YOLO-TRS: AN IMPROVED YOLO11 FOR TOMATO FRUIT RIPENESS AND STEM DETECTION
/
YOLO-TRS：一种用于番茄果实成熟度与果梗检测的改进 YOLO11 算法

Fuming MA[1)], Shaonian LI*[1)], Jing TAN[1)], Yue LI[2)]
[1)] College of Energy and Power Engineering, Lanzhou University of Technology, Lanzhou, Gansu/ China
[2)] College of Electrical and Information Engineering, Lanzhou University of Technology, Lanzhou, Gansu/ China
Tel: 0931-2973750; E-mail: Lsn19@163.com
Corresponding author: Shaonian Li
DOI: https://doi.org/10.35633/inmateh-77-91

Keywords: *tomato ripeness; YOLO11; dynamic snake convolution; CAA attention mechanism*

ABSTRACT
During field tomato harvesting, challenges such as stem-leaf occlusion, fruit overlap, and difficulties in stem localization significantly hinder the performance of harvesting robots. To address these issues, a joint detection model for fruits and fruit stems, termed YOLO-TRS, is proposed based on the YOLO11n network. First, a novel C3k2-DS module is designed and integrated into the backbone network, enhancing the model's ability to represent complex structural features of fruit stems. In addition, a CAA module is incorporated into the backbone to improve long-range feature modeling, thereby effectively reducing missed detections of fruits and fruit stems under occlusion conditions. The proposed model is evaluated using a self-constructed dataset. Experimental results show that YOLO-TRS achieves precision, recall, and mAP values of 89.9%, 91.5%, and 94.8%, respectively, outperforming the baseline YOLO11n model by 2.3%, 1.0%, and 2.4%. Compared with other classical object detection algorithms, YOLO-TRS demonstrates clear advantages in both detection accuracy and computational efficiency. These results confirm that the proposed model can effectively support fruit ripeness-related detection and accurately localize stem positions in complex field environments, providing a theoretical basis for intelligent agricultural harvesting.

摘要
田间番茄采收过程中，茎叶遮挡、果实重叠及果柄定位困难等问题严重影响采收机器人的作业性能。为解决上述挑战，本文基于 YOLO11n 网络提出一种果实与果柄联合检测模型 YOLO-TRS。首先，提出并将 C3k2-DS 模块集成于骨干网络中，增强模型对果柄复杂结构特征的建模能力；此外，在骨干网络中集成 CAA 模块，提升模型的长距离特征建模能力，进而有效降低遮挡场景下果实与果柄的漏检率；最后，基于自建田间番茄数据集对所提模型进行验证。实验结果表明，改进后的模型精度（Precision）、召回率（Recall）和平均精度均值（mAP）分别达到 89.9%、91.5%和 94.8%，较 YOLO11n 模型分别提升 2.3%、1.0%和 2.4%；与其他经典目标检测算法相比，该模型在检测精度与计算效率方面均展现出显著优势。这些实验结果验证了，YOLO-TRS 模型能够在复杂田间环境下有效检测果实成熟度并精准定位果柄位置，为农业智能采摘提供理论支撑。

INTRODUCTION

Tomato is one of the most economically significant crops in the world. China is a major producer of both fresh and processed tomatoes. *(Li et al., 2021)*. In contemporary agricultural production, precise detection of fruit ripeness and stem position are key to realizing intelligent harvesting. The accurate differentiation of ripeness stages is critical for maintaining fruit quality, minimizing storage costs, and optimizing harvesting efficiency. Additionally, the accurate detection of the stem position provides precise guidance for the robotic arms of harvesting machines, effectively avoiding fruit damage during the harvest. *(Zhou et al., 2022)*.

In the field of tomato phenotyping detection, early research primarily relied on traditional computer vision techniques *(Hou et al., 2015)*. *Feng et al., 2015,* developed a vision system based on line-structured light, using color feature extraction in a specific chromatic aberration model to identify red ripe tomatoes. *Li et al., 2021* utilized RGB-D images and improved clustering algorithms to enhance the accuracy and robustness of overlapping fruit recognition. *Goel et al., 2015,* developed a vision-based system that achieved fine classification of tomatoes into six maturity stages using a fuzzy rule-based classification method, with an

[1] *Fumin Ma, M.S. Stud. Eng.; Shaonian Li, Prof. Ph.D. Eng.; Jing Tan, M.S. Stud. Eng.; Yue Li, B.S. Stud. Eng.*

accuracy of up to 94.29%. However, these methods generally depended on handcrafted features, making them susceptible to background interference in complex natural environments and limiting their generalization capability. With advancements in technology, deep-learning-based convolutional neural networks (CNNs) have demonstrated powerful automatic feature-learning capabilities and have gradually become the mainstream approach *(Kamalesh et al., 2024)*. Among these, the YOLO series models have been widely applied in tomato detection tasks due to their excellent balance between speed and accuracy. Within the YOLOv3 framework, *Liu et al., 2020,* proposed a multi-scale feature fusion algorithm termed IMS-YOLO, increasing the detection accuracy to 97.13%. Liu et al. designed the YOLO-Tomato model to cope with complex environmental conditions *(Liu et al., 2020)*. Subsequently, YOLOv4 was adopted for its stronger feature extraction capability. *Li et al., 2021,* combined the HSV color space to improve the correct recognition rate to 94.77%. *Yang et al., 2022,* incorporated the CBAM module into backbone network of the YOLOv4-tiny, enabling accurate tomato ripeness classification with an average precision of 97.9%. *Liu et al., 2023,* proposed a tomato ripeness detection method that combines YOLOv4 with ICNet, achieving an average detection accuracy of 99.31%. YOLOv5 further optimized both accuracy and speed. *Gao et al., 2024,* introduced the CBAM attention module and Soft-NMS, effectively enhancing recognition robustness in complex environments. *He et al., 2022,* addressed the challenge of nighttime tomato recognition by improving the loss function. Recently, research has increasingly focused on model lightweighting and accuracy improvement. For instance, *Ge et al., 2022* incorporated ShuffleNetV2 and BiFPN to compress the model while maintaining performance. Zhang et al. and Wang et al. utilized attention mechanisms and optimized loss functions, respectively, both achieving high-precision tomato detection *(Zhang et al., 2023; Wang et al., 2023)*. The research frontier has now expanded to newer frameworks such as YOLOv8 and YOLO11. *Tian et al., 2024,* added detection layers and designed novel modules, constructing a TCAttn-YOLOv8 model that achieved 96.31% mAP. Wu et al. and Sun et al. innovated on the Neck layer, obtaining excellent comprehensive performance *(Wu et al., 2024; Sun et al., 2024)*. The latest YOLO11 model is also being explored. *Wei et al., 2024,* successfully constructed a lightweight and efficient detection model by introducing Ghost modules and feature refinement modules.

Despite these significant advances, current visual recognition systems for intelligent tomato harvesting face two prominent issues: First, most existing studies focus solely on fruit ripeness detection, lacking joint detection of both fruit ripeness and stem positions—the latter being critical for achieving efficient automated harvesting. Second, in complex natural environments, background interference and severe occlusion between leaves and fruits pose serious challenges to the detection accuracy of small targets like stems and fruits of specific maturity stages. To address these challenges, this study proposes an innovative lightweight model, YOLO-TRS, aimed at achieving accurate and robust joint detection of fruit ripeness and stems. The main contributions of this paper are as follows:

(1) A novel feature extraction module, C3k2-DS, is proposed, in which the standard convolution in the C3k2 module is replaced with dynamic snake convolution *(Qi et al., 2024)*. Using this module as a core component, the backbone network of YOLO11n is redesigned, significantly enhancing the model's ability to extract complex structural features such as fruit stems.

(2) Building upon the introduction of the C3k2-DS module, the CAA module *(Cai et al., 2024)* is integrated into the backbone network, resulting in the development of the novel YOLO-TRS algorithm. The incorporation of the CAA module further enhances the model's ability to capture contextual information and salient features, thereby improving its robustness to complex background interference and its capability to recognize occluded targets.

MATERIALS AND METHODS

Dataset and preprocessing

Data acquisition

The image dataset employed in the study was collected from the cultivation of a specific university. The tomato variety used is Provence, and the images were captured using an iPhone smartphone with a resolution of 4624×4624 pixels. To enhance the diversity of the tomato images, images were captured at various time intervals between 7:00 AM and 10:00 PM, employing different shooting angles (horizontal and top-down) and distances (with a direct camera-to-tomato distance varying from 300 to 550 mm). Additionally, variables such as lighting conditions, shading levels, and fruit quantities were systematically considered. A total of 1,000 tomato images were acquired. Representative tomato samples from different conditions are illustrated in Figure 1.

(a) backlight (b) insufficient lighting (c) direct sunlight (d) occlusion (e) dense distribution (f) sparse distribution

Fig. 1 - Examples of tomato images under different collection conditions

Data annotation and partitioning

The current national standard GH/T1193-2021 categorizes tomato ripeness into six stages based on color and size: unripe, green-ripe, breaker, pre-red ripe, mid-red ripe, and post-red ripe. The characteristics of tomatoes at different stages of ripeness are presented in Table 1. To simplify the classification process and improve training efficiency, tomato ripeness stages were redefined based on current national standards. Specifically, the unripe and green-ripe stages were merged into a green stage, the breaker and pre-red ripe stages were combined into a half-ripe stage, and the mid-red ripe and post-red ripe stages were merged into a red-ripe stage. This resulted in a three-stage ripeness classification scheme. All experimental data annotations strictly followed this redefined classification standard.

Table 1

Morphological characteristics of tomato fruits with different ripeness

Ripeness level	Morphological characteristics
unripe	Fruit and seeds have not yet fully developed and shaped, green pericarp, no luster, ripening difficulties
green-ripe	Fruit stereotypes, fruit surface has a glossy, from green to white green, the seed has grown, around the gelatinous, at this time can be artificially ripened, picking and storage
breaker	From green ripe to red ripe transition period, the umbilicus around the beginning of yellow or light red halo spot, fruit with red surface less than 10%
pre-red ripe	One to thirty percent red ripe, fruit with red surface 10%-30%
mid-red ripe	Forty to sixty percent red ripe, the fruit with red surface 40%-60%
post-red ripe	Seventy to ten percent red ripe, fruit with red surface 70%-100%

The dataset was randomly divided into training, validation, and tests set in a ratio of 8:1:1 for model training. Detailed information of the tomato dataset is provided in Table 2.

Table 2

Experimental dataset and data distribution

Set	Target Box				Number of Images
	green	half	red	stem	
Train	335	631	673	1282	800
Validation	45	74	85	168	100
Test	41	80	84	152	100
Total	421	785	842	1602	1000

Model Introduction

Network Architecture of YOLO11

YOLO11 marks a notable progression in object detection architectures. It introduces the C3k2 module, which integrates variable kernel convolution and channel separation to capture richer contextual information, thereby enhancing multi-scale feature extraction. The model also incorporates the C2PSA module to strengthen spatial correlations in feature maps, improving attention to key regions such as small or occluded objects. With a reduced parameter size and strong compatibility with edge devices, YOLO11 is well-suited for deployment on harvesting robots. Among its variants, the lightweight YOLO11n is selected as the baseline in this study. As illustrated in Figure 2, its architecture mainly comprises three parts: the Backbone, Neck, and Head.

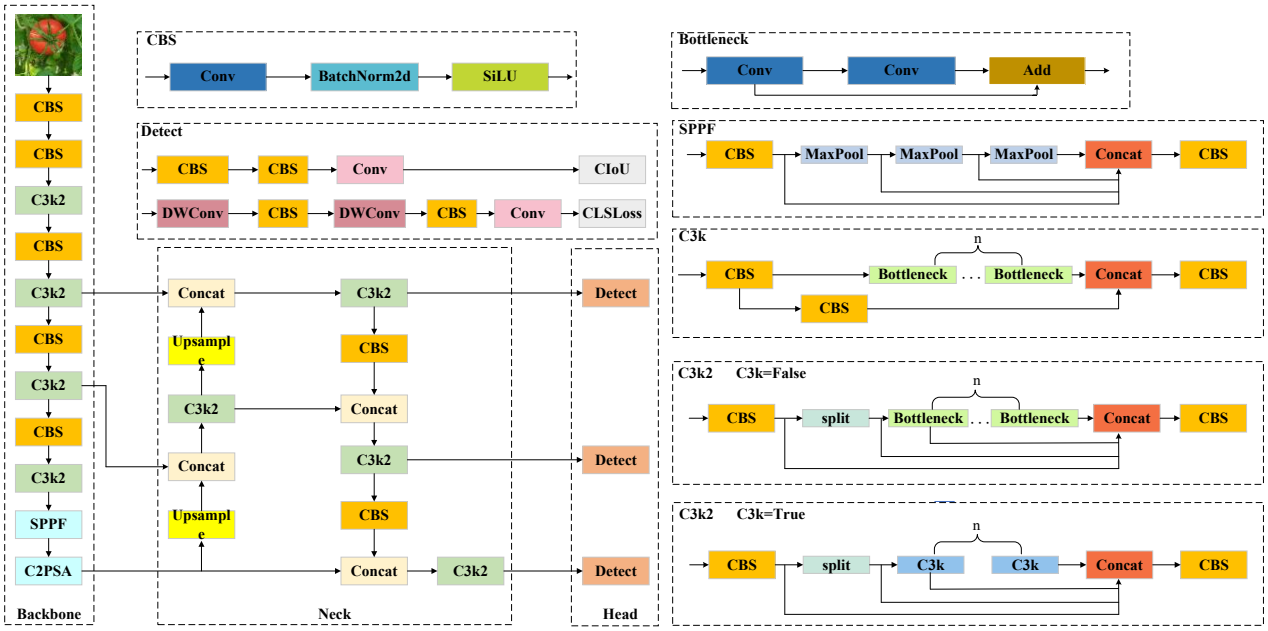


Fig. 2 - Model structure diagram of YOLO11

Network Architecture of YOLO-TRS

Figure 3 presents the architecture of the YOLO-TRS. The model enhances detection accuracy and feature representation through the synergistic integration of DS-Conv and the CAA module. Specifically, DS-Conv module can dynamically adjust the shape and parameters of convolutional kernels based on the specific input data, allowing it to better adapt to the complex edges and textures within the image. This capability significantly enhances its performance in detecting tomato stems in complex backgrounds. The CAA module, as a lightweight attention mechanism, effectively addresses background interference and occlusion issues, enhancing the model's performance in detecting tomato fruit ripeness and stems. The synergistic interaction between these components substantially improves the accuracy and robustness of YOLO-TRS in tomato fruit ripeness and stem detection, while preserving a relatively low computational overhead.

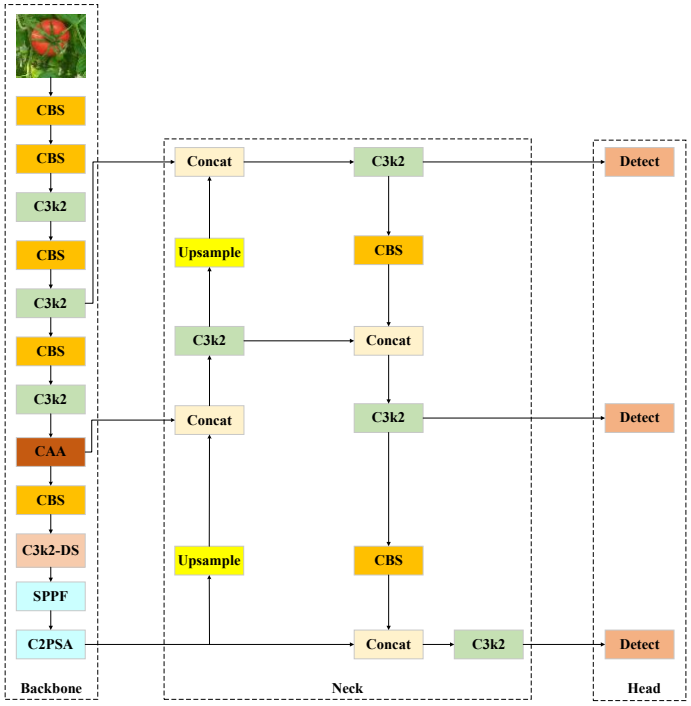
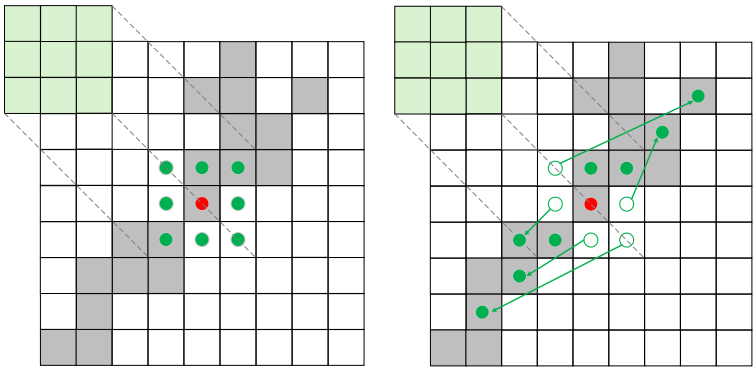


Fig. 3 - Model structure diagram of YOLO-TRS

C3k2-DS

The C3k2-DS module is designed by replacing the standard convolution in the C3k2 module with DS-Conv. Compared to standard convolution, the use of DS-Conv facilitates more effective extraction of features

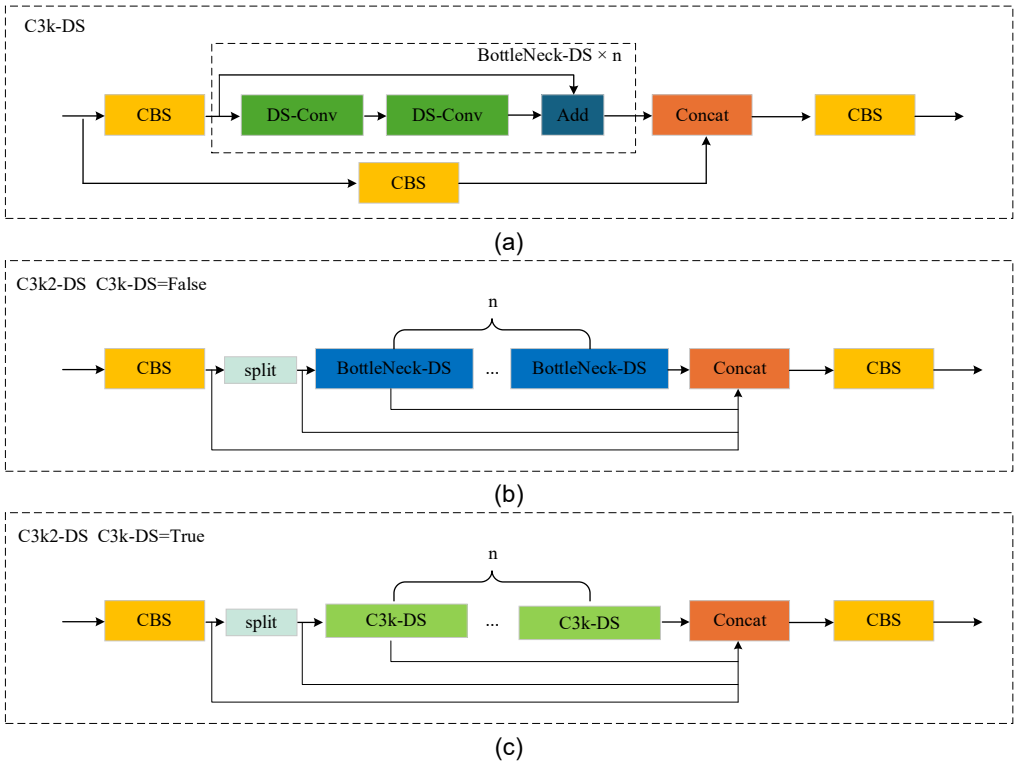
with varying geometric shapes. Due to its dynamic mechanism, the convolution kernel could adjust its shape dynamically to accommodate features from different regions. Consequently, the network can more effectively extract intricate features, capture irregularly shaped features with greater precision, adapt to input variations, expand the receptive field, and obtain a broader context. Figure 4 illustrates the difference between DS-Conv and standard convolution.



(a) Standard convolution (b) DS-Conv

Fig. 4 - Standard convolution and DS-Conv diagram

The proposed C3k2-DS module, integrated with DS-Conv, possesses a robust capability for extracting complex features. This enhanced capability provides a solid technical foundation for the accurate detection of fruit ripeness and stems in complex harvesting environments. The detailed structure of the C3k2-DS module is presented in Figure 5.



(a) the structure of C3k-DS. (b) the structure of C3k2-DS when C3k-DS is set to True. (c) the structure of C3k2-DS when C3k-DS is set to False.

Fig. 5 - The Structure of C3k2-DS

Context Anchor Attention (CAA)

To address the challenges of occlusion and background interference in complex orchard environments, the CAA attention mechanism was introduced, effectively mitigating these issues without significantly increasing the model's computational cost. The CAA mechanism employs global average pooling and one-dimensional strip convolutions to capture long-range pixel dependencies while simultaneously enhancing features in the central region. This design enables the CAA module to more effectively extract features of slender targets, such as fruit stems, and to reduce feature confusion caused by leaf occlusion or fruit overlap. The structure of the CAA module is illustrated in Fig. 6.

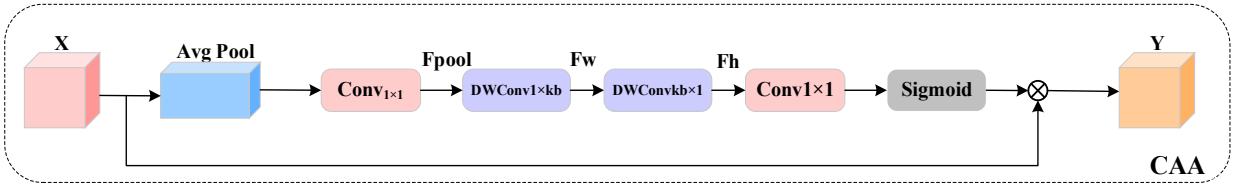


Fig. 6 - CAA module

First, local features are extracted through average pooling and 1×1 convolution, as illustrated in Equation 1.

$$F^{\text{pool}} = \text{Conv}_{1\times1}(\text{Pool}_{\text{avg}}(X)) \tag{1}$$

Next, vertical and horizontal strip depth convolutions are employed to approximate the conventional large-kernel depth convolution, as illustrated in Equations (2) and (3).

$$F^{w} = \text{DWConv}_{1\times k_b}(F^{\text{pool}}) \tag{2}$$

$$F^{h} = \text{DWConv}_{k_b\times1}(F^{w}) \tag{3}$$

The k_b parameter is configured as 11 + 2×1, which allows the strip convolution to acquire a sufficiently large receptive field for capturing strip-like structures, thus enhancing the feature extraction capability for elongated objects. This approach effectively reduces computational complexity while yielding performance comparable to that of standard large-kernel depth convolutions.

Ultimately, attention weights are computed via a 1×1 convolutional, followed by a Sigmoid activation function. The CAA operation is then executed by element-wise multiplication of the attention weights with the input features, as formalized in Equation (4)

$$Y = X \times \text{Sigmoid}(\text{Conv}_{1\times1}(F^{h})) \tag{4}$$

Model training

Experimental Environment and Parameter Settings

The performance and training efficiency of deep learning models are significantly influenced by configurations of hardware and the selection of hyperparameters. The appropriate choice of GPUs and CPUs plays a critical role in enhancing training efficiency. Regarding hyperparameters, settings including batch size, learning rate, epochs, and optimizer choice exert a considerable influence on model performance and generalization capability. This study was performed on a Windows 11 operating system with the PyTorch 2.0 development environment. **Error! Reference source not found.** and Table 4 provide a detailed overview of the model training environment and key parameter configurations.

Table 3

Model training environment

Environment	Details
GPU	NVIDIA GeForce RTX 4060Ti
CPU	Inter(R) Core i5-10200H@2.4GHz
Python	Python 3.8
CUDA	11.8
CuDNN	8.9.7

Table 4

Model training hyperparameters

Hyperparameters	Details
Epochs	350
Image Size	640×640
Batch Size	32
Optimizer	SGD
Initial learning rate	0.01

Model Evaluation Metrics

To evaluate the model's performance comprehensively, this study adopts a set of widely recognized metrics, including Precision (P), Recall (R), F1 score, mean Average Precision (mAP), gigaflops per second ($GFLOPs$), and the number of Parameters ($Params$).

$$P = \frac{TP}{TP + FP} \tag{5}$$

$$R = \frac{TP}{TP + FN} \tag{6}$$

where:

TP refers to the number of actual targets correctly identified and detected by the model. FP denotes the number of instances where the model incorrectly classifies a background or non-target region as a target. FN indicates the number of actual targets that the model fails to detect.

$$F1 = 2 \times \frac{P \times R}{P + R} \tag{7}$$

where:

$F1$ is calculated as the harmonic mean of precision and recall, offering a comprehensive evaluation of the model's overall performance.

$$AP = \int_0^1 P(r)dr \tag{8}$$

where:

P-R curve visually illustrates the trade-off between precision and recall at different decision thresholds. Typically, R is plotted on the x-axis and P on the y-axis. The closer the P-R curve is to the top-right corner, the better the model's performance in balancing precision and recall. The AP is determined by calculating the area under the curve for a given class, reflecting the model's ability to correctly identify the class across different confidence thresholds.

$$mAP = \frac{1}{N} \sum_{i=1}^{N} AP_i \tag{9}$$

where:

AP_i is the average precision for the i-th class, and N denotes the total number of classes. The mAP is obtained by calculating the arithmetic mean of the AP values across all categories, reflecting the overall performance of the model.

$$GFLOPs = \frac{2 \times H_{out} \times W_{out} \times K_h \times K_w \times C_{in} \times C_{out}}{10^9} \tag{10}$$

$$Params = (K_h \times K_w \times C_{in} + 1) \times C_{out} \tag{11}$$

where:

H_{in} and H_{out} are the output feature map sizes. K_h, K_w are the convolution kernel sizes. C_{in}, C_{out} are the number of input and output channels.

Experimental results and discussion

Design of the backbone network based on C3k2-DS

In this study, the C3k2 module in the original backbone network was replaced with the proposed C3k2-DS module. To explore the impact of its placement and quantity, comparative experiments were conducted with different configurations (as shown in Figure 7). As summarized in Table 5, the YOLO11-DS1 model—incorporating a single C3k2-DS module—achieved the highest detection accuracy. Compared to the baseline YOLO11, it improved the F1 score and mAP by 0.9% and 2.6%, respectively. This gain can be attributed to the dynamic snake convolution (DS-Conv), which adaptively adjusts the convolutional kernel shape based on input features, thereby enhancing the model's capacity to capture complex morphological traits of stems as well as texture and color characteristics of half-ripe fruits.

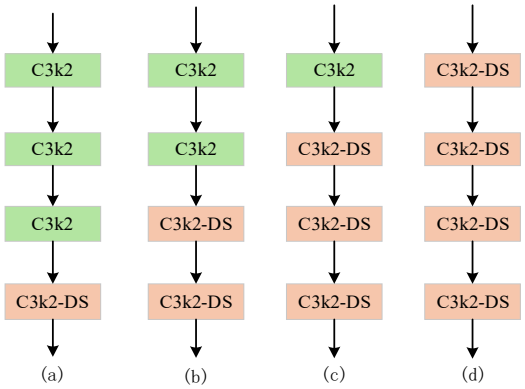
As the number of C3k2-DS modules increased, model accuracy began to fluctuate. For instance, when all four C3k2 modules were replaced (YOLO11-DS4), performance dropped relative to YOLO11-DS1. This decline is likely due to the increased model complexity and longer information pathways, where the structural nature of DS-Conv may impede efficient learning and retention of deep features.

Nevertheless, even YOLO11-DS4 still outperformed the baseline, with F1 and mAP rising by 0.8% and 0.6%, respectively, confirming the consistent benefit of integrating DS-Conv into the YOLO11 framework for tomato fruit and stem detection. Ultimately, this study selected YOLO11-DS1 as the new backbone network.

Table 5

Comparative results for different configurations of the backbone network

Model	F1/%	P/%	R/%	mAP/%	GFLOPs	Param/M
YOLO11	89.0	87.6	90.5	92.4	6.3	2.58
YOLO11-DS1	89.9	87.3	92.7	94.0	6.4	2.72
YOLO11-DS2	89.6	88.9	90.3	93.0	6.5	2.75
YOLO11-DS3	89.9	90.1	89.7	92.6	6.6	2.77
YOLO11-DS4	89.8	90.5	89.1	93.0	6.7	2.77



(a) One C3k2-DS structure, i.e., YOLO11-DS1, (b) Two C3k2-DS structure, i.e., YOLO11-DS2, (c) Three C3k2-DS structure, i.e., YOLO11-DS3, (d) Four C3k2-DS structure, i.e., YOLO11-DS4.

Fig. 7 - Different replacements of C3k2 in the backbone network

Experimental analysis of optimal placement of the CAA attention mechanism

The application of the CAA module can further enhance the performance of the YOLO11-DS1 algorithm, but its specific effect depends on the position of application and the specific requirements of the task. Therefore, this study conducted experiments by sequentially adding the CAA module to each layer of the backbone in the YOLO11-DS1 model. Table 6 presents the experimental results, where the numbers at the end of the model names indicate the layer number of the CAA module (with numbering starting from 0). It could be observed that the model's overall performance shows notable sensitivity to the variation in the position of the CAA. As the placement of the CAA module shifts from shallow to deep layers, the model's performance exhibits certain fluctuations. Specifically, the AP for the red and green categories shows a relatively low sensitivity to changes in the CAA position. In contrast, the AP for the half and stem categories demonstrates higher sensitivity to changes in the CAA position. This is likely due to the relatively stable appearance of the red and green fruits, whose features are simple and exhibit minimal shape variation across different layers. As a result, the change in CAA position has a limited impact on their detection precision. In contrast, the appearance features of semi-ripe fruits and stems are more complex and ambiguous, making them more susceptible to positional changes, leading to more significant fluctuations in AP values.

Furthermore, when the CAA module is applied to the 7th layer, the model's detection performance reaches its optimal level, with a mAP of 94.8% and an F1 score of 90.7%. The AP for the half and stem categories reaches 95.8% and 88.2%, respectively. The incorporation of the CAA module significantly improves the model's robustness in handling object detection tasks in complex backgrounds, making it more suitable for fruit ripeness and stem detection tasks in intricate environments. Based on the above analysis, the seventh layer of the backbone network was selected as the insertion point for incorporating the CAA module.

Table 6

Comparative results for different configurations of the CAA positions

Model	F1/%	P/%	R/%	mAP/%	GFLOPs	Param/M	AP/%			
							Red	Half	Green	Stem
YOLO11-DS1	89.9	87.3	92.7	94.0	6.4	2.72	97.0	93.5	98.7	86.9

YOLO11-DS1-CAA-4	87.8	90.6	85.1	93.6	6.5	2.73	97.2	96.3	98.5	82.3
YOLO11-DS1-CAA-5	89.9	89.4	90.4	92.9	6.5	2.76	96.5	93.3	98.6	83.3
YOLO11-DS1-CAA-6	89.9	88.7	91.2	93.0	6.5	2.76	97.0	93.1	98.4	83.6
YOLO11-DS1-CAA-7	90.7	89.9	91.5	94.8	6.5	2.76	96.3	95.8	98.9	88.2
YOLO11-DS1-CAA-8	86.3	89.6	89.0	92.2	6.5	2.86	96.7	90.6	97.9	83.4
YOLO11-DS1-CAA-9	90.2	88.4	92.1	94.0	6.5	2.86	96.6	93.8	98.4	87.4
YOLO11-DS1-CAA-10	89.7	88.8	90.7	92.5	6.5	2.86	96.6	92.0	97.8	83.6
YOLO11-DS1-CAA-11	88.9	87.1	90.8	92.7	6.5	2.86	96.8	91.9	98.3	83.6

Ablation experiments

Ablation experiments between different improvement methods

In this study, ablation experiments were conducted to evaluate the effectiveness of each improvement method. Specifically, the C3k2 module in the 9th layer was substituted with the C3k2-DS module, and the CAA module was added to the 7th layer of the backbone network. These two modifications were integrated into the original network model separately or in combination. The experimental results are presented in **Error! Reference source not found.**.

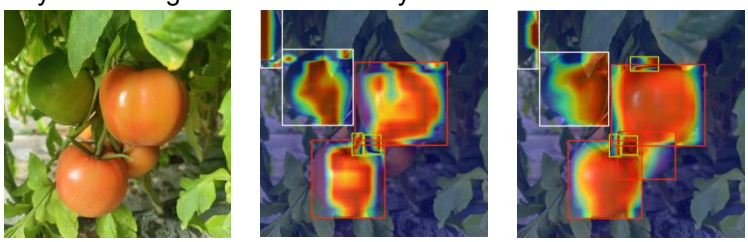
The introduction of the C3k2-DS module results in notable improvements in both the F1 score and mAP with only a slight increase in parameters and GFLOPs. In terms of individual categories, the C3k2-DS module significantly improves the model's detection performance for the stem and half-ripe tomatoes, with the AP value for stem increasing from 84.8% to 86.9%, and the AP value for half-ripe tomatoes increasing from 90.8% to 93.5%. The recognition accuracy of these two categories is particularly critical, as they are key bottlenecks that significantly limit the model's performance in practical applications. Furthermore, the incorporation of the CAA results in 1.1%, 0.4%, and 0.7% improvement in P, R, and mAP, respectively. While the inclusion of the CAA slightly increases the model's complexity, it significantly enhances its detection performance. Finally, when the C3k2-DS is combined with the CAA, the model's P, R, and mAP improve by 2.3%, 1.0%, and 2.4%, respectively. The AP of the half-ripe tomatoes and the stems, which were difficult to detect by the original YOLO11 model, increased by 5% and 3.4%, reaching 95.8% and 88.2%, respectively.

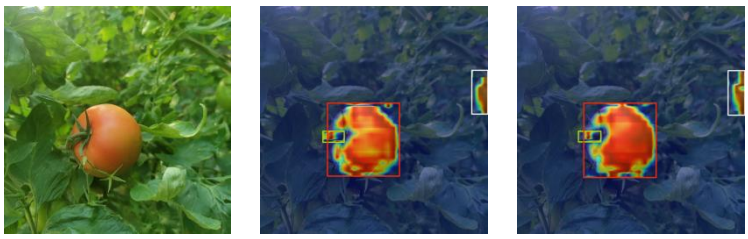
Table 7

Results of ablation experiments.

C3k2-DS	CAA	F1/%	P/%	R/%	mAP/%	GFLOPs	Param/M	AP/%			
								Red	Half	Green	Stem
×	×	89.0	87.6	90.5	92.4	6.3	2.58	95.5	90.8	98.4	84.8
√	×	89.9	87.3	92.7	94.0	6.4	2.72	97.0	93.5	98.7	86.9
×	√	89.8	88.7	90.9	93.1	6.4	2.62	96.2	92.1	98.2	85.7
√	√	90.7	89.9	91.5	94.8	6.5	2.76	96.3	95.8	98.9	88.2

To further analyze the impact of the CAA module, heatmaps at the CAA feature layer were generated. These heatmaps were then compared with those produced by the baseline model at the same feature layer. The results are presented in Figure 8. Each pixel in the heatmap corresponds to the activation level at the respective spatial location. Brighter regions, reflecting higher activation values, indicate a higher probability of the target's presence at those locations. It could be observed that after incorporating the CAA module, the heatmap shows a notable concentration in areas associated with tomatoes. This highlights the CAA module's effectiveness in refining the model's feature representation of tomatoes, enabling a more precise focus on the tomato fruit region, thereby enhancing detection accuracy.





(a) (b) (c)

(a) Original images. (b) Results without the incorporation of the CAA module. (c) Results with the incorporation of the CAA module.

Fig. 8 - Heatmaps of the CAA module ablation experiment

Ablation experiments of different attention methods

To identify the most effective attention mechanism for the proposed model, three attention modules—Coordinate Attention (CA) *(Hou et al., 2021)*, Efficient Multi-Scale Attention (EMA) *(Ouyang et al., 2023)*, and the proposed CAA—were comparatively evaluated in combination with both the standard C3k2 module and the enhanced C3k2-DS module. The experimental results are summarized in Table 8. When integrated with the standard C3k2 module, all three attention mechanisms improved model performance, though to varying degrees. The CAA module achieved the most balanced and significant gains in precision (P) and recall (R), along with a moderate improvement in mAP. In contrast, the CA module yielded the highest mAP among attention-enhanced variants but the lowest gains in P and R. After incorporating the C3k2-DS module, the model equipped with CAA achieved the best overall performance, attaining the highest F1-score and mAP among all configurations. These results validate that the CAA module provides more consistent and effective enhancement compared to CA and EMA, particularly when combined with structural feature extraction improvements.

Table 8

Results of ablation experiments with different attention methods.

Model	F1%	P/%	R/%	mAP/%	GFLOPs	Param/M	AP/%			
							Red	Half	Green	Stem
C3k2	89.0	87.6	90.5	92.4	6.3	2.58	95.5	90.8	98.4	84.8
C3k2-DS	89.9	87.3	92.7	94.0	6.4	2.72	97.0	93.5	98.7	86.9
C3k2+CA	89.1	87.9	90.4	93.5	6.3	2.59	98.0	94.3	97.6	84.3
C3k2+EMA	89.4	88.5	90.4	93.0	6.4	2.59	97.0	93.9	98.5	82.5
C3k2+CAA	89.8	88.7	90.9	93.1	6.4	2.62	96.2	92.1	98.2	85.7
C3k2-DS+CA	90.1	90.3	89.8	93.8	6.4	2.72	96.6	93.2	98.1	87.2
C3k2-DS+EMA	88.7	88.6	88.8	92.5	6.4	2.72	96.1	93.8	97.3	83.0
C3k2-DS+CAA	90.7	89.9	91.5	94.8	6.5	2.76	96.3	95.8	98.9	88.2

To identify the optimal convolutional enhancement strategy, the standard convolution in the bottleneck of the C3k2 module was replaced with three alternatives: DS-Conv, Dynamic Convolution (DNC) *(Chen et al., 2020)*, and Deformable Convolutional Networks v2 (DCNV2). Each variant was evaluated both with and without integration of the CAA module. The results are presented in Table 9. In the absence of the CAA module, the model equipped with DS-Conv achieved superior performance in both F1-score and mAP compared to those using DNC or DCNV2. After incorporating the CAA attention mechanism, the DS-Conv-based model continued to outperform all other configurations, attaining the highest F1-score (90.7%) and mAP (94.8%). These results demonstrate that DS-Conv provides more effective feature representation enhancement compared to current mainstream convolutional methods, making it the most suitable choice for improving detection performance in our task.

Table 9

Results of ablation experiments with different convolution methods

Model	F1/%	P/%	R/%	mAP/%	GFLOPs	Param/M	AP/%			
							Red	Half	Green	Stem
DS	89.9	87.3	92.7	94.0	6.4	2.72	97	93.5	98.7	86.9
DNC	88.1	86.2	90.1	92.3	6.3	2.81	96.0	90.6	98.0	84.7

DCNV2	88.9	89.4	88.4	92.3	6.3	2.63	95.2	89.8	98.1	85.9
DS+CAA	90.7	89.9	91.5	94.8	6.5	2.76	96.3	95.8	98.9	88.2
DNC+CAA	89.7	89.7	89.7	93.6	6.4	2.84	96.7	93.9	98.5	85.3
DCNV2+CAA	89.2	89.1	89.3	92.6	6.4	2.67	97.3	89.7	97.7	85.9

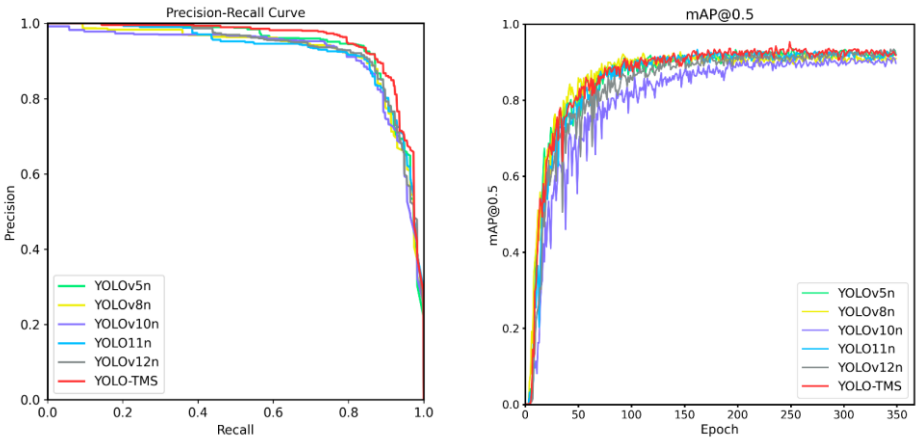
Comparative experiments with current mainstream models.

To evaluate the performance of the proposed YOLO-TRS model, comparative experiments were conducted against several current mainstream detection models under identical datasets and parameter settings. As summarized in Table 10, YOLO-TRS achieves the highest scores in mAP, precision, recall, and F1-score among all compared methods. This performance gain can be attributed to the enhanced capability of DS-Conv in extracting complex structural features, along with the CAA module's effectiveness in capturing rich contextual information. Together, these components enable the model to effectively handle challenges such as severe occlusion and background interference. Despite the performance improvements, YOLO-TRS maintains a moderate increase in parameter count, preserving real-time inference capability while achieving high detection accuracy. As illustrated in the P-R and mAP curves in Figure 9, YOLO-TRS consistently occupies the top position across evaluation metrics, further validating its overall superiority.

Table 10

Results of comparative experiments with current mainstream models.

Model	F1/%	P/%	R/%	mAP/%	GFLOPs	Param/M	AP/%			
							Red	Half	Green	Stem
YOLOv5n	89.3	88.4	90.3	93.7	7.1	2.50	96.4	93.8	97.6	86.8
YOLOv8n	88.5	87.3	89.8	92.1	8.1	3.01	95.7	91.0	98.4	83.3
YOLOv10n	87.9	86.8	89.1	91.5	8.2	2.70	94.6	92.0	97.6	82.0
YOLO11n	89.0	87.6	90.5	92.4	6.3	2.58	95.5	90.8	98.4	84.8
YOLOv12n	88.8	88.1	89.6	93.0	5.8	2.51	97.1	91.0	98.3	85.6
YOLO-TRS (ours)	90.7	89.9	91.5	94.8	6.5	2.76	96.3	95.8	98.9	88.2

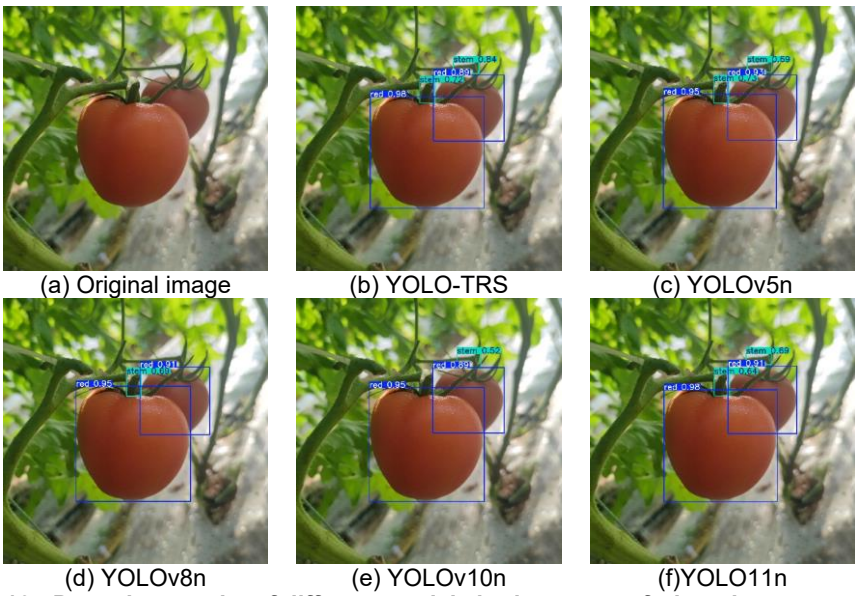


(a) P-R curves (b) mAP curves

Fig. 9 - P-R curves and mAP curves for different models

Analysis of Visual Results

To intuitively evaluate the detection performance of YOLO-TRS, experiments were conducted on the test sets, and the visual results were compared with those of several mainstream models. Figure 10 shows the comparative results in typical ripe tomato scenes during harvesting robot operation. The results indicate that, except for YOLO-TRS, all other models exhibit issues such as missed detections or inaccurate localization of tomato stems under occlusion and blur. Specifically, both YOLOv8n and YOLOv10n failed to detect all tomato stems, while YOLOv5n and YOLO11n detected all stems but with relatively low confidence. In contrast, our proposed YOLO-TRS not only accurately detects all tomato stems but also maintains higher confidence levels, demonstrating superior performance in stem detection among popular lightweight models.

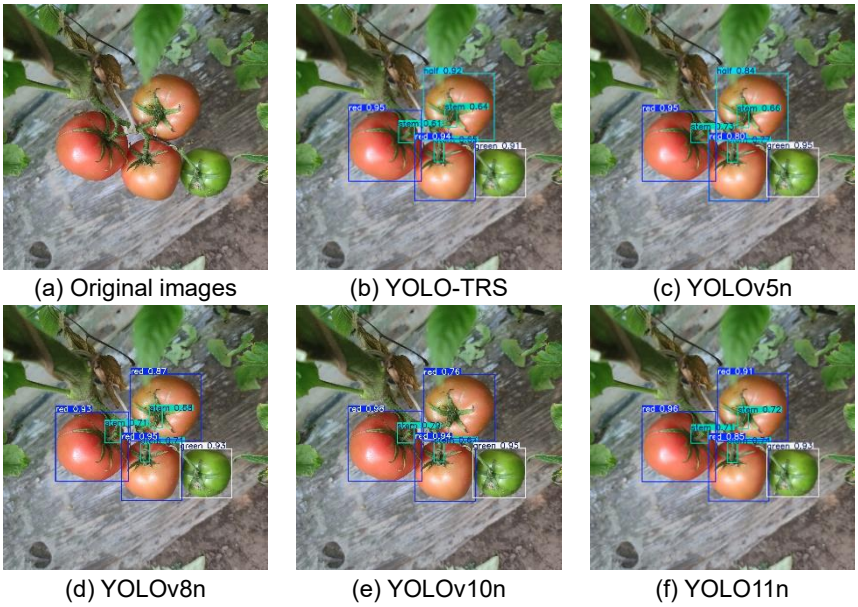


(a) Original image (b) YOLO-TRS (c) YOLOv5n

(d) YOLOv8n (e) YOLOv10n (f)YOLO11n

Fig. 10 - Detection results of different models in ripe tomato fruit and stems scenes

Figure 11 presents a detection scenario containing half-ripe, green, and red tomatoes. In the challenging task of identifying half-ripe tomatoes, YOLOv8n, YOLOv10n, and YOLO11n all produced false detections by misclassifying half-ripe tomatoes as red. Such errors could lead to the premature harvesting of half-ripe fruit, resulting in potential waste. Although YOLOv5n correctly recognized half-ripe tomatoes, its confidence was lower than that of YOLO-TRS. These results confirm the superior performance of YOLO-TRS in tomato ripeness detection. It is also noted that in this set of results, none of the models successfully detected stems on green tomatoes, which may be attributed to their smaller size and color similarity to the fruit. Overall, the above analysis validates that YOLO-TRS achieves accurate detection in both ripeness classification and stem localization.



(a) Original images (b) YOLO-TRS (c) YOLOv5n

(d) YOLOv8n (e) YOLOv10n (f) YOLO11n

Fig. 11 - Detection results of different models in various ripe tomato fruit and stems scenes.

Conclusions

This study proposes YOLO-TRS, an enhanced model based on YOLO11n, specifically designed for joint detection of tomato ripeness and stem position. With its compact architecture, the model is well-suited for deployment on resource-constrained harvesting robots, achieving short inference times while improving detection accuracy, thereby contributing to higher harvesting efficiency. The proposed improvements include a novel backbone architecture built around the C3k2-DS module, which significantly enhances the model's ability to capture complex structural features with only a minimal increase in parameters. Furthermore, the integration of the CAA module strengthens the model's focus on foreground objects and its contextual

understanding, enabling more reliable identification of tomatoes at different maturity stages and stems in complex field environments.

Experimental results demonstrate that YOLO-TRS outperforms several current mainstream lightweight models, achieving an F1-score of 90.7% and a mAP of 94.8%, while maintaining a compact model size (2.76 M parameters) and low computational cost (6.5 GFLOPs). Compared to YOLOv5n, YOLOv8n, YOLOv10n, YOLO11n, and YOLOv12n, YOLO-TRS improves F1-score by 1.4%, 2.2%, 2.8%, 1.7%, and 1.9%, and increases mAP by 1.1%, 2.7%, 3.3%, 2.4%, and 1.8%, respectively. Visualization results further validate the model's superior accuracy in detecting both tomato ripeness and stem positions under real-world conditions.

In conclusion, YOLO-TRS exhibits outstanding performance in the joint detection of tomato ripeness and stem location, offering an effective technical solution for intelligent fruit harvesting in agricultural applications.

ACKNOWLEDGEMENTS

This project is financially supported by the National Nature Science Foundation of China (52165006), Fundamental Research Program of Shanxi Province (202203021212450). The authors declare no competing interests.

REFERENCES

[1] Cai, X., Lai, Q., Wang, Y., Sun, Z., & Yao, Y. (2024). Poly kernel inception network for remote sensing detection. *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, USA, pp. 27706-27716.

[2] Chen, Y., Dai, X., Liu, M., Chen, D., Yuan, L., & Liu, Z. (2020). Dynamic convolution: attention over convolution kernels. *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, USA, pp. 11027-11036.

[3] Feng, Q., Cheng, W., Yang, Q., Xun, Y., & Wang, X. (2015). Identification and localization of overlapping tomatoes based on linear structured light vision system (基于线结构光视觉的番茄重叠果实识别定位方法研究). *Journal of China Agricultural University*, 20(04), 100-106.

[4] Gao, G., Shuai, C., Wang, S., & Ding, T. (2024). Using improved YOLOV5s to recognize tomatoes in a continuous working environment. *Signal, Image and Video Processing*, 18(05), 4019-4028.

[5] Ge, Y., Lin, S., Zhang, Y., Li, Z., Cheng, H., Dong, J., & Shao, S. (2022). Tracking and counting of tomato at different growth period using an improving YOLO-Deepsort network for inspection robot. *Machines*, 10(06), 489.

[6] Goel, N., & Sehgal, P. (2015). Fuzzy classification of pre-harvest tomatoes for ripeness estimation - An approach based on automatic rule learning using decision tree. *Applied Soft Computing*, 36, 45-56.

[7] He, B., Zhang, Y., Gong, J., Fu, G., Zhao, Y., & Wu, R. (2022). Fast recognition of tomato fruit in greenhouse at night based on improved YOLOv5 (基于改进 YOLOv5 的夜间温室番茄果实快速识别). *Transactions of the Chinese Society of Agricultural Machinery*, 53(05), 201-208.

[8] Hou, G., Chen, H., Niu, R., Li, T., Ma, Y., & Zhang, Y. (2025). Research on multi-layer model attitude recognition and picking strategy of small tomato picking robot. *Computers and Electronics in Agriculture*, 232, 110125.

[9] Hou, Q., Zhou, D., & Feng, J. (2021). Coordinate attention for efficient mobile network design. *In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Nashville, USA, pp. 13708-13717.

[10] Kanna, S., Kumaraperumal, R., Pazhanivelan, P., Jagadeeswaran, R., & Prabu, P. C. (2024). YOLO deep learning algorithm for object detection in agriculture: a review. *Journal of Agricultural Engineering - Italy*, 55(04), 1641.

[11] Li, J., Xiang, C., Wang, X., Guo, Y., Huang, Z., Liu, L., & Li, X. (2021). Current situation of tomato industry in China during “The Thirteenth Five-Year Plan” period and future prospect (“十三五”我国番茄产业现状及展望). *China Vegetables*, 1(02), 13-20.

[12] Li, T., Sun, M., Ding, X., Li, Y., Zhang, G., Shi, G., & Li, W. (2021). Tomato recognition method at the ripening stage based on YOLOv4 and HSV (基于 YOLOv4+HSV 的成熟期番茄识别方法). *Transactions of the Chinese Society of Agricultural Engineering*, 37(21), 183-190.

[13] Liu, F., Liu, Y., Lin, S., Guo, W., Xu, F., & Zhang, B. (2020). Fast recognition method for tomatoes under complex environments based on improved YOLO (基于改进型 YOLO 的复杂环境下番茄果实快速识别方法). *Transactions of the Chinese Society of Agricultural Machinery*, 51(06), 229-237.

[14] Liu, G., Nouaze, J., Mbouembe, P. L. T., & Kim, J. H. (2020). YOLO-tomato: A robust algorithm for tomato detection based on YOLOv3. *Sensors*, 20(07), 2145.

[15] Liu, J., He, J., Chen, H., Wang, X., & Zhai, H. (2023). Development of detection model for tomato clusters based on improved YOLOv4 and ICNet (基于改进 YOLO v4 和 ICNet 的番茄串检测型). *Transactions of the Chinese Society of Agricultural Machinery*, 54(10), 216-224+254.

[16] Qi, Y., He, Y., Qi, X., Zhang, Y., & Yang, G. (2023). Dynamic snake convolution based on topological geometric constraints for tubular structure segmentation. *In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, Paris, France, pp. 6047-6056.

[17] Ouyang, D., He, S., Zhang, G., Luo, M., Guo, H., Zhan, J., & Huang, Z. (2023). Efficient multi-scale attention module with cross-spatial learning. *In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, Rhodes Island, Greece, pp. 1-5.

[18] Sun, X. (2024). Enhanced tomato detection in greenhouse environments: a lightweight model based on S-YOLO with high accuracy. *Frontiers in Plant Science*, 15, 1451018.

[19] Tian, Z., Hao, H., Dai, G., & Li, Y. (2024). Optimizing tomato detection and counting in smart greenhouses: A lightweight YOLOv8 model incorporating high- and low-frequency feature transformer structures. *Neural Computing and Applications*, 2024, 1-37.

[20] Wang, C., Wang, C., Wang, L., Wang, J., Liao, J., Li, Y., & Lan, Y. (2023). A lightweight cherry tomato maturity real-time detection algorithm based on improved YOLOV5n. *Agronomy-Basel*, 13(08), 2106.

[21] Wei, J., Ni, L., Luo, L., Chen, M., You, M., Sun, Y., & Hu, T. (2024). GFS-YOLO11: a maturity detection model for multi-variety tomato. *Agronomy-Basel*, 14(11), 2644.

[22] Wu, M., Lin, H., Shi, X., Zhu, S., & Zheng, B. (2024). MTS-YOLO: A multi-task lightweight and efficient model for tomato fruit bunch maturity and stem detection. *Horticulturae*, 10(09), 1006.

[23] Yang, J., Qian, Z., Zhang, Y. J., Qin, Y., & Miu, H. (2022). Real-time recognition of tomatoes in complex environments based on improved YOLOv4-tiny (采用改进 YOLOv4-tiny 的复杂环境下番茄实时识别). *Transactions of the Chinese Society of Agricultural Engineering*, 38(09), 215-221.

[24] Zhang, J., Bi, Z., Yan, Y., Wang, P., Hou, C., & Lv, S. (2023). Fast recognition of greenhouse tomato targets based on attention mechanism and improved YOLO (基于注意力机制与改进 YOLO 的温室番茄快速识别). *Transactions of the Chinese Society of Agricultural Machinery*, 54(05), 236-243.

[25] Zhou, H., Wang, X., Au, W., Kang, H., & Chen, C. (2022). Intelligent robots for fruit harvesting: Recent developments and future challenges. *Precision Agriculture*, 23(05), 1856–1907.