YOLO11-RCSP: CHERRY TOMATO FLOWER POLLINATION STATUS RECOGNITION BASED ON MULTI-SCALE FEATURE FUSION
/
YOLO11-RCSP：融合多尺度特征的圣女果花授粉状态识别

Jianhua CUI, XinYU LI, FuZhong LI, Xiaoying ZHANG

Faculty of Software Technologies, Shanxi Agricultural University, Taigu, Shanxi / China
Corresponding authors: FuZhong Li; Tel: +86-13734008985; E-mail: lifuzhong@sxau.edu.cn
Corresponding authors: Xiaoying Zhang; Tel: +86-15803449361; E-mail: xiaoyingzhang@sxau.edu.cn

DOI: https://doi.org/10.35633/inmateh-79-39

Keywords: *Deep learning, Object detection, YOLO11, Pollination status recognition, Cherry tomato*

ABSTRACT

In greenhouse cultivation of cherry tomatoes, the accurate identification of pollination status is a key prerequisite for achieving automated robotic pollination. As the site of direct interaction during pollination, changes in the colour and morphology of the stigma most accurately indicate whether pollination has been successful. However, due to the minuscule size of the stigma and the extremely subtle visual differences among the three states—unpollinated, pollinated, and fruiting—combined with interference from fluctuating lighting conditions and complex backgrounds in greenhouse environments, traditional methods struggle to provide high-precision visual perception data for pollination robots. To address this issue, this paper proposes an improved model, YOLO11-RCSP, which focuses on the stigma as the core recognition object to achieve automatic classification of the three pollination states: introducing a novel attention mechanism, RFA (Receptive Field Attention), adding an enhancement module (CPA-Enhancer) to the original model, and incorporating SAConv switchable dilated convolutions. In addition, the loss function is improved using the Powerful-IoU method (adaptive penalty factor and gradient adjustment function based on anchor box quality). The precision rate of the improved model reached 88.40%, the recall rate reached 84.60%, and the average precision was 90.80% when the IoU was 0.50. The average precision ranged from 0.50 to 0.95, reaching 63.20%. These metrics represent improvements of 5.20%, 3.30%, 4.40%, and 5.30%, respectively, compared to the original YOLO11 network model, demonstrating stronger feature extraction capabilities and robustness. Ablation experiments further confirm that the attention mechanism and multi-scale optimization have a synergistic effect on enhancing the performance of stigma pollination status recognition. This research will be applied to pistil-level pollination status recognition tasks, helping to advance the development and application of pollination robots, and providing critical technical support for the development of automated pollination equipment in protected agriculture.

摘要

在圣女果温室栽培中，授粉状态的精准识别是实现机器人自动化授粉作业的核心前提。柱头作为授粉过程的直接作用部位，其颜色、形态变化能够最准确地反映授粉是否成功。然而柱头目标尺度微小，未授粉、授粉和结果三类状态的视觉差异极为细微，而且受温室环境中光照变化与复杂背景的干扰，传统方法难以为授粉机器人提供高精度的视觉感知决策依据。针对该问题，本文提出一个改进模型 YOLO11-RCSP，以柱头为核心识别对象，实现三类授粉状态的自动分类：引入一种新型的注意力机制 RFA（感受野注意力）；在原有模型中添加一种增强模块（CPA-Enhancer）；添加 SAConv 可切换空洞卷积；改进损失函数采用 Powerful-IoU(自适应惩罚因子和基于锚框质量的梯度调节函数）的方法。改进后模型的准确率达到 88.40%，召回率达到 84.60%，平均精度在 IoU=0.50 时为 90.80%，在 0.50 至 0.95 的 IoU 平均精度达到 63.20%，各项指标相较于原 YOLO11 网络模型分别提高了 5.20%，3.30%，4.40%，5.30%，拥有更强的特征提取能力和鲁棒性。消融实验进一步证实注意力机制与多尺度优化对提升柱头授粉状态识别性能具有协同效应。本研究将应用于柱头级别的授粉状态识别任务，有助于推动授粉机器人的研发与应用，为设施农业自动化授粉装备的研发提供关键技术支撑。

INTRODUCTION

Cherry tomatoes are an important fruit and vegetable crop widely cultivated in China's protected agriculture, known for their extensive planting area and significant economic benefits. According to statistics, the area under protected tomato cultivation in China has exceeded 1 million hectares, with cherry tomatoes being particularly favoured by the market due to their excellent taste and rich nutritional content. Cherry tomatoes are monoecious plants with unisexual flowers. In natural environments, pollination can be achieved through natural wind or insect vectors; however, in enclosed environments such as greenhouses, the absence of natural wind and insect activity leads to poor pollination, which can result in flower and fruit drop, as well as deformed fruit, severely affecting the yield and quality of cherry tomatoes *(Wang et al., 2016)*. Currently, greenhouse cherry tomato production primarily relies on manual pollination, bumblebee pollination or hormone-dipped flowers. Manual pollination is inefficient, labour-intensive and difficult to control for uniformity; bumblebee pollination is affected by environmental factors such as temperature and humidity within the greenhouse and is relatively costly *(Wang et al., 2013)*; hormone dipping, meanwhile, tends to increase the incidence of deformed fruit, severely affecting the economic efficiency of cherry tomato production. Consequently, pollination robots have emerged as a solution. Pollination robots can significantly reduce labour costs, accelerate pollination speed and improve cherry tomato pollination efficiency. However, the complex growing environment of greenhouse cherry tomatoes—characterised by small, numerous and overlapping flowers—complicates their identification and detection. This represents both a key challenge and a focal point in pollination robot research *(Hou et al., 2024)*.

In recent years, object detection methods based on deep learning have been widely used in crop detection, providing technical support for the precise localization and recognition of crops *(Kang et al., 2023; Hu et al., 2023; Jia et al., 2024)*. In the field of agricultural robot vision, most existing studies focus on fruit picking or flower localization, while studies related to pollination mostly take the whole flower as the recognition unit. To meet the demand for pollination detection of kiwifruit flowers, Pan et al. proposed a lightweight model KIWI-YOLO, which enhances the model’s perception of flower contour details through a frequency-domain feature fusion module, and effectively recognized four states including bud, female flower, male flower, and successful pollination on a self-built kiwifruit flower dataset, with mAP@50 reaching 91.6% *(Pan et al., 2024)*. To solve the detection difficulty caused by the dense distribution of kiwifruit flowers in complex orchard environments, Tang et al. proposed an improved model YOLOv11-TYW based on the YOLOv11n architecture, which verified the applicability of the YOLOv11 architecture in orchard flower object detection tasks and provided a technical reference for flower recognition in subsequent pollination robots *(Tang et al., 2025)*.

Isabel Pinheiro et al. adopted deep learning methods to detect kiwifruit flowers and determine their gender to address the problems that kiwifruit is a dioecious plant with no nectar to attract pollinators and existing technologies lack gender-based flower detection methods; four pre-trained models including YOLOv5, YOLOv8, RT-DETR, and DETR were evaluated, K-fold cross-validation was used to test model performance, and a manually annotated kiwifruit flower gender classification dataset was constructed, among which the YOLOv5s model performed best with a detection accuracy of 97%, providing technical support for the precise operation of automated kiwifruit pollination robots *(Pinheiro et al., 2024)*. Manzoor et al. proposed an improved model based on YOLOv5s; by replacing the backbone network with MobileNetV3 and introducing Ghost convolutions from GhostNet, the model achieved a lightweight effect in apple blossom detection with about 85% reduction in model size and about 89% reduction in GFLOPs, while maintaining an accuracy of 90.6% and mAP@50 of 91.2%, verifying the applicability and lightweight potential of the YOLO series in flower detection for pollination scenarios *(Manzoor et al., 2024)*. Touko Mbouembe et al. proposed an improved YOLOv5s model, which significantly enhanced feature representation ability and detection robustness for occluded objects by integrating the SE attention mechanism, BiFPN multi-scale feature fusion network, CARAFE upsampling operator, and Soft-NMS algorithm *(Touko Mbouembe et al., 2024)*. However, using the whole flower as the recognition target has shortcomings: although the overall size of the flower is large, its morphological characteristics are affected by petal opening angles, variety differences, and environmental factors; the key changes determining successful pollination actually occur at the stigma level, such as stigma colour, surface gloss, and ovary expansion, and these subtle changes are often blocked by petals or disturbed by background in whole-flower images, making them difficult for models to capture accurately. Therefore, methods using the whole flower as the recognition object have obvious deficiencies in the accuracy and interpretability of pollination status recognition.

As the direct site of pollination, the stigma's colour and morphological changes most sensitively reflect whether pollination has been successful. Lorberboym et al. employed a YOLOv8 segmentation model to perform instance segmentation of cannabis stigmas, assessing flower maturity through the proportion of stigma colour, further demonstrating the potential of stigma visual analysis in automated agriculture *(Lorberboym et al., 2026)*.

The Swin Transformer proposed by *Gao et al.* utilises a hierarchical self-attention mechanism to construct robust multi-scale feature representations, offering a new approach to small-object detection in complex visual tasks *(Gao et al., 2024)*. Addressing the issues of low efficiency and high costs associated with manual saffron stigma harvesting, *Priyadharshini et al.* proposed a saffron harvesting system based on an autonomous mobile robot. This system utilises a CNN module to identify mature flowers and the position of their stigmas, and integrates a 6-degree-of-freedom robotic arm onto a mobile platform to achieve precise grasping and separation of the stigmas via vision-guided servo control *(Priyadharshini et al., 2023)*.

Yang et al. addressed the challenges of detecting the minute scale and difficult-to-identify rotational orientation of forsythia pistils by proposing the PSTL-Orient method to detect the area and orientation of rotating pistils. By combining servo technology with flower detection, they achieved precise autonomous pollen delivery to the pistils by a robot *(Yang et al., 2023)*.

However, the identification of stigmas presents multiple technical challenges: due to the minute scale of the stigma target, which occupies very few pixels in an image, traditional object detection methods struggle to effectively extract subtle features; the visual differences between the three states—unpollinated, pollinated and fruiting—are extremely subtle, manifesting only as gradual changes in stigma colour, glossiness and the morphology of the surrounding ovary; these differences require the model to possess a high level of discriminatory ability; lighting conditions in greenhouse environments are complex, with changes in the angle of light in the morning and evening, and fluctuations in brightness caused by shade net coverage, whilst the varied background further increases the difficulty of recognition.

Based on the above analysis, this paper proposes the YOLO11-RCSP method for identifying the pollination status of cherry tomato stigmas, based on an improved YOLO11. This is achieved by introducing a novel attention mechanism, RFA (Receptive Field Attention), adding an enhancement module (CPA-Enhancer) to the original model, adding SAConv switchable dilated convolutions, and improving the loss function by adopting Powerful-IoU (adaptive penalty factors and gradient adjustment functions based on anchor box quality). These four improvement strategies enable the model to adaptively cope with changes in greenhouse lighting and occlusion, whilst also allowing the receptive field to be adjusted according to stigma scale, balancing small-target details with large-scale context to enhance recognition accuracy and robustness. Validation results show that the model achieves a precision of 88.40% and a recall of 84.60%. The average precision at IoU=0.50 is 90.80%, and the average precision across the IoU range of 0.50 to 0.95 is 63.20%. Compared to the original YOLO11 network model, these metrics have improved by 5.20%,3.30%, 4.40% and 5.30% respectively compared to the original YOLO11 network model.

MATERIALS AND METHODS

MATERIALS

DATA AND ACQUISITION

The cherry tomato flower dataset used in this study was captured in an earthen-walled greenhouse at the Innovation and Entrepreneurship Practice Base of the School of Software Engineering, Shanxi Agricultural University, Shanxi Province, China, as shown in Figure 1.

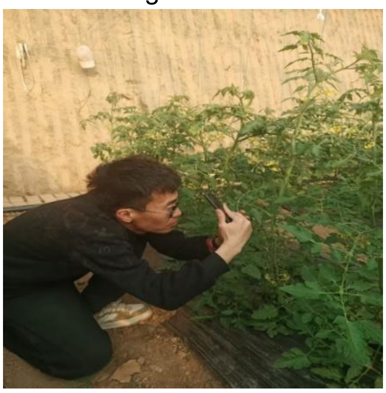

Fig. 1 - Scene of data capture

To enable the model to adapt to changes in stigma scale at different distances, three shooting distances were adopted: close-up (approximately 10–20 cm), medium (approximately 30–40 cm) and far (approximately 50–60 cm). The shooting angles included front, side and upward views to cover the various orientations a pollination robot might encounter during operation. Furthermore, to account for variations in greenhouse lighting conditions, data was collected during three time slots: 8:00–9:30 am, 12:00–1:30 pm and 3:30–5:00 pm. A total of 2,044 images of cherry tomato flowers were captured.

Photographs taken at the three different distances are shown in Figure 2:

Close-up *Medium range* *Long distance*

Fig. 2 - Images captured at different shooting distances

CONSTRUCTION OF THE DATASET

The 2,044 photographs collected were divided into training, validation and test sets in a ratio of 8:1:1, comprising 1,636 images in the training set, 204 in the validation set and 204 in the test set. Adhering to the principle of avoiding duplication and omissions, the LabelImg annotation tool was used to label the stigma of each image, categorising them into three states: unpollinated, pollinated and fruiting *(Li et al., 2024; Song et al., 2024; Sun et al., 2024)*. The specific data labels are shown in Figure 3.

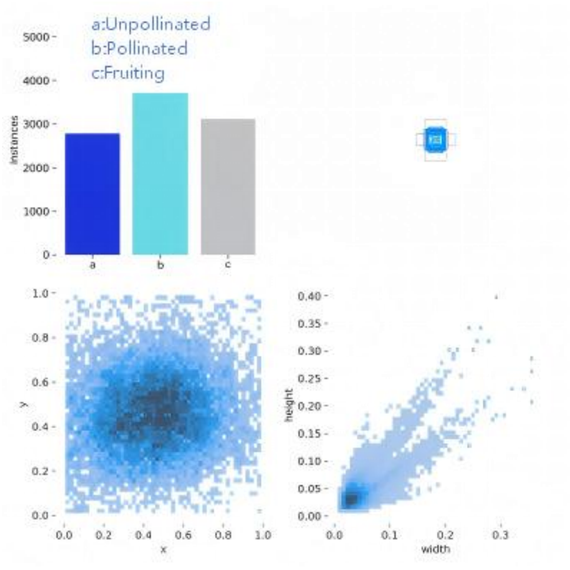

Fig. 3 - Data Labelling

IMPROVED METHOD

To address multiple technical challenges, including the minute scale of the stigma target, subtle visual differences between the three states, and complex lighting conditions in greenhouse environments, this paper proposes YOLO11-RCSP, a stigma pollination status recognition method based on an improved YOLO11. (1) Introducing RFAConv receptive field attention, which resolves the issue of parameter sharing by dynamically adjusting convolution kernel parameters, thereby enhancing the model's ability to perceive both local details and spatial context of the stigma; (2) Adding a CPA-Enhancer adaptive enhancement module to the original model, utilising thought chain prompts to encode degraded information, enabling the model to adaptively cope with changes in greenhouse lighting and occlusion; (3) Replacing the original C3k2 module with the

C3k2_SAConv module, which deeply integrates switchable dilated convolutions with C3k2, enabling the model to adaptively adjust the receptive field according to the scale of the stigma, balancing small-object details with large-scale context; (4) Modifying the loss function to Powerful-IoU, introducing an adaptive penalty factor and a gradient adjustment mechanism based on anchor box quality to improve small-object localisation accuracy and training stability. The diagram of the improved framework is shown in Figure 4.

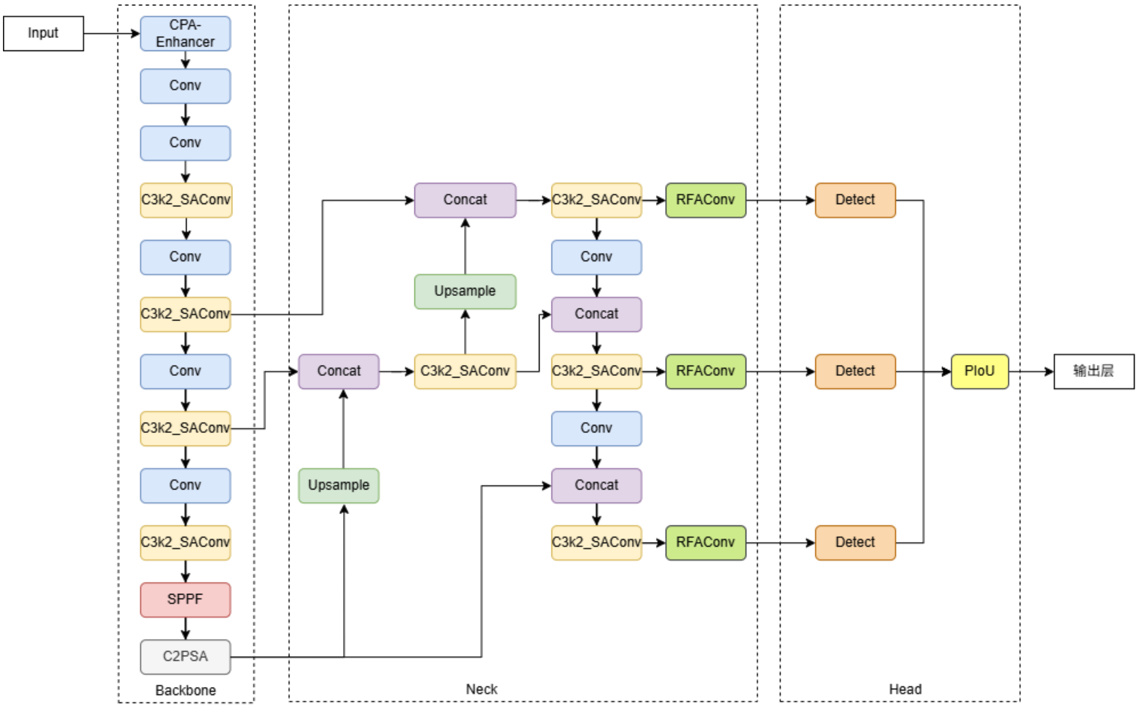

Fig. 4 - Diagram of the improved framework

RFACONV RECEPTIVE FIELD ATTENTION

The core idea of RFAConv is to introduce receptive field attention, allowing convolution kernel parameters to be dynamically adjusted based on the position and contextual information of the receptive field region, thereby addressing the issue of shared convolution kernel parameters *(Zhang et al., 2023)*. Unlike traditional spatial attention mechanisms, RFA not only focuses on spatial features within the image but also places greater emphasis on the spatial feature distribution within each receptive field. By dynamically weighting the importance of various features within the receptive field, it achieves the effect of 'different convolutional kernel parameters for different regions', as illustrated in Figure 5.

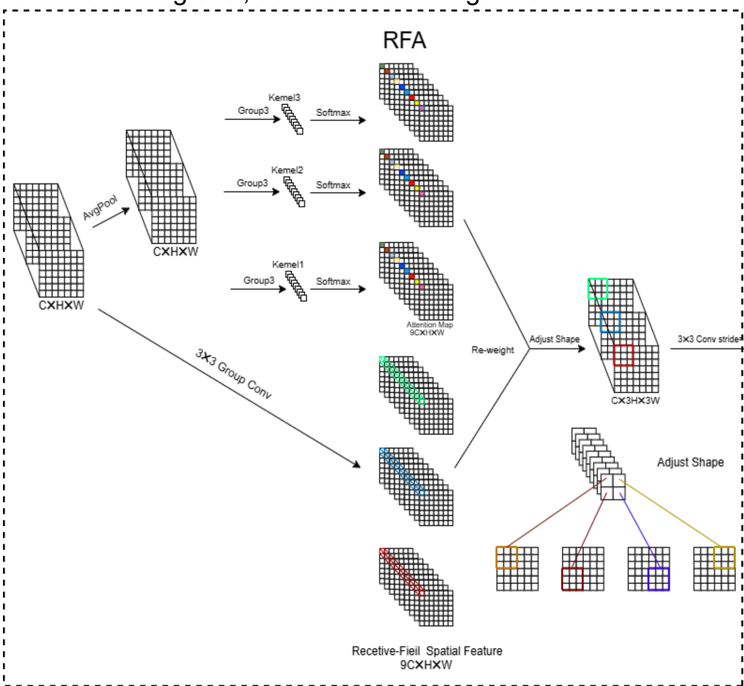

Fig. 5 - RFAConv architecture diagram

This study integrates the RFAConv module into the multi-scale feature fusion stage of YOLO11's neck network. This not only further enhances feature discrimination capabilities and strengthens the model's multi-scale perception but also controls computational overhead whilst improving performance. This design enables the model to dynamically adjust the size of the receptive field based on the scale characteristics of the stigmas in the input image. For smaller, unpollinated stigmas, the model assigns higher weights to branches with smaller receptive fields to preserve fine-scale details; for pollinated stigmas, it assigns higher weights to branches with medium and large receptive fields to capture broader spatial context.

CPA-ENHANCER MODULE

The CPA-Enhancer is specifically designed to enhance object detection performance under unknown degradation conditions. Its core innovation lies in utilising Chain of Thought (CoT) prompts to encode degradation-related information and gradually adjust the enhancement strategy under the step-by-step guidance of the CoT prompts, enabling the model to adaptively address unknown degradation types without prior knowledge of the degradation type *(Zhang et al., 2024)*.This is the first work to apply CoT prompts to object detection tasks; its modular design allows it to be seamlessly integrated into any general-purpose detector, achieving significant performance gains under various unknown degradation conditions. Its structure is shown in Figure 6.

The characteristics of the CPA-Enhancer module are well-suited to the requirements of this study. The identification of pollination status in tomato stigmas faces various complex factors, such as changes in greenhouse lighting and leaf occlusion, which traditional enhancement methods struggle to comprehensively address. Upon adding the CPA-Enhancer to the YOLO11 backbone network, the model can adaptively adjust its enhancement strategy based on the degraded features of the input image, thereby improving the robustness of detection for the small stigmas. Furthermore, as this module has a small number of parameters and low computational overhead, it effectively enhances the accuracy of pollination status recognition whilst maintaining YOLO11's real-time inference capability, providing a more reliable visual perception module for pollination robots.

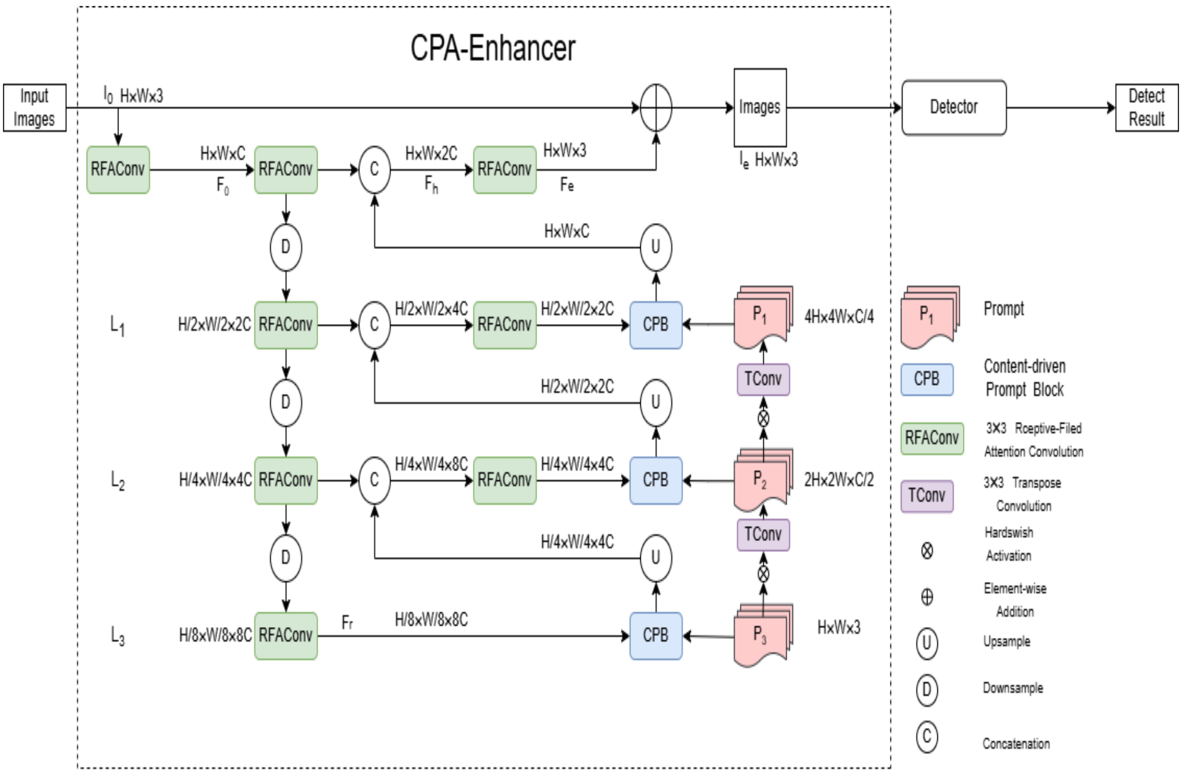

Fig. 6 - CPA-Enhancer architecture diagram

SACONV MODULE

SAConv is a convolutional operation capable of dynamically switching the stranding rate. Its core concept involves introducing a learnable routing mechanism that allows the convolutional kernel to select from or weight-blend multiple stranding rates *(Qiao et al., 2020)*.

This routing module first performs global average pooling on the input features, compressing them into a feature vector, then generates weight coefficients for each branch via two fully connected layers, and finally performs a weighted summation of the outputs from each branch. The expression is given by Eq. (1):

$$F_{out}=\sum_{i=1}^{3} \mathrm{comv}_{d_i} (F_{in}) \tag{1}$$

where '3' denotes the number of parallel dilated convolution branches configured in the SAConv module, i.e. three dilated convolution branches with different dilation rates; d_i is the dilation rate of the i-th branch; w_i is the weight coefficient generated by the dynamic routing module, satisfying Eq. (2):

$$\sum w_i=1 \tag{2}$$

This improvement enables the model to dynamically adjust the void ratio based on input features, allowing it to flexibly capture multi-scale context whilst preserving fine-grained details; its structure is shown in Figure 7.

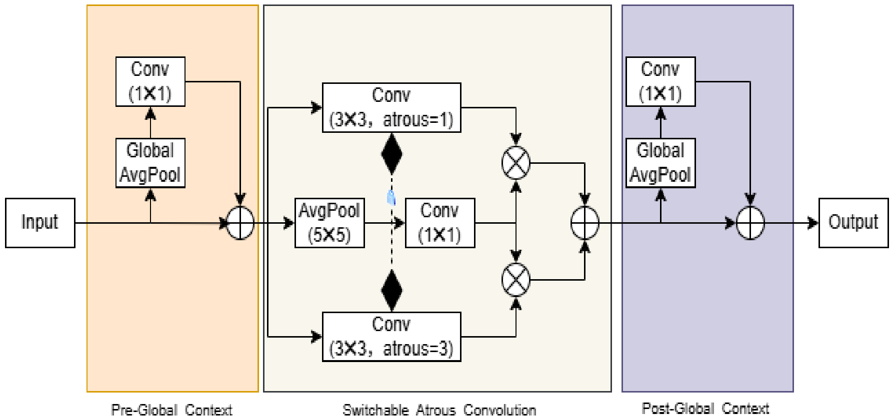

Fig. 7 - SAConv architecture diagram

Wang et al. integrated SAConv into YOLOv8n to construct the LSD-YOLO model, which was applied to the detection of surface diseases on lemons, thereby validating the effectiveness of this module in agricultural small-object detection tasks *(Wang et al., 2024)*. In this study, we propose the C3k2_SAConv module by deeply integrating the SAConv module with the C3k2 module. This enables the C3k2_SAConv module to retain the original lightweight advantages of C3k2 whilst enhancing the model's dynamic perception of features at different scales through the introduction of the SAConv module. It is particularly suitable for fine-grained recognition tasks, such as stigma detection, where scale varies significantly and the requirements for receptive fields change dynamically.

POWERFUL-IOU LOSS FUNCTION

Existing IoU-based loss functions first increase the size of the anchor boxes during regression to achieve overlap with the target box. Even when the area of the anchor box is already larger than that of the target box, this regression method is complex and slow, requiring more iterations to converge. Furthermore, their penalty terms contain unreasonable components that fail to accurately reflect the differences between the anchor and target boxes, and they do not fully account for the target size, which may lead to degradation in certain cases. To address issues such as anchor box enlargement, this paper proposes a size-adaptive penalty term that guides the anchor boxes to regress directly and efficiently. This paper combines this penalty term with a function that adjusts the gradient based on anchor box quality, resulting in a new IoU-based loss term called the Powerful-IoU (PIoU) loss.

PIoU is an improved bounding box regression loss function proposed by Liu et al., which has achieved superior performance compared to traditional loss functions such as CIoU and GIoU on mainstream detectors including YOLOv8 and DINO *(Liu et al., 2023)*. Its core innovation lies in the introduction of an adaptive penalty factor and a gradient adjustment mechanism based on anchor quality, enabling the model to dynamically adjust the loss contribution according to anchor quality, thereby enhancing its adaptability to small objects and high-precision localisation tasks. Its expression is given by Eq. (3):

$$L_{PIoU} = \gamma(1 - IoU)^2 \tag{3}$$

Here, '$(1\text{-IOU})^2$' serves as the base loss term, assigning a higher loss weight to samples with low IoU; ` $\gamma$ ` acts as a gradient adjustment factor, dynamically adjusting the gradient contribution based on the quality of the anchor boxes.

This design enables PIoU to combine the dual advantages of rapid convergence and stable refinement, making it particularly suitable for the high-precision localisation requirements of small targets such as the stigma.

Evaluation Metrics

For object detection models, this paper primarily employs precision (P), recall (R), average precision (mAP@50), and mean average precision (mAP@50–95) as the main metrics for evaluating model performance. P represents the proportion of all targets correctly identified by the algorithm's bounding boxes, reflecting the model's accuracy; R represents the proportion of correctly identified samples out of the total number of samples; mAP@50 is a key metric reflecting the comprehensive performance of the detection model; mAP@50-95 denotes the average of the mean average precision across different IoU thresholds. The calculations for precision, recall, average precision and mean average precision are shown in Eqs. (4) to (7):

$$\mathrm{P}=\frac{TP}{TP+FP} \tag{4}$$

$$\mathrm{R}=\frac{TP}{TP+FN} \tag{5}$$

$$\mathrm{AP}=\int_0^1 P(r)dr \tag{6}$$

$$mAP=\frac{1}{C}\sum_{i=1}^{c} AP_i \tag{7}$$

TP denotes the number of correctly detected positive samples; FP denotes the number of incorrectly detected negative samples; FN denotes the number of missed positive samples; C represents the number of classes; AP_i represents the average precision of the i-th class.

EXPERIMENTAL ENVIRONMENT

For detailed experimental configuration parameters, see Table 1.

Table 1

Experimental Environment Configuration

Configuration Name	Version and Model
GPU Model and Quantity	NVIDIA GeForce RTX 4080
CPU Model	13th Gen Intel® Core™ i7-13700K
System Memory	32.0GB
Operating System	Windows 11 Home Chinese Edition (22H2 build)
Deep learning framework	Torch 2.2.0 + cu118
CUDA and cuDNN versions	CUDA 11.8 + cuDNN 8.9.7
Programming Languages and Dependencies	timm == 0.9.12, mmcv-full == 2.2.0, numpy=1.26.3

The training parameters are summarized in Table 2.

Table 2

Training Parameter Configuration

Parameter Name	Parameter Value
epoch	200
batch	8
workers	4
optimizer	SGD
size	640
patience	30
amp	False
iou	0.7
lr0	0.01
lrf	0.01

RESULTS AND DISCUSSION

To ensure the validity and reliability of both ablation and comparative experiments, this study strictly controlled extraneous variables by employing identical training configurations for all participating models, and fixed random seeds across all runs to eliminate stochastic fluctuations, thus ensuring that any observed performance variations arise exclusively from network architectural improvements.

ABLATION STUDIES

In the ablation experiments, individual modules (A, B, C, D) were incorporated into the base model to assess their impact on performance. A denotes the RFAConv module; B denotes the CPA-Enhancer module; C denotes the C3k2_SAConv module; D denotes the PIoU loss function. The results are shown in Table 2.

Table 2

Results of the ablation experiments

Model	Module A	Module B	Module C	Module D	P	R	mAP@50	mAP@50–95	GFLOPs	FPS
1					0.832	0.813	0.864	0.579	6.3	210.1
2	√				0.826	0.817	0.87	0.574	14.8	37.7
3		√			0.83	0.819	0.867	0.6	9.6	10.3
4	√	√			0.843	0.829	0.882	0.614	13.87	8.4
5			√		0.84	0.833	0.881	0.568	5.9	93.6
6				√	0.832	0.813	0.865	0.581	6.3	207.1
7			√	√	0.867	0.807	0.878	0.569	5.9	97.6
8	√		√		0.855	0.821	0.868	0.576	14.6	33.8
9	√		√	√	0.838	0.827	0.878	0.584	14.6	33.9
10	√	√	√	√	0.884	0.846	0.908	0.632	13.75	36.3

Based on the results of the comprehensive ablation experiments, the final improved scheme is Model 10: A+B+C+D. It achieved the best performance across core accuracy metrics such as precision, recall and mAP, significantly outperforming the baseline model and other module combinations, thereby fully validating the effectiveness of the multi-module joint optimisation strategy. However, compared to the baseline model, the computational cost of this scheme increased from 6.3 to 13.75, whilst inference speed decreased from 210.1 to 36.3. This change is primarily attributable to the inherent complexity of the introduced modules: RFAConv and C3k2_SAConv enhance feature representation capabilities but increase computational overhead; the serial processing mechanism of the CPA-Enhancer reduces real-time frame rates; and the PIoU loss function introduces additional floating-point operations. Nevertheless, these trade-offs do not detract from the value of the findings of this study. This study prioritises the improvement of detection accuracy; exchanging controllable computational redundancy for a 4.4% improvement in mAP@50 and a 5.3% improvement in mAP@50-95 represents a typical reasonable trade-off between accuracy and efficiency. Furthermore, compared to other high-accuracy combinations, such as Model 4, which achieves only 8.4 FPS, this approach maintains 36.3 FPS whilst achieving the highest accuracy, demonstrating a superior overall balance. Finally, the baseline model itself is designed to be extremely lightweight, with its 210 FPS far exceeding practical application requirements, whilst the computational load of this approach remains within an efficient and acceptable range, providing a clear direction for subsequent engineering optimisation. Therefore, variations in computational load and speed do not diminish the innovative contribution of this study at the algorithmic level; rather, they further highlight the significant advantages in terms of accuracy improvement.

MULTI-MODEL COMPARISON

To validate the superiority of YOLO11 in detecting the pollination status of cherry tomatoes, this paper trained seven models for detecting the pollination status of cherry tomato flowers using the same cherry tomato dataset under identical conditions. Precision (P), Recall (R), Average Precision (mAP@50) and Mean Average Precision (mAP@50-95) were adopted as the primary metrics for evaluating model performance. The comparison results are shown in Table 3.

Table 3

Comparison of model performance

Model	Precision (P)	Recall (R)	mAP@50	mAP@50–95	GFLOPs	FPS
YOLOv5	0.827	0.798	0.865	0.552	5.8	260.1
YOLOv6	0.834	0.796	0.855	0.533	11.5	242.8
YOLOv8	0.829	0.788	0.858	0.547	6.8	256.6
YOLOv9t	0.837	0.815	0.873	0.549	6.4	103.6
YOLOv10n	0.835	0.79	0.858	0.545	8.2	184.1
YOLO11	0.832	0.813	0.864	0.579	6.3	210.1
YOLO11-RCSP	0.884	0.846	0.908	0.632	13.75	36.3

As summarised in Table 3, which compares the results of several mainstream models on the same cherry blossom pollination dataset, the YOLO11-RCSP model proposed in this paper leads comprehensively across core accuracy metrics, achieving a precision of 88.4% and a recall of 84.6%, with mAP@50 and mAP@50-95 reaching as high as 90.8% and 63.2%, respectively, significantly outperforming six comparison models ranging from YOLOv5 to YOLO11. For example, compared to the original YOLO11, mAP@50 is improved by 4.4%, and mAP@50-95 by 5.3%; compared to the similarly performing YOLOv9t, mAP@50 is improved by 3.5%, fully demonstrating the significant advantage of this improved approach in detection accuracy. In terms of computational efficiency, YOLO11-RCSP achieves 13.75 GFLOPs and 36.3 FPS. Although this exceeds some lightweight models, the improvement in accuracy far outweighs the computational cost when compared to YOLOv6 and YOLOv10n. However, whilst YOLOv9t maintains 103.6 FPS, its mAP@50 and mAP@50-95 are both significantly lower than those of our model. The 36.3 FPS fully meets the real-time requirements of agricultural detection scenarios, whilst the 210.1 FPS of the baseline YOLO11 is over-engineered. Therefore, trading a moderate loss in efficiency for a significant gain in accuracy aligns with the core requirements for detection accuracy in practical applications, fully validating the superiority of YOLO11-RCSP in the comprehensive trade-off between accuracy and efficiency.

VISUALISATION

VISUAL ANALYSIS OF DETECTION RESULTS

In this experiment, the performance of the YOLO11-RCSP model and other models in the object detection task was evaluated using six sets of cherry tomato images with distinct characteristics, as shown in Figure 8.

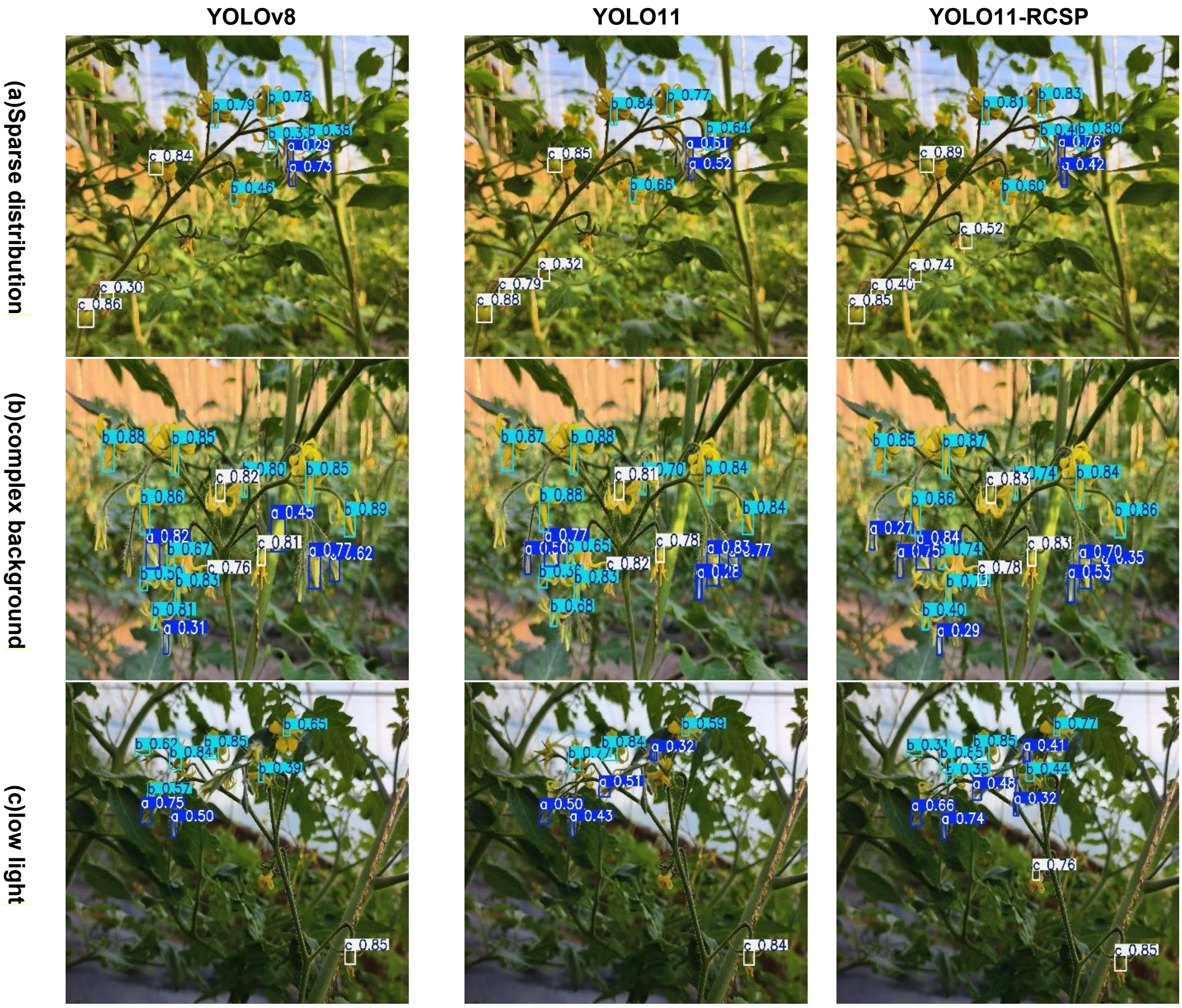

Fig. 8 - Detection comparison among different models and the improved model under various cherry tomato scenarios

The blue, green, and white bounding boxes indicate detected unpollinated styles, detected pollinated styles, and detected cherry tomato fruits, respectively.

In sparsely distributed and complex background scenarios, other models generally suffer from missed detections or misaligned bounding boxes. For example, YOLOv5 and YOLOv8 exhibit a weak response to unpollinated stigmas obscured by leaves, or fail to detect them entirely, whereas YOLO11-RCSP can accurately localise the stigma region, with bounding boxes closely matching the actual targets and a significantly higher accuracy rate than other models.

In scenes with dense targets, such as Column e, YOLOv6 and YOLOv9t exhibited overlapping detection boxes and false positives between multiple targets, misidentifying adjacent petals or ovaries as stigmas; whilst YOLOv10n and the original YOLOv11 detected most stigmas, their localisation accuracy was insufficient, and the detection boxes failed to tightly enclose the stigmas. In contrast, YOLO11-RCSP not only correctly detected all stigmas but also produced compact detection boxes with high accuracy.

In low-light images, the accuracy of other models generally declined, with some models even failing to detect stigmas, whilst YOLO11-RCSP maintained a high detection rate and localisation accuracy.

Visual Analysis of Heatmaps

In this experiment, to intuitively evaluate the areas of focus of different models regarding the pollination status of cherry tomato pistils, this study utilised the GradCAM method to generate heatmaps for each model and conducted a comparative analysis, as shown in Figure 9.

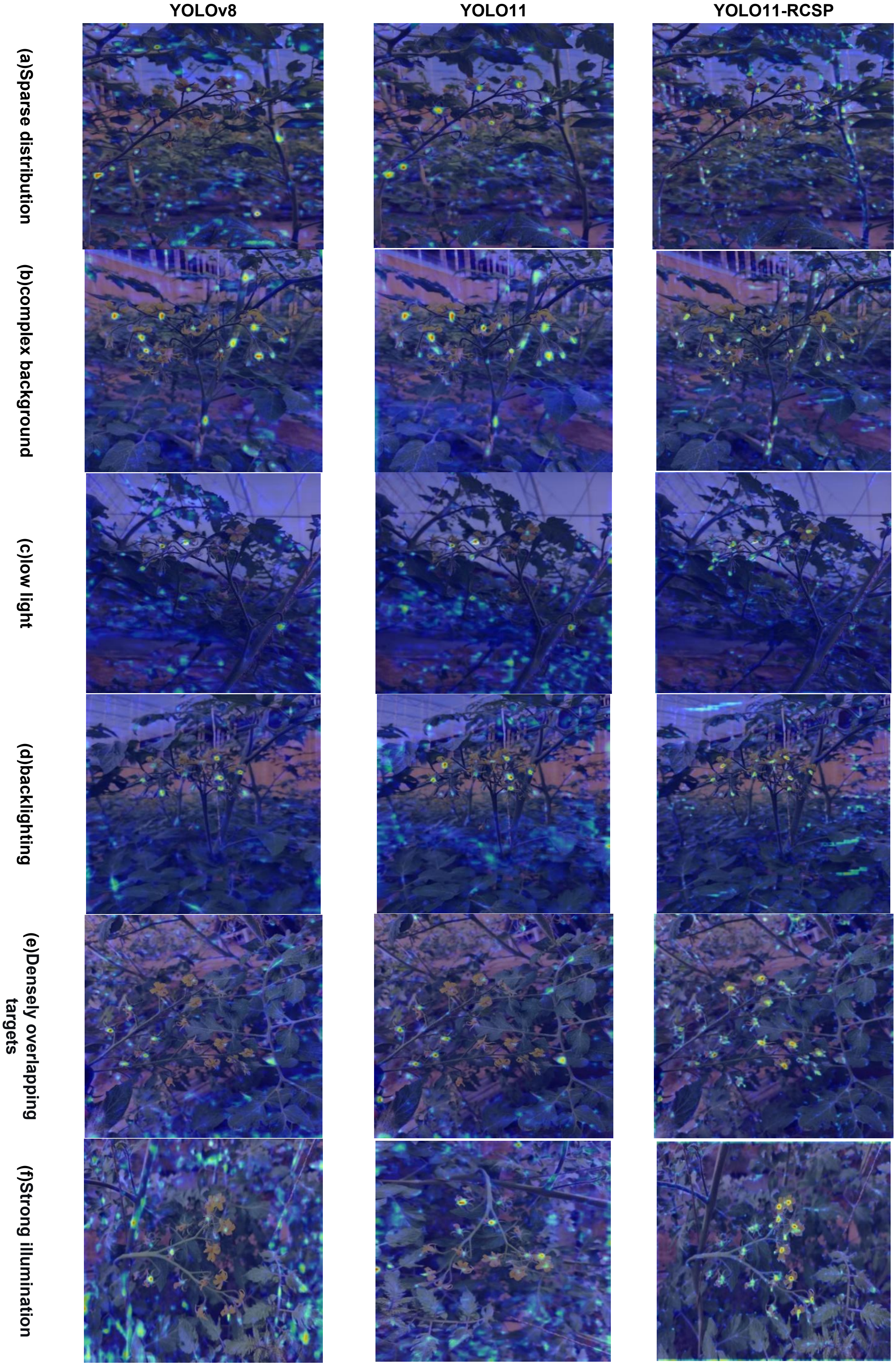

Fig. 9 - Comparison of heatmaps generated by different models and the improved model on various cherry tomato scenes

In scenarios featuring sparse cherry tomato distributions and complex backgrounds, the heatmaps demonstrate that the YOLO11-RCSP model focuses more accurately on the target regions compared to other models, effectively reducing unnecessary background activations. This indicates that the improved model can more effectively extract and focus on key features when handling complex backgrounds or sparse target distributions, exhibiting stronger robustness against interference.

For dense cherry tomato distributions, the YOLO11-RCSP model maintains high recognition accuracy even when targets are densely packed, accurately distinguishing individual targets and reducing false activations or missed detections. In contrast, other models are more susceptible to interference between targets in such scenarios, leading to a decline in detection performance.

Scene Application

Real-scene image collection was carried out in a ridge-planted facility tomato greenhouse using a DJI mobile phone gimbal combined with a smartphone, to verify the actual deployment effect of the model. During collection, the camera advanced uniformly along the planting ridges with wide-angle framing to include multiple tomato flowers in a single frame. By traversing each ridge, moving evenly and slightly adjusting the gimbal angle, repeated frames and overlapping regions were avoided, and the flowering and fruiting areas of the whole greenhouse were fully covered. A total of 218 field test images were obtained.

Inference deployment and acceleration of the optimal weight model best.pt were implemented based on PyTorch. A tomato flower detection system was constructed using Python and OpenCV, which enabled automatic image inference and detection. Unpollinated flowers, pollinated flowers and developed fruits were marked with bounding boxes of different colours, and quantity statistics and visual output were completed.

A field comparison test was conducted with reference to manually annotated ground truth. A total of 1530 valid targets were detected, including 445 unpollinated flowers, 820 pollinated flowers and 265 developed fruits, corresponding to the ground truth values of 454, 811 and 271 respectively. The results show that the model can accurately identify tomato organs at different developmental stages in the field environment with low quantity statistics error, and can provide intelligent support for flowering management, pollination effect evaluation and yield prediction of facility tomatoes.

CONCLUSIONS

To address the core requirement for the accurate recognition of stigma pollination status in cherry blossom pollination robots, along with technical challenges including the tiny size of stigma targets, subtle visual differences among three states, and complex lighting conditions in greenhouse environments, this paper proposes a YOLO11-RCSP model based on the improved YOLO11. The model incorporates four key enhancements: introducing RFAConv receptive field attention convolution to solve parameter sharing issues; adding a CPA-Enhancer adaptive enhancement module to improve adaptability to greenhouse lighting variations and occlusions; replacing the original C3k2 module with the C3k2_SAConv module to dynamically adjust the receptive field; and adopting the Powerful-IoU loss function to optimize small-object localization accuracy. On the self-built cherry tomato pistil dataset, the model achieves 88.40% precision, 84.60% recall, 90.80% mAP@50, and 63.20% mAP@50–95, representing improvements of 5.20%, 3.30%, 4.40%, and 5.30% respectively over the baseline YOLO11, and significantly outperforms mainstream detectors such as YOLOv5, YOLOv6, YOLOv8, YOLOv9t, and YOLOv10n, thereby validating the effectiveness and superiority of the proposed method.

Although the YOLO11-RCSP model has achieved significant improvements in detection accuracy, computational complexity has increased and inference speed has decreased following the model enhancements; whilst it meets real-time requirements, there is still room for optimisation. Future work will focus on model lightweighting and adaptive methods for cross-scenario data augmentation to further enhance the model’s deployment efficiency and generalisation robustness.

This research is expected to be deployed in the visual perception module of a tomato pollination robot, enabling precise determination of pollination status and providing a basis for decision-making in automated pollination operations.

ACKNOWLEDGEMENT

This study received support from two projects: the Science and Technology Innovation Enhancement Project under the Scientific Research Initiatives of Shanxi Agricultural University (No. CXGC2025057), the Smart Agricultural Machinery and Rural Governance Project funded by the Shanxi Provincial Key Research and Development Project (No. 202202140601021). The authors express sincere gratitude for the financial and institutional support provided by these initiatives.

REFERENCES

[1] Gao, P., Zhang, X.-Y., Yang, X.-L., et al. (2024). Robust Visual Tracking Using Hierarchical Vision Transformer with Shifted Windows Multi-Head Self-Attention: Regular Section. *IEICE Transactions on Information and Systems*, E107.D(1), 161--164.

[2] Hou, B. F., Li, X. M., Mou, X. W., et al. (2024). Design and experiment of cherry tomato greenhouse inspection robot system. *Chinese Journal of Agricultural Mechanization*, *45*(1), 285--294. https://doi.org/10.13733/j.jcam.issn.2095-5553.2024.01.039

[3] Hu, Y. F., Zhao, X. L., Li, P. J., et al. (2023). Tomato fruit detection in natural environment based on improved YOLOv5. *Chinese Journal of Agricultural Mechanization*, *44*(10), 231--237. https://doi.org/10.13733/j.jcam.issn.2095-5553.2023.10.032

[4] Jia, G. G., Wang, B., & Zhao, Y. (2024). Simulation of apple picking machinery based on visual recognition. *Journal of Agricultural Mechanization Research*, *46*(12), 22--27+112. https://doi.org/10.13427/j.cnki.njyi.2024.12.004

[5] Kang, Y., Li, J. Y., Qiu, J. K., et al. (2023). Soybean seedling identification from ultra-low altitude UAV based on EfficientDet network. *Journal of Agricultural Mechanization Research*, *45*(9), 7--13. https://doi.org/10.13427/j.cnki.njyi.2023.09.013

[6] Li, Z. C., Luo, W. P., Ling, X. J., et al. (2024). Detection method of chestnut burr in natural environment based on improved YOLOv8. *Chinese Journal of Agricultural Mechanization*, *45*(12), 251--258. https://doi.org/10.13733/j.jcam.issn.2095-5553.2024.12.037

[7] Liu, C., Wang, K., Li, Q., et al. (2023). Powerful-IoU: More straightforward and faster bounding box regression loss with a nonmonotonic focusing mechanism. *Neural Networks*, *170*, 276--284.

[8] Lorberboym, E., Lazare, S., Golshmid, P., & Shani, G. (2026). Estimating Cannabis Flower Maturity in Greenhouse Conditions Using Computer Vision. *Agriculture*, *16*, 460. https://doi.org/10.3390/agriculture16040460

[9] Manzoor, S. H., Zhang, Z., Li, X., et al. (2024). Lightweight and Robust YOLOv5s with MobileNetV3 and GhostNet for Precision Apple Flower Detection for Pollination Drones. In Y. Liu, L. Yang, Y. Shi, et al. (Eds.), New Technologies Applied in Apple Production: Sensing and Autonomous Systems (pp. 81--108). *Singapore: Springer Nature Singapore*.

[10] Pan, F., Hu, M., Duan, X., Zhang, B., Xiang, P., Jia, L., Zhao, X., & He, D. (2024). Enhancing kiwifruit flower pollination detection through frequency domain feature fusion: a novel approach to agricultural monitoring. *Frontiers in Plant Science*, *15*, 1415884. https://doi.org/10.3389/fpls.2024.1415884

[11] Pinheiro, I., Moreira, G., Magalhães, S., et al. (2024). Deep learning based approach for actinidia flower detection and gender assessment. *Scientific Reports*, *14*(1), 24452. https://doi.org/10.1038/s41598-024-75432-0

[12] Priyadharshini, J., Arunn, T., Rithick Roshan, B., et al. (2023). Saffron Harvesting using Robotic Arm Embedded on a Rover. *In 2023 2nd International Conference on Automation, Computing and Renewable Systems (ICACRS) (pp. 1745--1750)*. Pudukkottai: IEEE.

[13] Qiao, S., Chen, L.-C., & Yuille, A. (2020). DetectoRS: Detecting Objects with Recursive Feature Pyramid and Switchable Atrous Convolution. *CoRR*, *abs/2006.02334*.

[14] Song, Y. L., Wang, C., Li, D. Y., et al. (2024). UAV small target detection algorithm based on improved YOLOv5s. *Journal of Zhejiang University (Engineering Edition),* *58*(12), 2417--2426.

[15] Sun, J. F., Wang, L., & Wang, J. P. (2024). Research on fruit freshness detection algorithm based on improved YOLOv5s. *Modern Electronic Technology*, *47*(22), 37--43. https://doi.org/10.16652/j.issn.1004-373x.2024.22.007

[16] Tang, Y., Sun, J. Y., Zhang, Y. F., et al. (2025). Kiwifruit flower detection using an optimized YOLOv11n architecture. *INMATEH Agricultural Engineering*, *77*(3), 44--53.

[17] Touko Mbouembe, P. L., Liu, G., Park, S., & Kim, J. H. (2024). Accurate and fast detection of tomatoes based on improved YOLOv5s in natural environments. *Frontiers in Plant Science*, *14*, 1292766. https://doi.org/10.3389/fpls.2023.1292766
[18] Wang, J. S., Liu, Q. Y., & Zhao, P. S. (2016). Study on effect of vibrator pollination on tomato in winter-spring greenhouse. *Vegetables*, (4), 21--23.
[19] Wang, S., Li, Q., Yang, T., et al. (2024). LSD-YOLO: Enhanced YOLOv8n Algorithm for Efficient Detection of Lemon Surface Diseases. *Plants*, *13*(15), 2069. https://doi.org/10.3390/plants13152069
[20] Wang, X. F., Zhang, Q., & Su, A. H. (2013). Application effect of bumblebee pollination technology on greenhouse cherry tomato. *Northern Horticulture*, (11), 45--47.
[21] Yang, M., Lyu, H., Zhao, Y., et al. (2023). Delivery of pollen to forsythia flower pistils autonomously and precisely using a robot arm. *Computers and Electronics in Agriculture*, *214*.
[22] Zhang, X., Liu, C., Yang, D., Song, T., Ye, Y., Li, K., & Song, Y. (2023). RFAConv: Innovating Spatial Attention and Standard Convolutional Operation. *CoRR*, *abs/2304.03198*.
[23] Zhang, Y., Wu, Y., Liu, Y., & Peng, X. (2024). CPA-Enhancer: Chain-of-Thought Prompted Adaptive Enhancer for Object Detection under Unknown Degradations. *arXiv preprint arXiv*:2403.11220.