# TOMATO MATURITY DETECTION METHOD Based ON YOLOv11n-SDS
# /
# 基于 YOLOv11n-SDS 的番茄成熟度检测方法

**ZiLu HUANG[1], ChengJun ZHAI[2], YueYing GUO[3], HongBo WANG*[1]**
[1] College of Mechanical and Electronic Engineering, Inner Mongolia Agricultural University, Hohhot / China
[2] Inner Mongolia Autonomous Region Education Examination Authority, Hohhot / China
[3] College of Food Science and Engineering, Inner Mongolia Agricultural University, Hohhot / China
*Tel: +86 13739981395; E-mail: wanghb@imau.edu.cn*
*DOI: https://doi.org/10.35633/inmateh-78-80*

***Keywords:*** *tomato maturity, YOLOv11n, object detection, SPD-Conv, C3K2_DLKA, SDI*

**ABSTRACT**

*To address the low speed and limited accuracy of tomato maturity detection in complex greenhouse environments characterized by dense distribution, overlap, and occlusion, this study proposes YOLOv11n-SDS, an improved algorithm based on YOLOv11n. The key enhancements include: (1) a Spatial Pyramid Depthwise Separable Convolution (SPD-Conv) module; (2) the integration of a Deformable Large-Kernel Attention (DLKA) mechanism into the backbone C3K2 module; and (3) a Semantic and Detail Injection (SDI) module replacing the Concat operation in the neck network. These improvements enhance the detection of small and low-resolution targets, as well as occluded fruits under challenging lighting and background conditions. Experimental results show that YOLOv11n-SDS improves mAP@0.5 by 2.4%, recall by 1.3%, and precision by 3.2%, while maintaining a low computational cost of 9.1 GFLOPs. Compared with existing models, including RT-DETR, Faster R-CNN, SSD, and other YOLO variants, the proposed model achieves a superior balance between accuracy, efficiency, and practical applicability. Furthermore, the model was deployed and validated on a mobile robotic platform in a greenhouse environment, enabling real-time tomato maturity detection and 3D target localization. These results demonstrate its strong potential for practical applications, such as ripeness monitoring and harvesting-oriented perception.*

**摘要**

*针对复杂温室环境下番茄存在密集分布、重叠以及叶片遮挡等情况，使用现有的目标检测算法进行番茄成熟度检测存在速度慢、识别准确率低等问题，本研究提出一种基于 YOLOv11n 的改进算法 YOLOv11n-SDS。其核心改进包括：（1）引入空间金字塔深度可分离卷积（SPD-Conv）模块；（2）在主干网络 C3K2 模块中加入可变形大核注意力（DLKA）机制；（3）采用语义细节注入（SDI）模块替代颈部网络中的 Concat 连接。这些改进有效提升了对低分辨率图像和小尺寸番茄的识别能力，显著增强了在复杂光照与背景干扰下对遮挡果实的检测性能。实验表明，YOLOv11n-SDS 的 mAP@0.5 提升 2.4%，召回率提高 1.3%，精确率增长 3.2%，计算量仅为 9.1 GFLOPs。与 RT-DETR、Faster-RCNN、SSD 及 YOLO 系列等模型相比，本算法在检测精度、计算效率和实际应用性方面取得更优平衡。此外，本文在温室环境下将所提模型部署到移动机器人平台上进行了应用验证，实现了番茄成熟度的在线检测与三维定位，表明该方法在成熟度监测和面向采摘的目标感知方面具有良好的实际应用潜力。*

## INTRODUCTION

Tomatoes are highly valued for their rich nutritional content and flavor, making them one of the most extensively cultivated crops in China *(Collins et al., 2022; Gonzali et al., 2021)*. Currently, manual harvesting remains the predominant method, which is characterized by high cost, low efficiency, and significant labor shortages during peak seasons. The realization of efficient automated harvesting therefore relies on accurate and robust target detection technology *(Chen et al., 2025; Liu et al., 2024)*. However, in natural orchard environments, tomato fruits are frequently occluded by leaves and branches, grow in dense clusters, and are presented against complex backgrounds. Moreover, their appearance is highly sensitive to varying lighting conditions, making the precise identification of maturity stages particularly challenging.

*ZiLu HUANG, M.S. Stud.; HongBo Wang, Professor, Correspondent author*

Consequently, the development of a detection algorithm capable of rapid and accurate maturity assessment is crucial for advancing automated tomato harvesting systems.

In recent years, the swift evolution of deep learning technology has led to substantial improvement in agriculture, particularly in the classification of fruit and vegetable maturity and precise detection, showcasing the considerable benefits of intelligent harvesting *(Hua et al., 2022)*. Presently, the predominant algorithms in object detection comprise one-stage detection algorithms grounded in regression and two-stage detection strategies reliant on candidate areas. Two-stage detection methods initially provide possible item candidate regions, subsequently followed by categorization and positional refining of these regions. Common algorithms comprise R-CNN *(Girshick et al., 2014)*, Fast R-CNN *(Girshick et al., 2015)*, and Faster R-CNN *(Ren et al., 2015)*. *Seo D et al., (2021)*, proposed a real-time robotic detection system utilizing Faster R-CNN for monitoring tomato growth, employing a color model resilient to varying lighting circumstances to establish an image-based standard for tomato fruit ripening. *Wang Z et al., (2022),* developed an enhanced Faster R-CNN model, MatDet, for the detection of tomato maturity. They utilized RoIAlign during the feature mapping phase to get more precise bounding boxes, hence tackling the difficulty of identifying tomato maturity in intricate environments. *Wang C et al., (2023),* proposed an R-CNN model for tomato recognition and segmentation, utilizing the Swin Transformer as the backbone network to improve feature extraction. This method not only accurately identifies tomatoes among cherry tomato varieties but also differentiates various maturity stages. Although these methods exhibit remarkable accuracy and robustness, they are hindered by substantial processing demands, considerable disk space requirements, and extended detection periods, rendering them inappropriate for real-time field applications.

Single-stage detection algorithms directly anticipate item categories and locations within photos, bypassing the generation of candidate regions, exemplified by the YOLO series. In contrast to two-stage detection algorithms, single-stage methods provide expedited detection rates and enhanced scalability, rendering them more appropriate for real applications. *SN Appe et al., (2023),* developed a tomato identification model with YOLOv5, incorporating the CBAM attention mechanism into the network architecture to proficiently identify overlapping tiny tomatoes, attaining an average accuracy of 88.1%. This investigation encountered difficulties due to limited detection precision. *Miao et al., (2023),* proposed an enhanced lightweight YOLOv7-based approach for detecting the maturity of cherry tomatoes. Experiments revealed that this lightweight model attained visually acceptable outcomes, with the enhanced model's mean average precision (mAP) rising by 0.1%. *Wu et al., (2024),* presented the MTS-YOLO model, attaining a mAP@0.5 of 92.0% for the identification of tomato maturity and stems. *Wang S et al., (2024),* proposed an enhanced YOLOv8 algorithm for detecting tomato maturity in intricate environments, attaining a mean Average Precision (mAP) of 86.9% and a recall rate of 82.0%. These investigations on YOLO-based enhancement algorithms further corroborate the adaptability of the YOLO series and illustrate the framework's significant compatibility and versatility.

Presently, despite the notable advancements in maturity detection algorithms, their accuracy remains significantly inadequate when utilized for tomato maturity assessment. This is especially applicable in intricate environmental conditions, including leaf and stem obstruction, fruit overlap, and light interference, as well as when identifying neighboring fruits with comparable maturity traits. This manuscript suggests enhancements to the YOLOv11n algorithm. The incorporation of Spatial Pyramid Depthwise Separable Convolution (SPDConv) and Deformable Large Kernel Attention (DLKA) into the backbone network, together with the substitution of the Concat module in the neck network with an SDI module, improves the model's responsiveness to geometric deformations of occluded tomatoes. This efficiently mitigates intricate lighting and background disturbances, markedly enhancing detection precision for diminutive tomato targets.

## MATERIALS AND METHODS

### DATA SAMPLE COLLECTION AND DATASET CREATION

The tomato image data were collected from a tomato picking garden located in the suburbs of Hohhot, where standardized cultivation techniques are applied. Considering that varying angles can impact recognition accuracy, 1,650 tomato images were photographed under diverse weather and lighting conditions to ensure dataset diversity and representativeness. These images encompassed different shooting angles, distances, mixed maturity levels, fruit occlusions, and overlapping foliage. Greenhouse tomatoes under various conditions are shown in Figure 1.

To enhance the diversity of training data, prevent model overfitting, and improve model generalization, data augmentation was applied to the collected raw images. This included randomly combining techniques such as rotation, scaling, cropping, noise addition, and brightness adjustment, expanding the dataset to 4,650 images. Tomato images were manually annotated using the LabelImg tool. Following national standards, the tomato dataset was categorized into three stages: green stage (red coverage 0-40%), semi-ripe stage (red coverage 40%-70%), and ripe stage (red coverage 70%-100%), labeled as green, salmon pink, and red, respectively. To meet the training requirements of the detection model, image pixels were resized to 640×640. The ratio of the training, validation, and test sets was set to 7:1:2.

a. Fruit occlusion b. Front light c. Mixed maturity levels d. Different shooting angles e. overlapping foliage

**Fig. 1 - Images of tomatoes in a greenhouse in different scenes**

## IMPROVEMENTS TO THE TOMATO OBJECT DETECTION MODEL

### YOLOv11n OBJECT DETECTION MODEL

This paper utilizes YOLOv11n as its foundational model, showcasing its architecture in Figure 2.

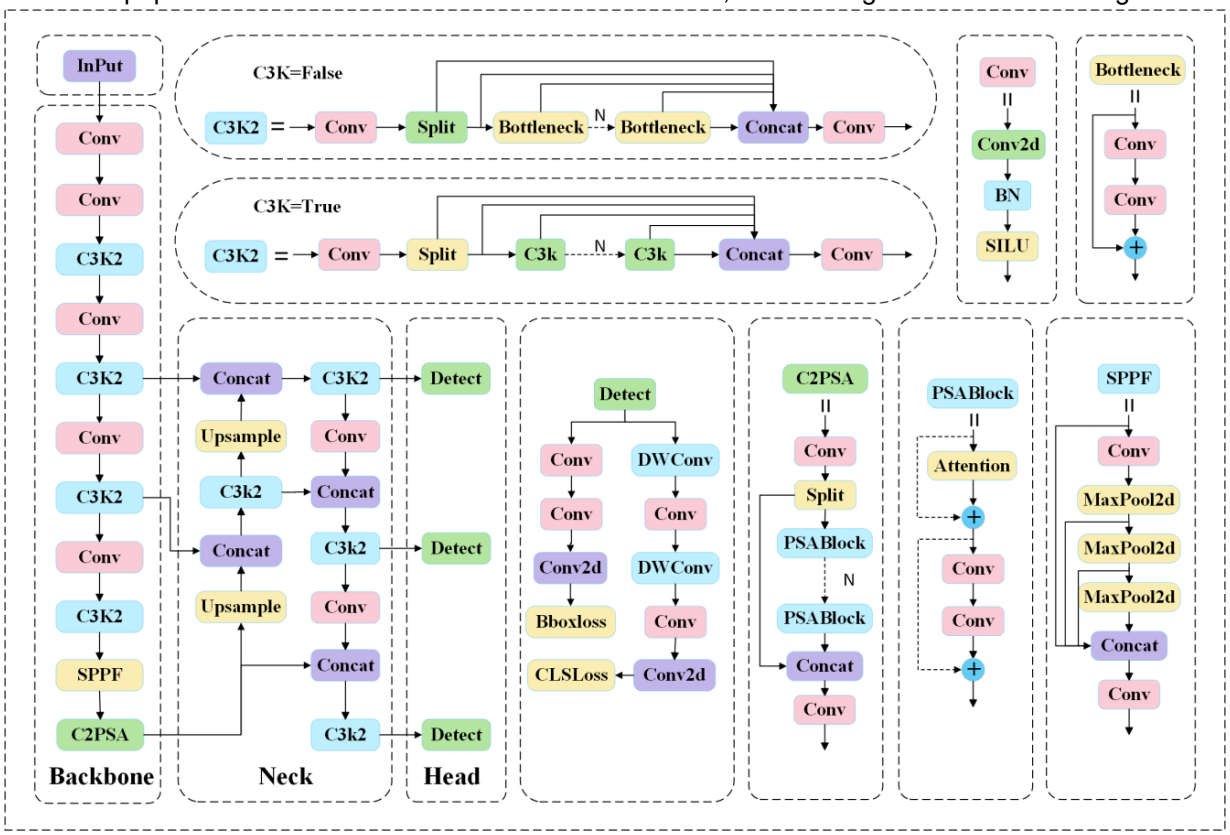



**Fig. 2 - Structural diagram of the improved YOLOv11**

The YOLOv11 model is constructed based on the CSP network and enhances the C2F module of YOLOv8 by introducing the C3K2 module. By utilizing the C3K2 module as the core structure for feature extraction and fusion, the ability of the model to capture key features and recognize detailed object characteristics is significantly strengthened, thereby improving its flexibility and configurability *(Zhihao et al., 2025)*. It retains the Spatial Pyramid Pooling (SPPF) module. Furthermore, building upon the original YOLOv10 architecture, the PSA module following SPPF is upgraded to the C2PSA module to further enhance the model's feature extraction capabilities.

YOLOv11 incorporates the head design philosophy of YOLOv10, employing depthwise separable convolutions to reduce redundant computations and improve efficiency *(Gan et al., 2025).* Selectively introducing residual structures optimizes gradient propagation, further enhancing training effectiveness. YOLOv11 strengthens its feature extraction capabilities and boosts overall performance by integrating attention mechanisms and feedforward neural networks. Compared to previous YOLO models, YOLOv11 demonstrates outstanding performance *(Sapkota et al., 2025).* The YOLOv11 series comprises five distinct architectures: YOLOv11n, YOLOv11s, YOLOv11m, YOLOv11l, and YOLOv11x. The primary distinction among these variants lies in the configuration of feature extraction modules and convolutional filters at specific positions within the network. From YOLOv11n to YOLOv11x, both the model size and the number of parameters progressively increase.

## IMPROVED YOLOV11N OBJECT DETECTION MODEL

The improved model architecture is illustrated in Figure 3. The improved model architecture is illustrated in Figure 3. The convolutional layers of the backbone network are restructured using Spatial Pyramid Depthwise Separable Convolution (SPD-Conv) to establish a multi-scale feature fusion mechanism. This mechanism effectively optimizes the model, enhancing its detection performance for low-resolution images and small objects. Within the backbone network, the original attention mechanism is replaced with Deformable Large-Kernel Attention (DLKA). By dynamically adjusting the sampling positions of large convolutional kernels, this approach enhances the model's adaptability to geometric deformations of occluded objects. For the feature concatenation layer in the neck network, a Spatial Depth Interaction (SDI) module is introduced. This module enhances the contextual relevance of important features through cross-level feature interaction and channel reweighting mechanisms, thereby improving the model's robustness to complex lighting conditions. Simultaneously, it reduces the model's dependence on background noise, significantly boosting detection accuracy and scene adaptability.

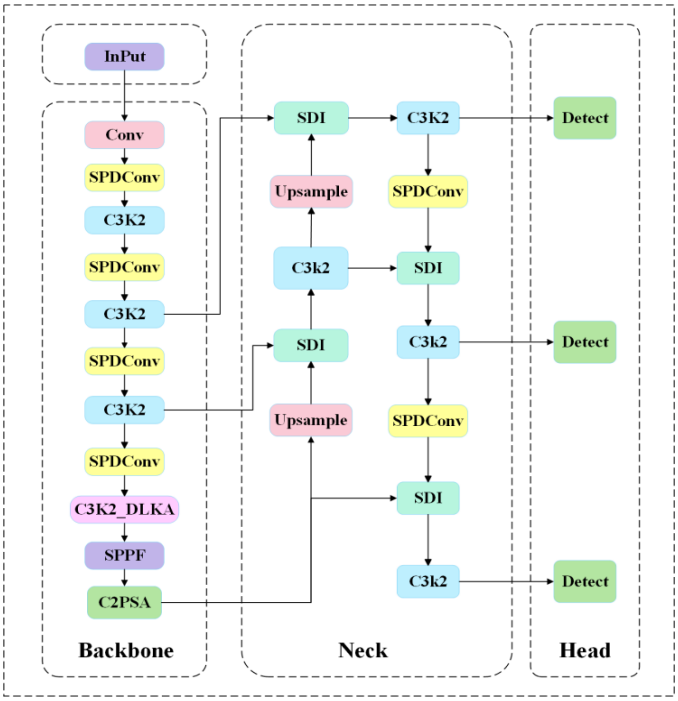



**Fig. 3 - Improved YOLOv11n object detection module**

### SPD-Conv module

In tomato detection under complex conditions, the accuracy for small targets such as unripe tomatoes remains considerably lower than for normal-sized objects. Small targets occupy fewer pixels and provide limited background information during training. Moreover, they often appear together with larger objects, which tend to dominate the learning process, leading to missed detections of smaller instances and reduced reliability of convolutional neural networks. This issue stems primarily from the use of strided convolutions and pooling layers in conventional architectures. While these operations work well for high-resolution images with moderate-sized objects by efficiently skipping redundant pixels, they cause significant loss of fine-grained details and impair feature learning when dealing with small, blurry targets.

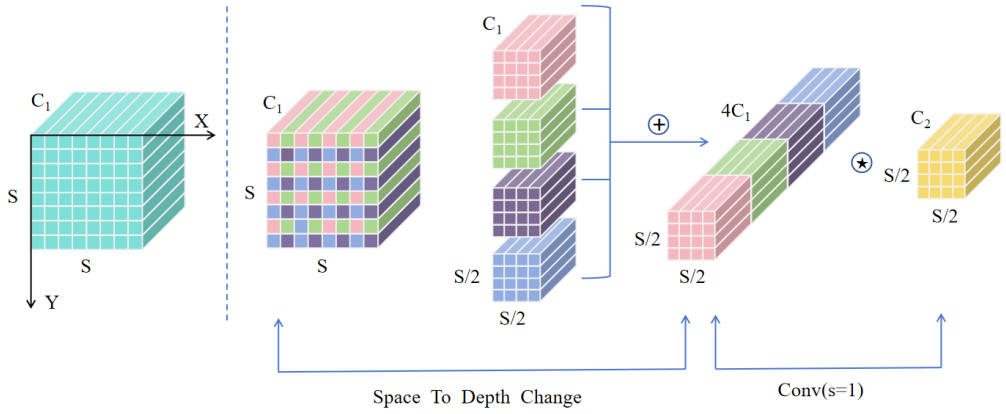



**Fig. 4 - SPD-Conv module**

To address this issue, a space-to-depth convolution (SPDConv) module *(Sunkara et al., 2022; Wenjie et al., 2025)* is introduced to replace standard convolutional layers in both the backbone and neck networks of YOLOv11n. SPDConv captures more detailed features through sparse sampling, converting spatial information into channel-wise information, thereby improving detection performance on small objects and low-resolution images. The module consists of a space-to-depth (SPD) layer followed by a non-strided 1×1 convolution. As shown in Figure 4, the SPD component partitions the input feature map of size $S \times S \times C_1$ into scale×scale sub-maps of size (S/scale)×(S/scale)×$C_1$ using skip sampling. These are then concatenated into a feature map of size (S/scale)×(S/scale)×$scale^2C_1$. A subsequent 1×1 convolution reduces dimensionality while retaining critical features. With scale set to 2, this approach preserves feature map channels while halving spatial size, maximizing feature retention and effectively improving the discrimination of small targets from background interference.

**Deformable-LKA Module**

In tomato maturity recognition, variations in fruit shape, size, and orientation, as well as geometric deformations caused by leaf occlusion, require the model to have strong local feature extraction capabilities. The conventional attention mechanism in the C3K2 module employs a fixed receptive field, which limits its ability to handle occluded targets and accurately represent deformed tomato features. To overcome this limitation, a Deformable Large Kernel Attention (DLKA) module *(Jigang et al., 2025)* is introduced, as illustrated in Figure 5.

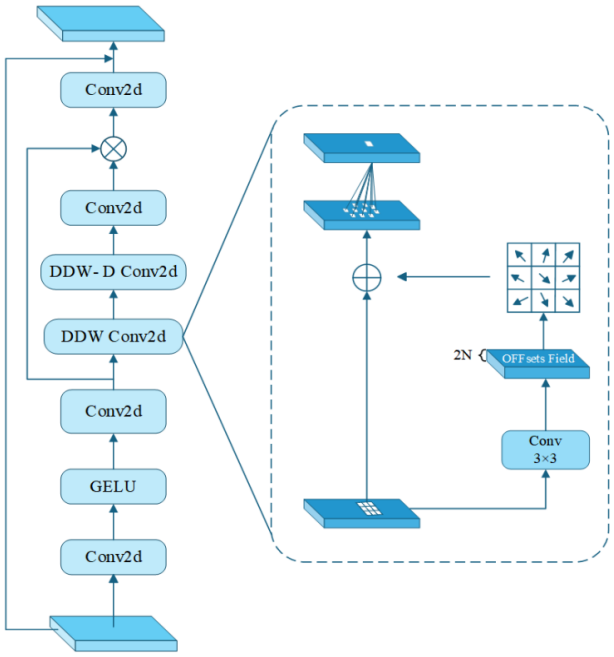



**Fig. 5 - Deformable-LKA module**

The DLKA leverages large-kernel attention to infer target information from broader contextual cues, while deformable convolutions allow the sampling grid to adaptively adjust through learnable offsets derived from the feature maps. This adaptability enables the generation of feature-specific convolutional kernels, enhancing the clarity and accuracy of target boundaries. As a result, the model improves its capability to segment and identify tomatoes against complex backgrounds and better captures features from deformed objects. By integrating the DLKA into the C3K2 module of the YOLOv11 backbone network, the model focuses more effectively on key characteristics of tomato fruits, thereby increasing detection accuracy and robustness for occluded targets.

The deformable large kernel attention structure is defined as shown in Equation (1) *(Azad et al., 2024)*:

$$\begin{cases} \text{Attention} = \text{Conv1} \times 1(\text{DDW} - \text{D} - \text{Conv}(\text{DDW} - \text{Conv}(\text{F}'))) \\ Output = Conv1 \times 1(\text{Attention} \otimes \text{F}') + F \end{cases} \quad (1)$$

The input features are denoted by $F \in R^{C \times H \times W}$. $\text{F}' \in \text{GELU}(Conv(F))$. DDW-Conv represents the Deform-DW two-dimensional convolution operation in the figure, while DDW-D-Conv denotes the Deform-DW-D two-dimensional convolution operation. The operator $\otimes$ signifies element-wise multiplication.

**SDI Module**

To address the challenges of complex lighting variations and background interference in tomato maturity detection under greenhouse conditions, this study introduces an improved feature fusion module for the YOLOv11 model. Conventional concatenation operations often fail to effectively distinguish fruit features from background noise when integrating multi-level features, resulting in increased misdetection under challenging conditions such as backlighting and leaf occlusion. To overcome this limitation, a Semantic and Detail Injection (SDI) module, adapted from the UNetV2 architecture *(Peng et al., 2025; Meng et al., 2025)*, is proposed. The module employs a dual-path design to enhance feature integration. In the semantic path, a channel attention mechanism dynamically weights high-level features, emphasizing color attributes associated with maturity. Input features are first compressed using global average pooling to obtain channel-wise statistics, followed by two fully connected layers that generate adaptive channel weights. In the detail path, an enhanced spatial attention mechanism utilizes parallel max-pooling and average-pooling operations to capture textural details of the tomato surface. Combined with 3×3 convolutions, this mechanism produces spatial weight maps that prioritize retaining fine details near the calyx region. The fusion stage employs the Hadamard product to enable deep interaction between semantic and detail features. To facilitate network optimization, the SDI module incorporates residual connections to mitigate feature degradation in deep layers. By leveraging both spatial and channel attention mechanisms, the module effectively aggregates multi-level features, enriching semantic representation and preserving structural details while strengthening the model's resistance to interference. The overall architecture of the SDI module is depicted in Figure 6. These improvements not only enhance detection accuracy but also bolster the model's robustness in complex environments.

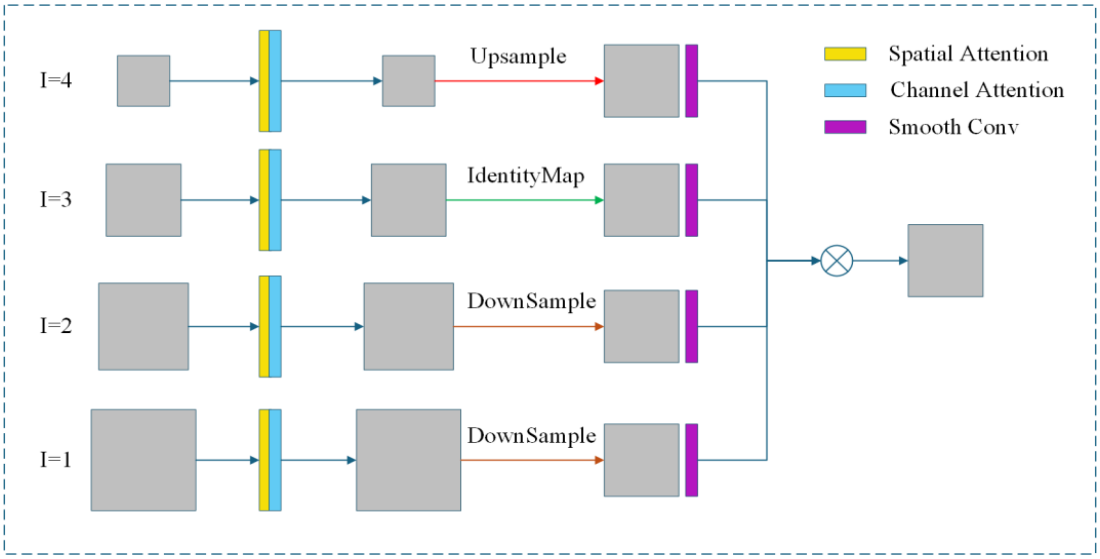



**Fig. 6 - SDI module**

## RESULTS AND ANALYSIS

### Test environment

This study was conducted in a standardized experimental environment. The operating system used for training and evaluation was Windows 11, with a 13th Gen Intel® Core™ i7-13620H CPU (2.40 GHz), an NVIDIA GeForce RTX 4060 Laptop GPU, and 16 GB of RAM.

The implementation was based on Python 3.9.19 and the PyTorch 2.3.0 deep learning framework, with CUDA version 11.8. The image data were subjected to standardized preprocessing procedures, including resolution adjustment and data augmentation, to reduce potential bias and improve model generalization. The experimental platform adhered to reproducibility and verifiability principles in parameter configuration, ensuring the reliability of the experimental results and the reproducibility of the research findings *(JunMao et al., 2025; Liqing et al., 2025)*.

### Evaluation indicators

To evaluate the effectiveness of the improved YOLOv11n network architecture, this study selected precision (P), recall (R), average precision (AP), mean average precision (mAP), model parameter count (Parameters), model floating-point operations per second (FLOPs), and frames per second (FPS) as metrics to assess the model's recognition performance. Precision represents the proportion of true positive instances among all instances classified as positive by the model, while recall reflects the percentage of actual positive instances correctly identified by the model relative to the total number of actual positives. Mean Average Precision (mAP), one of the most critical evaluation metrics for multi-class detection tasks, provides a unified performance measure across all categories, comprehensively reflecting the model's detection capabilities across different classes. Here, mAP@0.5 denotes the average precision at an intersection-over-union ratio of 0.5 Model parameters refer to the number of weights and biases requiring learning within the model. Model floating-point operations describe the model's computational efficiency, representing the number of floating-point operations executed per second. FPS reflects the model's inference speed. The calculation formula is as follows:

$$P = \frac{T_P}{T_P + F_P} \times 100\% \quad (2)$$

$$R = \frac{T_P}{T_P + F_N} \times 100\% \quad (3)$$

$$AP = \int_0^1 P(R)dR \quad (4)$$

$$mAP = \frac{\sum_{i=1}^{m} AP(n)}{3} \quad (5)$$

In the equation, TP denotes true positives (the number of correctly predicted positive samples), FP signifies false positives (the number of incorrectly predicted positive samples), and FN represents false negatives (the number of missed positive samples). n=3 corresponds to the three classification levels of tomato maturity.

### Ablation Test

To validate the rationality and effectiveness of the three improved modules based on the YOLOv11n model and their final combination, ablation tests were conducted on different combinations of these improvements under consistent testing environments and parameters to ensure fairness. The test results are shown in Table 1.

**Table 1**

**Dissolution test results**

| Model | SPDConv | C3K2_DLKA | SDI | Precision (P) / % | Recall (R) / % | MAP@0.5 / % | Parameter quantity / M | FLOPs / G | FPS |
|---|---|---|---|---|---|---|---|---|---|
| YOLOv11n | × | × | × | 91.7 | 90.2 | 93.4 | 2.58 | 8.3 | 51.21 |
| A | √ | × | × | 93.9 | 89.6 | 94.1 | 3.08 | 8.1 | 61.24 |
| B | × | √ | × | 94.7 | 92.2 | 94.7 | 3.35 | 8.9 | 60.80 |
| C | × | × | √ | 94.5 | 88.1 | 94.5 | 2.63 | 7.7 | 53.86 |

| Model | SPDConv | C3K2_DLKA | SDI | Precision (P) / % | Recall (R) / % | MAP@0.5 / % | Parameter quantity / M | FLOPs / G | FPS |
|---|---|---|---|---|---|---|---|---|---|
| D | √ | √ | × | 94.3 | 91.3 | 95.3 | 4.31 | 8.8 | 65.36 |
| E | √ | × | √ | 93.3 | 88.5 | 94.6 | 3.11 | 8.3 | 61.24 |
| F | × | √ | √ | 94.5 | 88.9 | 94.9 | 3.15 | 8.5 | 58.07 |
| YOLOv11n-SDS | √ | √ | × | 94.9 | 91.5 | 95.8 | 3.28 | 9.1 | 71.12 |

As summarized in Table 1, the ablation study was conducted to evaluate the contribution of each proposed module. In Experiment A, the SPD-Conv module was integrated into the YOLOv11n backbone, resulting in a 0.7% increase in mAP@0.5 and a 2.2% gain in precision, though recall decreased by 0.6%.

This suggests that the module strengthens the model's ability to detect low-resolution images and small tomato targets, though the lack of contextual optimization limits overall performance gains. Experiment B replaced the original attention mechanism with the C3K2_DLKA module, leading to a 1.3% improvement in mAP@0.5 and a 3.0% rise in precision, indicating that the deformable large-kernel sampling strategy improves adaptability to occluded and deformed tomatoes, thereby enhancing localization accuracy. Experiment C introduced the SDI module, which brought a 1.1% increase in mAP@0.5 and a 2.8% gain in precision, although recall decreased by 2.1%, reflecting a trade-off between context-aware feature refinement and detection coverage.

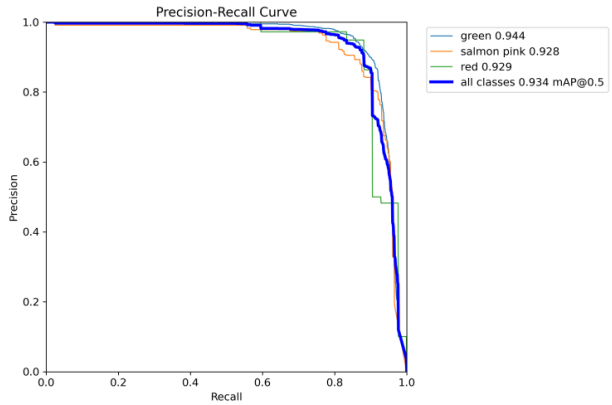



Fig. 7 - PR Curve of YOLOv11n

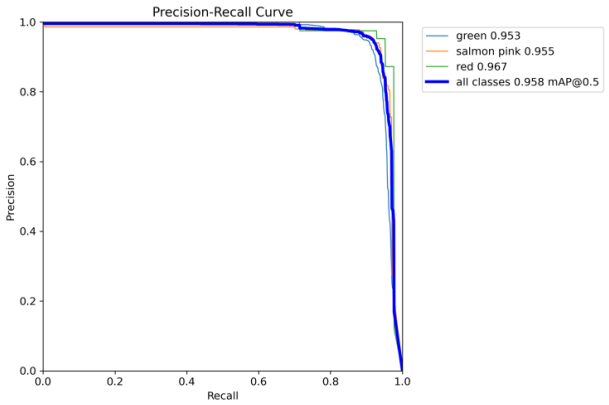



Fig. 8 - PR Curve of YOLOv11n-SDS

In Experiment D, the combined use of SPD-Conv and C3K2_DLKA modules yielded a 1.9% improvement in mAP@0.5, demonstrating their functional complementarity, though contextual modeling remained suboptimal. Experiment E, which combined SPD-Conv and SDI, achieved a 1.2% gain in mAP@0.5, indicating that the SDI module further enhances the multi-scale feature representation from SPD-Conv, particularly for small targets in complex backgrounds. Experiment F, integrating C3K2_DLKA and SDI, led to a 1.5% increase in mAP@0.5, illustrating that semantic enhancement from SDI synergizes with geometric deformation modeling from C3K2_DLKA, significantly improving robustness under occlusion.

The full model incorporating all three modules achieved a 2.4% improvement in mAP@0.5, a 3.2% gain in precision, a 1.3% increase in recall, and a 20.09 FPS improvement over the baseline YOLOv11n. This configuration effectively balances detection performance and lightweight efficiency, making it suitable for real-world agricultural applications. The precision-recall (PR) curves of YOLOv11n-SDS and the baseline, shown in Figures 7 and 8, confirm the consistent performance improvement across different confidence thresholds, validating the effectiveness of the proposed approach in tomato maturity detection.

## COMPARATIVE TEST

To evaluate the performance of the proposed YOLOv11n-SDS model in detecting tomato fruit ripeness, a comparative analysis was conducted between the improved model and mainstream object detection algorithms under consistent experimental conditions. The experimental results are presented in Table 2.

Table 2

Comparison results of different models

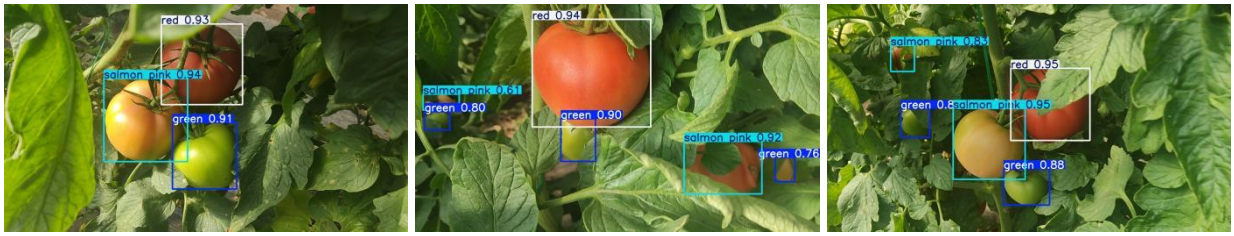
| Model | Precision/% | Recall/% | MAP@0.5/% | Parameters/M | FLOPs/G |
|---|---|---|---|---|---|
| RT-DETR | 87.8 | 83.1 | 86.6 | 18.95 | 54.7 |
| Faster-RCNN | 79.2 | 77.6 | 88.7 | 89.31 | 240.2 |
| SSD | 87.8 | 87.1 | 90.4 | 25.64 | 65.3 |
| YOLOv5n | 89.2 | 83.5 | 90.2 | 2.89 | 8.8 |
| YOLOv6n | 85.0 | 85.0 | 81.7 | 4.16 | 13.5 |
| YOLOv8n | 90.3 | 81.3 | 89.9 | 2.98 | 8.9 |
| YOLOv10n | 89.3 | 85.7 | 92.9 | 3.45 | 10.1 |
| YOLOv11n | 91.7 | 90.2 | 93.4 | 2.58 | 8.3 |
| YOLOv11n-SDS | 94.9 | 91.5 | 95.8 | 3.28 | 9.1 |

Experimental results demonstrate that the improved YOLOv11n-SDS model attains a precision of 94.9%, a recall of 91.5%, and a mAP@0.5 of 95.8%. It exceeds the mAP@0.5 performance of competing models—RT-DETR, Faster-RCNN, SSD, YOLOv5n, YOLOv6n, YOLOv8n, YOLOv10n, and YOLOv11n—by margins of 9.2%, 7.1%, 5.4%, 5.6%, 14.1%, 5.9%, 2.9%, and 2.4%, respectively, while also leading in both precision and recall. The model maintains high efficiency, with a parameter count of 3.28 M and computational cost of 9.1 GFLOPs.

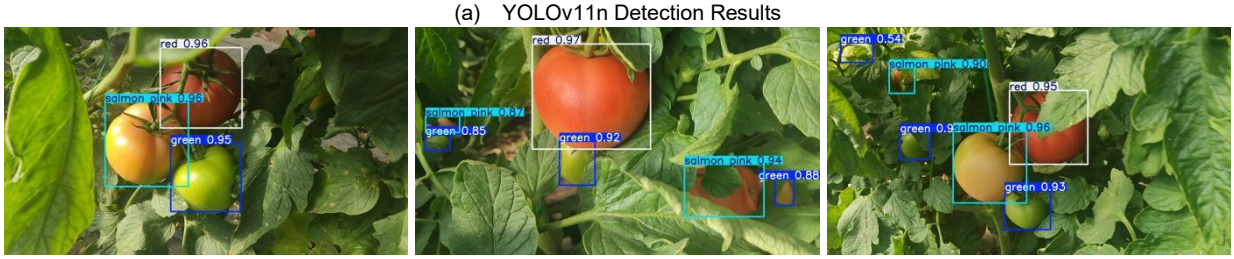
A comparative analysis with the lightweight YOLOv5n shows that a minimal increase of 0.39 M parameters and 0.3 GFLOPs yields a substantial 5.4% gain in mAP@0.5, underscoring the model's optimal balance of accuracy, lightweight design, and computational efficiency for tomato ripeness detection.

**Evaluation of Testing Results**

Figure 9 presents a comparative analysis of tomato maturity detection performance between the baseline YOLOv11n and the improved YOLOv11n-SDS model. The results indicate a marked enhancement in detection capability following the proposed improvements. Specifically, the original YOLOv11n model failed to detect low-resolution small-sized tomatoes under strong backlighting in distant scenes, whereas YOLOv11n-SDS successfully identified all targets. This improvement is attributed to the multi-scale feature fusion capability of the SPD-Conv module and the enhanced contextual understanding provided by the SDI module, which collectively improve small target detection and robustness against complex background interference. Furthermore, in cases of tomatoes heavily obscured by leaves, the original model exhibited low confidence scores and imprecise bounding box localization. In contrast, YOLOv11n-SDS, leveraging the robust geometric deformation modeling capabilities of the C3K2_DLKA module, achieved more accurate localization and higher confidence levels, thereby increasing the detection rate of occluded tomatoes. The visualization results confirm that the proposed enhancements effectively address practical challenges such as small targets, complex backgrounds, and occlusions, validating the efficacy of the model optimization approach.




(a) YOLOv11n Detection Results




(b) YOLOv11n-SDS Detection Results

**Fig. 9 - Tomato maturity detection results**

**Test verification**

To further verify the practical applicability of the proposed YOLO11n-SDS model, it was deployed on a mobile robotic perception platform. As shown in Figure 10, the platform consisted of a TAObotics Turtle intelligent mobile chassis, a robotic manipulator, an onboard computing unit, and an Intel RealSense D435i RGB-D camera. During operation, the RealSense D435i captured both RGB images and depth information of the tomato plants. The RGB images were used as the input of the proposed model for tomato detection and maturity classification, while the depth data were used to estimate the spatial position of the detected targets, enabling simultaneous maturity recognition and 3D localization. Due to the limitations of the available experimental conditions, the practical deployment test was conducted in a laboratory environment.

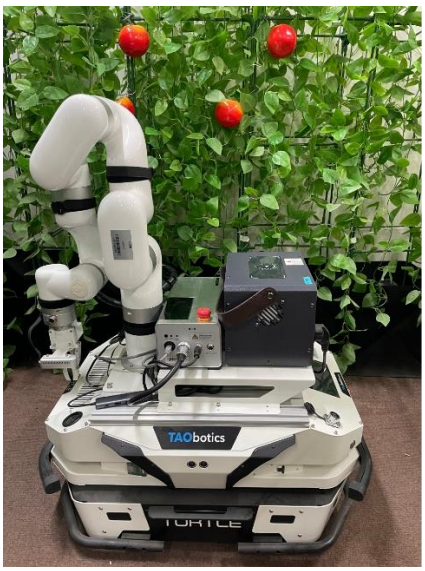



**Fig. 10 - Robotic deployment platform**

As shown in Figure 11, under greenhouse conditions with leaf occlusion and complex background interference, the deployed system was able to perform online detection of tomatoes at different maturity stages and output the target category, confidence score, and 3D coordinates in real time. In the representative example, the average confidence score was 0.93 for red tomatoes and 0.925 for salmon-pink tomatoes. These results indicate that the proposed model is not only effective in offline evaluation, but can also operate stably on a robotic platform, providing useful perception support for tomato ripeness monitoring and harvesting-oriented target localization, and demonstrating good potential for practical application.

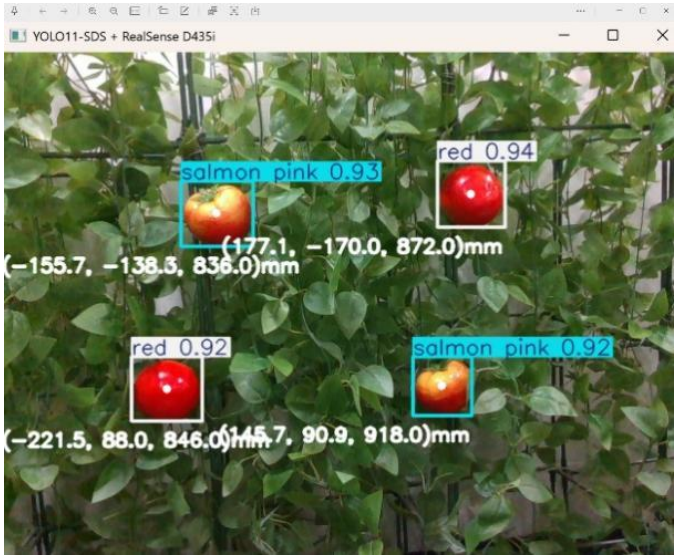



**Fig. 11 - Online detection and localization results**

## CONCLUSIONS

This study addresses the practical needs of tomato ripeness detection in agricultural environments by introducing three key improvements to the YOLOv11n framework: the Spatial Pyramid Depthwise Separable Convolution (SPD-Conv) module, the Deformable Large-Kernel Attention (DLKA) mechanism within the C3K2 module, and the Semantic and Detail Injection (SDI) module.

The SPD-Conv module enhances the detection of low-resolution images and small-sized tomatoes through multi-scale feature fusion. Experiments show that adding this module alone increases mAP@0.5 by 0.5%, demonstrating its particular usefulness in orchard settings where image quality is degraded by leaf occlusion.

The C3K2_DLKA module uses a deformable large-kernel sampling strategy to significantly improve the model's adaptability to geometric variations in tomatoes, thereby raising detection accuracy under occlusion. The SDI module reduces the interference from complex lighting and background clutter by optimizing cross-level feature interactions.

A notable synergistic effect is observed among the three modules. The multi-scale features from SPD-Conv provide a solid foundation for the deformation modeling in C3K2_DLKA, while the contextual feature refinement of SDI further enhances the performance of the preceding modules. After integration, YOLOv11n-SDS achieves a mAP@0.5 of 95.8% (a 2.4% gain over YOLOv11n), a recall of 91.5% (+1.3%), and a precision of 94.9% (+3.2%), while maintaining high inference speed. Despite these enhancements, the model contains only 3.28 M parameters, an increase of 1.3 M compared to the baseline, making it suitable for deployment on resource-constrained agricultural devices.

Furthermore, the proposed YOLOv11n-SDS model was deployed and validated on a mobile robotic platform in a greenhouse environment. The deployment results showed that the model could perform online tomato maturity detection and output 3D target coordinates in real time under conditions with foliage occlusion and complex background interference. This demonstrates that the proposed method is not limited to offline image analysis, but can also provide effective perception support for tomato ripeness monitoring and harvesting-oriented target localization. Future work will further validate the proposed method in more diverse and large-scale practical greenhouse production scenarios to improve its robustness and generalization ability.

**ACKNOWLEDGEMENT**

This research was supported by the Inner Mongolia Autonomous Region Science and Technology Innovation Guidance Project (Kcj1-202205).

**REFERENCES**

[1] Appe, S.N., Arulselvi, G., Balaji, & G.N. (2023). CAM-YOLO: tomato detection and classification based on improved YOLOv5 using combining attention mechanism. *PeerJ Computer Science, 9*, e1463. https://doi.org/10.7717/peerj-cs.1463

[2] Chen, B. J., Bu, J. Y., Xia, J. L., & Li, M. X. (2025). AFBF-YOLO: An Improved YOLO11n Algorithm for Detecting Bunch and Maturity of Cherry Tomatoes in Greenhouse Environments. *Plants, 14*(16), 2587. https://doi.org/10.3390/plants14162587

[3] Collins, E. J., Bowyer, C., Tsouza, A., & Chopra, M. (2022). Tomatoes: An extensive review of the associated health impacts of tomatoes and factors that can affect their cultivation. *Biology, 11*(2), 239. https://doi.org/10.3390/biology11020239

[4] Gan, X., Chen, C., & Zhang, S. (2025). QMDF-YOLO11: Rice pests detection algorithm in complex scenarios. *Computer Engineering and Applications*, *61*(13), 113-123. https://doi.org/10.3778/j.issn.1002-8331.2501-0120

[5] Girshick, R. (2015). Fast R-CNN. *Proceedings of the IEEE international conference on computer vision* (pp. 1440-1448). https://doi.org/10.1109/ICCV.2015.169

[6] Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 580-587). https://doi.org/10.1109/CVPR.2014.81

[7] Gonzali, S., & Perata, P. (2021). Fruit colour and novel mechanisms of genetic regulation of pigment production in tomato fruits. *Horticulturae*, *7*(8), 259. https://doi.org/10.3390/horticulturae7080259

[8] Hua, X., Li, H., Zeng, J., Han, C., Tang, L., & Luo, Y. (2023). A review of target recognition technology for fruit picking robots: From digital image processing to deep learning. *Applied Sciences*, *13*(7), 4160. https://doi.org/10.3390/app13074160

[9] Ji, W., Song, T., Rong, X., & Zhou, B. (2025). PATD-YOLO: Road obstacle object detection algorithm based on YOLOv11. *Computer Engineering and Applications*. (Preprint/Early Access). https://doi.org/10.3778/j.issn.1002-8331.2504-0038

[10] Li, J., Huang, Z., Xia, L., Sun, H., & Wang, H. (2025). Tomato maturity detection based on improved YOLOv8n. *INMATEH-Agricultural Engineering*, *75*(1). https://doi.org/10.35633/inmateh-75-53

[11] Liu, K., Yu, J., Huang, Z., Liu, L., & Shi, Y. (2024). Autonomous navigation system for greenhouse tomato picking robots based on laser SLAM. *Alexandria Engineering Journal*, *100*, 208-219. https://doi.org/10.1016/j.aej.2024.05.032

[12] Meng, K., Si, J., Yang, T., & Tian, H. (2025). Facial expression recognition based on YOLOv9 improvement. *International Conference on Computer Vision and Image Processing (CVIP 2024)*. SPIE, *13521*, 247-253. https://doi.org/10.1117/12.3058029

[13] Miao, R., Li, Z., & Wu, J. (2023). Lightweight maturity detection of cherry tomato based on improved YOLOv7. *Transactions of the Chinese Society of Agricultural Machinery*, *54*(10), 225-233. https://doi.org/10.6041/j.issn.1000-1298.2023.10.022

[14] Peng, Y., Chen, D. Z., & Sonka, M. (2025). U-net v2: Rethinking the skip connections of U-net for medical image segmentation. *2025 IEEE 22nd International Symposium on Biomedical Imaging (ISBI)* (pp. 1-5). IEEE. https://doi.org/10.1109/isbi60581.2025.10980742

[15] Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster R-CNN: Towards real-time object detection with region proposal networks. *Advances in Neural Information Processing Systems*, *28*. https://doi.org/10.1109/tpami.2016.2577031

[16] Sapkota, R., Flores-Calero, M., Qureshi, R., Badgujar, C., & Nepal, U. (2025). YOLO advances to its genesis: A decadal and comprehensive review of the You Only Look Once (YOLO) series. *Artificial Intelligence Review*, *58*(9), 274. https://doi.org/10.1007/s10462-025-11253-3

[17] Seo, D., Cho, B. H., & Kim, K. C. (2021). Development of monitoring robot system for tomato fruits in hydroponic greenhouses. *Agronomy*, *11*(11), 2211. https://doi.org/10.3390/agronomy11112211

[18] Sunkara, R., & Luo, T. (2022). No more strided convolutions or pooling: A new CNN building block for low-resolution images and small objects. *Joint European Conference on Machine Learning and Knowledge Discovery in Databases* (pp. 443-459). Cham: Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-26409-2_27

[19] Wang, C., Yang, G., Huang, Y., Liu, Y., & Zhang, Y. (2023). A transformer-based mask R-CNN for tomato detection and segmentation. *Journal of Intelligent & Fuzzy Systems*, *44*(5), 8585-8595. https://doi.org/10.3233/jifs-222954

[20] Wang, S., Xiang, J., Chen, D., & Zhang, C. (2024). A method for detecting tomato maturity based on deep learning. *Applied Sciences*, *14*(23), 11111. https://doi.org/10.3390/app142311111

[21] Wang, Z., Ling, Y., Wang, X., Meng, D., Nie, L., An, G., & Wang, X. (2022). An improved Faster R-CNN model for multi-object tomato maturity detection in complex scenarios. *Ecological Informatics*, *72*, 101886. https://doi.org/10.1016/j.ecoinf.2022.101886

[22] Wu, J., Zhou, L., Zeng, J., Shan, Q., Wang, R., & Wang, R. (2025). Object detection and recognition method of ships based on IRS-YOLO. *Journal of Optoelectronics·Laser*, *36*(08), 848-856. https://doi.org/10.16136/j.joel.2025.08.0188

[23] 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*(9), 1006. https://doi.org/10.3390/horticulturae10091006

[24] Zhang, Z., Li, X., & Chen, S. (2025). Small object detection algorithm in UAV aerial images based on improved YOLO11. *Chinese Journal of Liquid Crystals and Displays*, *40*(06), 915-930. https://doi.org/10.3969/j.issn.1003-3106.2025.08.002

[25] Zhao, L., Qian, R., Liu, C., Wang, S., & Xia, J. (2025). Wheat impurity detection algorithm based on improved YOLO v8. *INMATEH-Agricultural Engineering*, *75*(1). https://doi.org/10.35633/inmateh-75-62