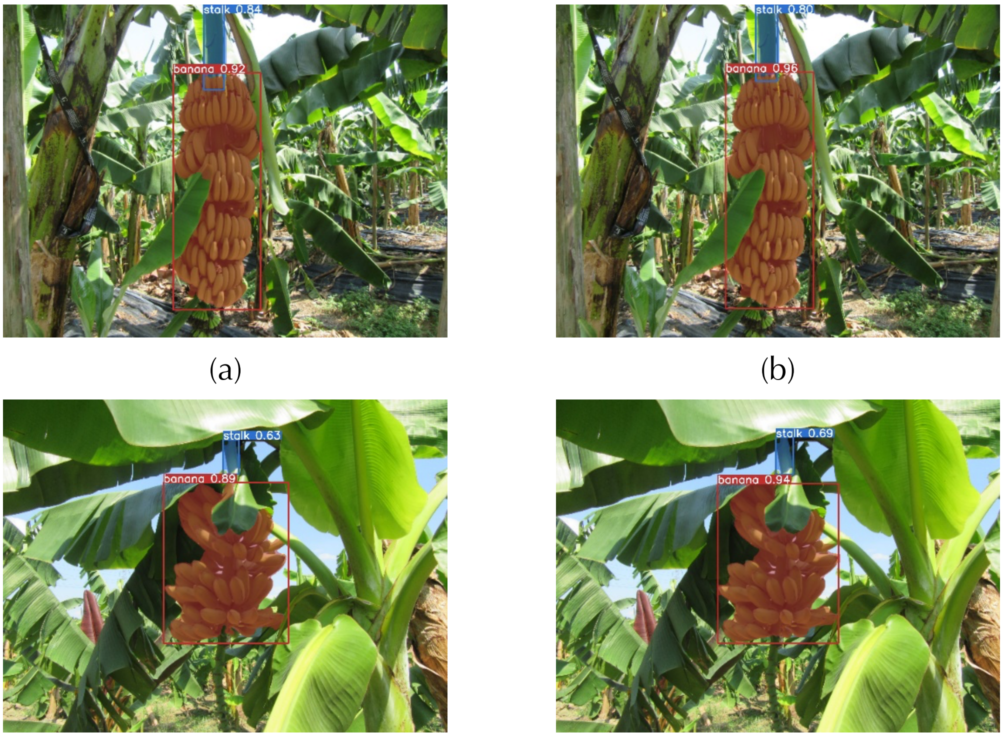
Figure 4. Context Anchor Attention (CAA).

(a) (b)

Figure 5. A comparison under sunny conditions shows: (a) YOLOv11n-seg misclassified leaves as banana bunches, while (b) YOLO-Banana-Seg accurately segmented banana bunches and stalks.

(a) (b)

Figure 6. A comparison under cloudy conditions shows: (a) YOLOv11n-seg misclassified the petiole as part of the banana bunch, while (b) YOLO-Banana-Seg accurately segmented the bunch and stalk.



(a) (b)

(c) (d)

Figure 7. An instance of misclassification occurred for banana leaves positioned at the boundary: the regions incorrectly identified by YOLOv11n-Seg (a and c) are more extensive than those in YOLO-Banana-Seg (b and d).

Table 1. Comparison of the performance with different models

Model	AP		mAP_M@50(%)	mAP_M@50-95(%)	Parameters
	Banana(%)	Stalk(%)			
YOLOv11n-seg	82.0 [b]	92.1 [b]	87.0 [b]	63.5 [b]	2834958
YOLO-Banana-Seg	**82.7** [a]	**93.3** [a]	**88.0** [a]	**64.7** [a]	**2362466**

Different lowercase letters within a column indicate significant difference ($p < 0.05$, Wilcoxon signed-rank test). Parameters are not tested.

Table 2. Evaluation of various model.

Model	P_M(%)	R_M(%)	mAP_M(%)		Parameters	GFLOPs(M)
			50	50-95		
YOLOv5s-seg	96.3	84.1	88.6	66.6	9766326	37.8
YOLOv8s-seg	95.0	85.2	89.2	68	11790854	42.7
YOLOv11s-seg	96.3	83.2	88.2	67	10067598	35.3
YOLOv5n-seg	96.8 [b]	81.1 [c]	86.3 [c]	62.3 [c]	2755750	11.0
YOLOv8n-seg	96 [c]	82.4 [b]	87.4 [b]	63.5 [b]	3258454	12.0
YOLOv11n-seg	95.7 [c]	82.9 [a]	87 [c]	63.5 [b]	2834958	10.2
YOLO-Banana-Seg	**97.4** [a]	**82.9** [a]	**88** [a]	**64.7** [a]	**2362466**	10.5

Letters indicate significant differences (Friedman test + Nemenyi post-hoc; FDR corrected $q < 0.05$). Parameters and GFLOPs are not compared statistically.

Table 3. Evaluation of various backbone.

Models	mAP_B@50(%)	mAP_M@50(%)	GFLOPs(M)
YOLOv11n	87.2 [a]	87.0 [a]	10.2
YOLOv11n + CMUNeXt block	86.4 [b]	86.3 [b]	11.6
YOLOv11n + Context Guided block	87.1 [ab]	86.7 [ab]	11.6
YOLOv11n + DySnakeConv	87.1 [ab]	86.7 [ab]	11.6
YOLOv11n + MLCA	**87.5** [a]	**87.1** [a]	**10.2**

Different letters within a column indicate significant difference ($p < 0.05$, Friedman test + Nemenyi). GFLOPs are not tested.

Table 4. Evaluation of various neck.

Models	mAP_B@50(%)	mAP_M@50(%)	Parameters	GFLOPs(M)
YOLOv11n	87.2 ^{a}	87.0 ^{a}	2834958	10.2
YOLOv11n + BiFPN	86.8 ^{ab}	86.2 ^{ab}	2251251	10.5
YOLOv11n + ASF	86.4 ^{ab}	86.8 ^{ab}	2926617	11.0
YOLOv11n + HSFPN	86.4 ^{b}	85.7 ^{b}	2002558	9.3
YOLOv11n + SMFN	86.8 ^{ab}	86.2 ^{ab}	2154814	9.6

Letters denote significant differences ($p < 0.05$, Friedman test + Nemenyi). Parameters and GFLOPs are not included in hypothesis testing.

Table 5. Evaluation of various attention mechanisms.

Models	mAP_B@50(%)	mAP_M@50(%)	GFLOPs(M)
YOLOv11n	87.2 ^{ab}	87.0 ^{ab}	10.2
YOLOv11n + CAFM	87.3 ^{ab}	86.2 ^{b}	10.3
YOLOv11n + SEAM	86.5 ^{b}	86.0 ^{b}	10.2
YOLOv11n + MSAA	87.0 ^{ab}	86.8 ^{ab}	22.1
YOLOv11n + CAA	**87.6** ^{a}	**87.2** ^{a}	10.4

Different letters within a column indicate significant difference ($p < 0.05$, Friedman test + Nemenyi). GFLOPs are not tested.

Table 6. Ablation experiment.

Model	SMFN	CAA	MLCA	mAP_B(%)		P_M	R_M	mAP_M(%)		Parameters	GFLOPs
				50	50-95	(%)	(%)	50	50-95		(M)
YOLOv11n-seg	×	×	×	87.2 ^{a}	63.8 ^{c}	95.7 ^{c}	82.9 ^{a}	87 ^{b}	63.5 ^{b}	2834958	10.2
	√	×	×	86.8 ^{a}	63.2 ^{c}	96.5 ^{b}	82 ^{b}	86.2 ^{c}	62.5 ^{c}	2154814	9.6
	√	√	×	87.3 ^{a}	63.7 ^{c}	97.4 ^{a}	82.5 ^{a}	87.1 ^{b}	63.5 ^{b}	2362430	10.5
	√	√	√	**87.5** ^{a}	**65.4** ^{a}	**97.4** ^{a}	**82.9** ^{a}	**88** ^{a}	**64.7** ^{a}	2362466	10.5

"×"means the module was not used; "√" means the module was included. Different lowercase letters in the same column denote significant differences (Friedman test + Nemenyi, $q < 0.05$). Parameters and GFLOPs are not compared statistically.