RESEARCH ON A TOMATO RIPENESS DETECTION METHOD BASED ON CMLE-YOLO

/

基于 CMLE-YOLO 的番茄成熟度检测方法研究

Shuo LIU, Pengzhi HOU, Linqiang DENG, Lijun CHENG*, Jia LV*
Faculty of Software Technologies, Shanxi Agricultural University, Taigu 030801, China;
Tel: +86-13835441585; E-mail: cljzyb@sxau.edu.cn

DOI: https://doi.org/10.35633/inmateh-78-62

Keywords: *Tomato ripeness detection; YOLOv11; lightweight object detection; computer vision; deep learning in agriculture; smart agriculture; fruit maturity classification*

ABSTRACT

Improper determination of tomato harvest maturity often leads to uneven ripening, overripening, decay, and softening damage during transportation, resulting in substantial postharvest losses. To provide an objective basis for tomato maturity classification and fruit counting, this study developed an improved lightweight model for tomato maturity detection and counting, named CMLE-YOLO. Built upon YOLOv11, the proposed model incorporates a Cross-Fusion and Multi-scale Attention (CFMA) module into the backbone and neck to enhance spatial feature interaction and global context modeling. In addition, a Lightweight Quality-Aware Detection Head (LQAD) head was designed to improve the consistency between classification confidence and localization accuracy while reducing parameter redundancy. A dataset containing 2,000 images and 8,593 annotated tomato instances was constructed for model training and evaluation. Experimental results showed that CMLE-YOLO achieved strong performance in detecting three tomato maturity stages, namely green, half_ripened, and fully_ripened, with a mAP@50 of 0.8508, outperforming several mainstream detectors, including YOLOv5, YOLOv6, and YOLOv8. The model also remained lightweight, with only 2.13 M parameters and 5.2 GFLOPs, indicating lower computational complexity than most comparative models. Overall, CMLE-YOLO achieved a favorable balance between detection accuracy and efficiency, providing technical support for real-time harvesting management, automated grading, and yield estimation in tomato production systems.

摘要

番茄采收成熟度的错误判定常导致运输过程中出现不均匀成熟、过熟、腐烂及软化损伤，造成重大采后损失为番茄成熟度分类与果实计数提供客观依据，本研究开发了改进型轻量级番茄成熟度检测与计数模型 CMLE-YOLO。该模型基于 YOLOv11 构建，在主干网络与颈部区域融入跨阶段快速混合注意力（CFMA）模块，以增强空间特征交互与全局上下文建模能力。此外，设计了轻量级质量感知检测（LQAD）头部，在降低参数冗余的同时，提升了分类置信度与定位精度的协调性。构建包含 2000 张图像及 8593 个标注番茄实例的数据集用于模型训练与评估。实验结果表明，CMLE-YOLO 在检测三个番茄成熟阶段（未成熟、半成熟、成熟）时表现优异，mAP@50 达 0.8508，超越 YOLOv5、YOLOv6、YOLOv8 等主流检测器。该模型同时保持轻量化特性，仅需 213 万参数和 5.2 GFLOPs 计算量，计算复杂度低于多数对比模型。总体而言，CMLE-YOLO 在检测精度与运行效率间实现了良好平衡，为实时采收管理、自动化分级等应用提供了技术支持。

INTRODUCTION

Tomatoes are among the most important protected horticultural crops worldwide, and post-harvest losses remain a major factor affecting industry profitability (*Food and Agriculture Organization of the United Nations, 2019*). Because tomatoes are highly perishable, harvest ripeness directly influences storage performance, transportability, and arrival quality. In cross-regional distribution, harvesting too early may lead to uneven ripening and unstable market quality (*Al-Dairi et al., 2021*), whereas harvesting too late increases the risk of softening, bruising, and decay during transport. Therefore, accurate ripeness identification is important for post-harvest grading, transport allocation, and loss reduction. In this study, tomato ripeness was grouped into three operationally meaningful stages—unripe, half-ripened, and fully ripened—because this categorization is more consistent with post-harvest handling and transport decision-making than finer commercial subclasses.

However, automated tomato ripeness detection in real production environments remains challenging. Tomatoes are typically small, densely distributed, and frequently occluded by leaves, stems, and neighboring fruits. In greenhouse scenes, uneven illumination, background clutter, and specular reflections further increase detection difficulty (*Li et al., 2023; Li et al., 2025*). In addition, adjacent ripeness stages often exhibit subtle visual differences, making ripeness analysis a fine-grained detection task rather than a conventional fruit localization problem (*Ma et al., 2025; Wu et al., 2024*). These characteristics require a model that can simultaneously achieve accurate classification, reliable counting, and efficient deployment.

Recent advances in deep learning have significantly promoted the application of computer vision in agricultural perception tasks, especially in fruit detection and classification (Koirala et al., 2019). Object detection frameworks represented by the YOLO series have shown advantages in inference efficiency and have been widely studied in agricultural scenarios (*Badgujar et al., 2024*). At the same time, ripeness grading in post-harvest management is usually based on surface color changes, and industry grading standards have provided a practical basis for stage definition and handling decisions (*Al-Dairi et al., 2021; Huang et al., 2022*). Nevertheless, existing studies still face several limitations when applied to tomato ripeness analysis under real agricultural conditions. First, adjacent ripeness stages are still easily confused under occlusion and illumination variation (*Li et al., 2023*). Second, most studies emphasize detection accuracy while providing limited quantitative evaluation of counting reliability (*Fu et al., 2024; Zhao et al., 2024*). Third, the balance between detection performance and lightweight deployment remains insufficient for practical harvesting and grading scenarios (*Wu et al., 2024; Ma et al., 2025*). These limitations restrict the direct application of existing methods in post-harvest management and transport-oriented decision support.

To address these issues, this study proposes CMLE-YOLO, a lightweight tomato ripeness detection model for complex agricultural environments. Considering the practical requirements of post-harvest grading and transport allocation, tomato maturity was defined using three operational categories: unripe, half-ripened, and fully ripened. The main contributions of this study are as follows. (1) A three-stage ripeness detection framework was established to better match post-harvest grading and transport decision needs. (2) A lightweight detection model, CMLE-YOLO, was developed to improve ripeness discrimination under dense distribution, occlusion, and complex illumination conditions. (3) An integrated evaluation strategy covering both ripeness detection and instance-level counting was introduced to support practical quantity estimation and deployment in agricultural scenarios.

MATERIALS AND METHODS

Data Collection

The tomato image data for this study were captured using a Canon EOS 70D digital SLR camera equipped with an EF-S 18–135 mm f/3.5–5.6 IS STM lens. The native output resolution is 5472×3648 pixels, and the images were taken in the standard tomato cultivation area at the experimental base of Shanxi Agricultural University. The acquisition process was designed to reflect typical agricultural vision conditions encountered in practical fruit detection tasks (*Koirala et al., 2019; Badgujar et al., 2024*).

All images were captured under natural daylight conditions, spanning both morning and evening periods. Although light intensity was not quantitatively measured, the natural random distribution of collection times objectively covered diverse daylight conditions—including soft light and strong direct sunlight—enabling the model to learn tomato appearance variation features under varying illumination. It should be noted that this study did not collect nighttime scene images or utilize artificial lighting equipment. The research scope is limited to the task of detecting tomato ripeness under natural daylight conditions.

A total of 2,100 raw images were collected. After quality screening (excluding severely blurred images and samples with missing targets) and annotation completeness review, 2,000 images were ultimately retained for subsequent experiments. Based on this, tomatoes in the images were annotated into three maturity categories: “fully_ripened”, “half_ripened”, and “green” forming a dataset comprising 8,593 annotated instances.

In the data annotation process, LabelImg tool was used to manually annotate the tomato targets in the images, categorized into "fully_ripened", "half_ripened", and "green" based on their maturity stages. All annotation results were saved in YOLO format for subsequent model training and experimental analysis. After annotation, the dataset was divided into training, validation, and test sets in an 8:1:1 ratio to ensure the validity of model training, parameter tuning, and performance evaluation. Meanwhile, to enhance the model's adaptability to complex environments and improve data diversity, online data augmentation techniques were applied during training, including random flipping of images, hue, saturation, and brightness augmentation, as well as mosaic augmentation, further improving the model's robustness and generalization ability.

To visually illustrate the distribution characteristics of the annotated targets, Fig. 1 displays the distribution patterns of the center coordinates (x, y), width, and height of the annotated bounding boxes. The center coordinates exhibit a relatively clustered distribution, reflecting the dense growth patterns of tomatoes within the images, while both width and height values skew toward smaller ranges and exhibit a clear positive correlation. Combined with category-specific scale statistics (fully_ripened: average bounding box area 0.0193, half_ripened: 0.0179, green 0.0123), this visually confirms the high proportion of small objects in the dataset and the prevalence of small objects in the green category. It also provides data-driven insights for subsequent model improvements targeting small objects and dense scenes.

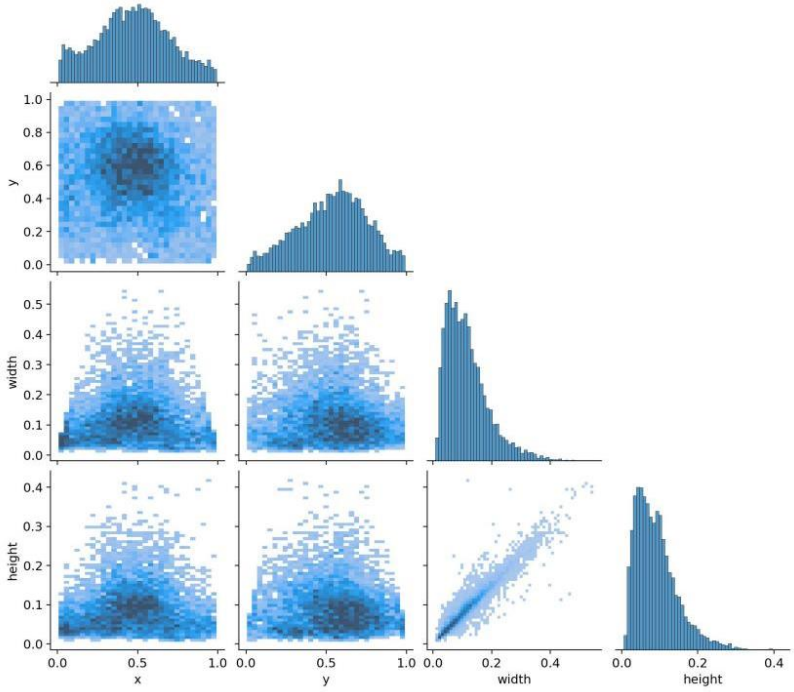

Fig. 1 - Heatmap and Histogram of Target Box Parameter Distribution for Tomato Ripeness Dataset

Model Structure

Although the YOLO series demonstrates high accuracy and inference efficiency in general object detection, its classic “backbone-neck-detection head” architecture still faces structural limitations in complex field scenarios. Particularly under conditions of multi-scale dense objects, strong occlusions, and lighting perturbations, there remains room for improvement in the consistency between feature representation and prediction quality (*Xu et al., 2024)*.

This study focuses on ordinary tomato fruits, whose natural harvesting environment exhibits typical characteristics such as subtle and continuously graded maturity differences, dense fruit distribution, frequent foliage occlusions, and complex background textures. These factors cause the original YOLOv11 to frequently encounter issues like boundary confusion between adjacent stages, localization shifts, and missed detections when distinguishing fine-grained maturity levels and locating densely packed small objects.

From a feature modeling perspective, YOLOv11's bottleneck-stacked architecture (e.g., C3k2) in its backbone and neck layers prioritizes local texture extraction, resulting in relatively insufficient modeling of cross-region dependencies and global context. Under complex backgrounds and occlusion conditions, the network is susceptible to interference from distracting information such as leaf textures, reflections, and shadows, thereby weakening its stable representation of key maturity cues (subtle color variations, local textures, and morphological edges). To enhance selective attention to critical regions and contextual integration, existing research indicates that channel/spatial attention mechanisms and global relationship modeling (e.g., Non-local) can effectively improve feature discriminative power and robustness (*Hu et al., 2018; Wang et al., 2018*). Simultaneously, multi-scale object detection typically relies on top-down and bottom-up feature fusion architectures (e.g., FPN, PANet, BiFPN) to mitigate semantic inconsistencies caused by scale variations. However, under conditions of dense small objects and occlusions, stronger cross-scale interactions and complementary information mechanisms are still required to improve the separability and localizability of intermediate-level features (*Lin et al., 2017; Liu et al., 2018; Tan et al., 2020*).

From the detection head perspective, the multi-scale branches in YOLOv11 are structurally independent, leading to parameter redundancy and computational overhead. Additionally, a potential "quality mismatch" exists between classification confidence and localization quality—classification scores do not reliably reflect localization accuracy, thereby affecting NMS ranking and retention strategies. This increases the risk of false suppression and missed detections in dense target scenarios (*Feng et al., 2021*). Furthermore, detection tasks often employ small batch sizes due to GPU memory constraints. BatchNorm's statistical instability under small batches further compromises training convergence and generalization robustness. Alternative normalization methods like Group Normalization typically exhibit greater stability in small batch scenarios (*Ioffe and Szegedy, 2015)*.

To address the key challenges in tomato ripeness detection, this paper proposes the improved CMLE-YOLO model based on YOLOv11: First, the Cross-Fusion and Multi-scale Attention (CFMA) module replaces the original C3k2 in both the backbone and neck networks. By enhancing cross-scale feature interaction and contextual fusion capabilities, the network more effectively focuses on fruit regions while suppressing complex background interference. This improves its ability to represent subtle color and texture differences across various ripening stages. Second, a lightweight quality-aware detection head, Lightweight Quality-Aware Detection Head (LQAD), is designed to reduce multi-scale branch redundancy and strengthen consistency between classification and localization predictions. This enhances detection and counting stability and reliability under dense occlusion conditions. The overall model architecture is shown in Fig. 2.

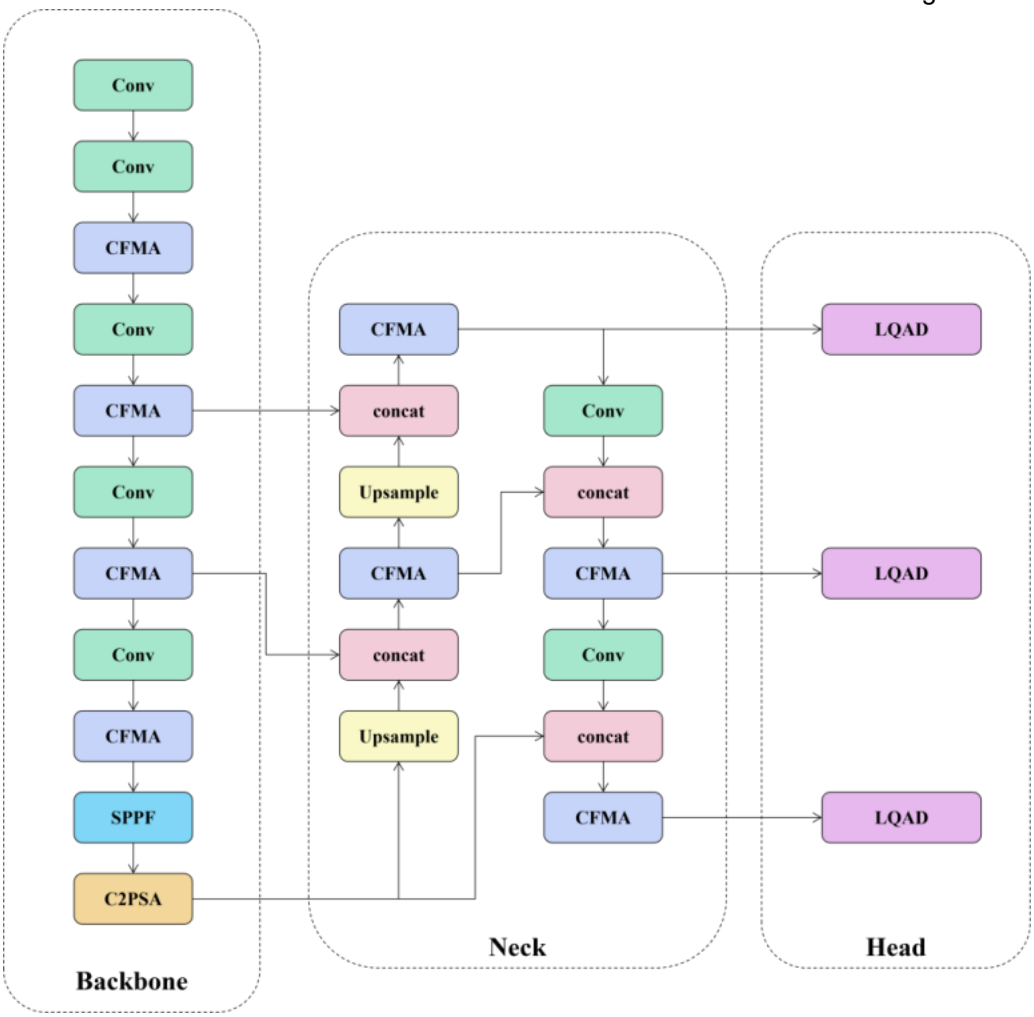

Fig. 2 - CMLE-YOLO Network Architecture

CFMA（Cross-Fusion and Multi-scale Attention）

This study replaces the bottleneck module in the C3k2 architecture of YOLOv11 with an improved FEM-Block, constructing a novel CFMA module. Traditional bottleneck modules primarily rely on standard convolutions for feature extraction, resulting in limited expressiveness. The CFMA module innovatively introduces an efficient local convolution-spatial mixing mechanism, combined with an MLP-based channel mixing structure, achieving deeper feature interaction and reconstruction across both spatial and channel dimensions.

Specifically, the FEM-Block first employs PConv (Partial Convolution) to perform spatial convolutions only on selected channels, reducing computational load while preserving the convolutional receptive field (PConv structure shown in Fig. 3). Subsequently, an MLP composed of 1×1 convolutions expands and compresses the channel dimension, achieving stronger feature representation capabilities. Additionally, this introduces an EMA attention mechanism in the residual branch (as shown in Fig. 3) to globally enhance MLP-output features, promoting multi-scale information fusion and strengthening expression capabilities in critical regions. Finally, the combination of residual connections and the DropPath random depth strategy improves training stability and model generalization. The overall FEM-Block architecture is illustrated in Fig. 4.

Overall, the CFMA module retains the lightweight structural advantages of C3k2 while introducing more efficient spatial mixing and attention mechanisms. This significantly enhances the network's modeling capabilities for complex backgrounds, fine-grained objects, and scale variations. By providing more discriminative and expressive intermediate feature representations, the CFMA module strengthens the feature extraction capabilities of both the backbone and neck networks, delivering richer and more stable feature foundations for subsequent layers.

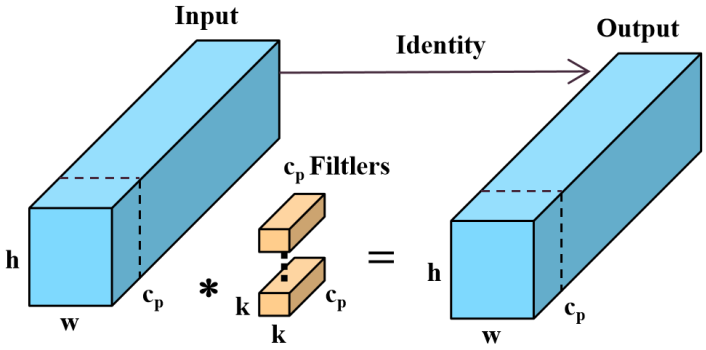

Fig. 3 - Pconv Block Diagram

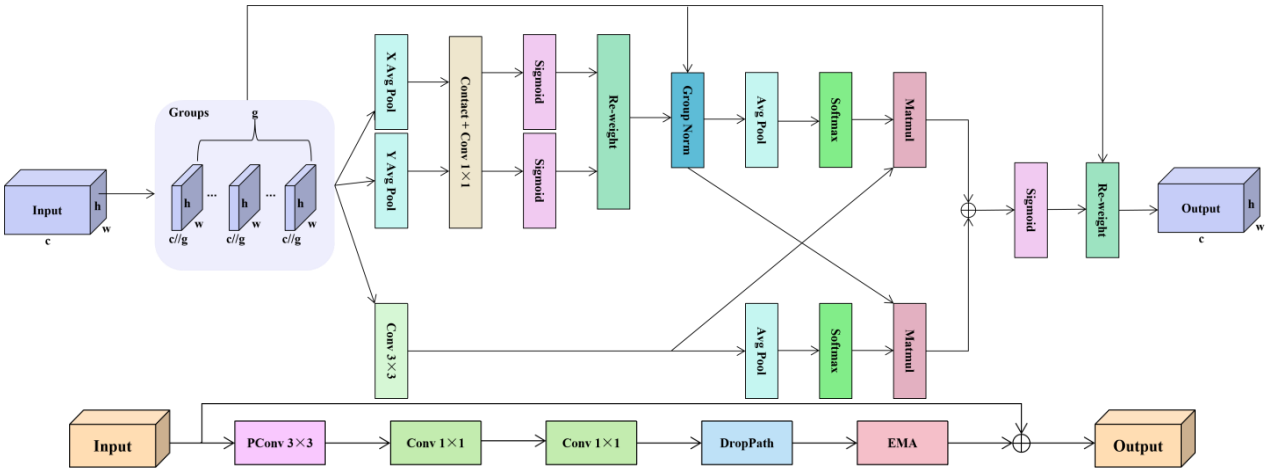

Fig. 4 - Structure of FEM-Block and EMA Attention Mechanism

LQAD (Low-Quality Image Analysis and Detection)

To enhance the detection accuracy and computational efficiency of YOLOv11 on multi-scale objects, this study reconstructs its original detection head, designing a lightweight quality-aware detection head (LQAD). This detection head primarily consists of three components: a standard convolution module based on GroupNorm (Conv_GN), a lightweight shared convolution structure (LSCS), and a localization quality estimator (LQE). These components synergistically reduce computational overhead while enhancing classification score reliability and bounding box prediction accuracy. The overall architecture is illustrated in Fig. 5.

In traditional YOLO detection heads, convolutional units typically rely on BatchNorm for feature normalization. However, BatchNorm exhibits high sensitivity to batch size, leading to statistical fluctuations during high-resolution training or small batches, which compromises model convergence stability. To mitigate this issue, this study introduces Conv_GN as the foundational building block for detection heads. This module comprises a convolutional layer, GroupNorm, and the SiLU activation function.

GroupNorm operates independently of batch statistics, maintaining stable normalization effects during small-batch training while demonstrating superior generalization capabilities across data domains or scenarios. Simultaneously, SiLU's smooth nonlinearity enhances feature expression capabilities, aiding in capturing fine-grained structures and small object information.

By systematically adopting Conv_GN in the detection head, this study achieves modular unification in normalization strategies and activation mechanisms, thereby further improving the training stability and overall performance of the detection head.

The original YOLOv11 detector employs independent convolutional branches at each scale, resulting in significant structural redundancy and parameter bloat. To address this, this study proposes LSCS. First, it uses Conv_GN3×3 to uniformly map multi-scale features into a consistent channel space. Subsequently, it adopts a shared convolutional structure composed of Conv_GN3×3 and Conv_GN1×1 to perform unified feature enhancement modeling across all scales. This cross-scale shared lightweight design reduces the detector head's parameter count and computational load while preserving expressive power, further enhancing prediction consistency across resolutions. Simultaneously, the introduction of GN improves training stability and consistency in multi-scale feature distribution, resulting in a more compact and efficient overall structure. It also provides a more stable feature foundation for the subsequent LQE module. Furthermore, the structural design maintains simplicity in both regression and classification branches, employing convolutional layers as core building blocks. The regression branch incorporates a learnable scale factor (Scale) for adaptive scaling of regression outputs, enhancing consistency and numerical stability across detection layers.

The YOLO series models commonly suffer from inconsistencies between classification scores and localization quality, where high classification scores do not necessarily correspond to high localization accuracy. This can lead to the retention of prediction boxes with significant positional deviations during the NMS stage, thereby compromising overall detection performance. To address this issue, this study introduces a Localization Quality Estimator (LQE) into the detection head. It explicitly models the confidence of bounding boxes by leveraging the discrete distribution information of the regression output via Distribution Focal Loss (DFL). This approach converts the regression distribution into a probabilistic form and extracts statistical features reflecting both the concentration of the distribution and prediction uncertainty. Subsequently, a multi-layer perceptron generates a quality adjustment value, enabling quality-aware correction of classification scores. This ensures scores simultaneously reflect category confidence and localization quality. This mechanism enables the model to more effectively distinguish between high-quality and low-quality prediction boxes during the NMS phase, thereby enhancing detection accuracy.

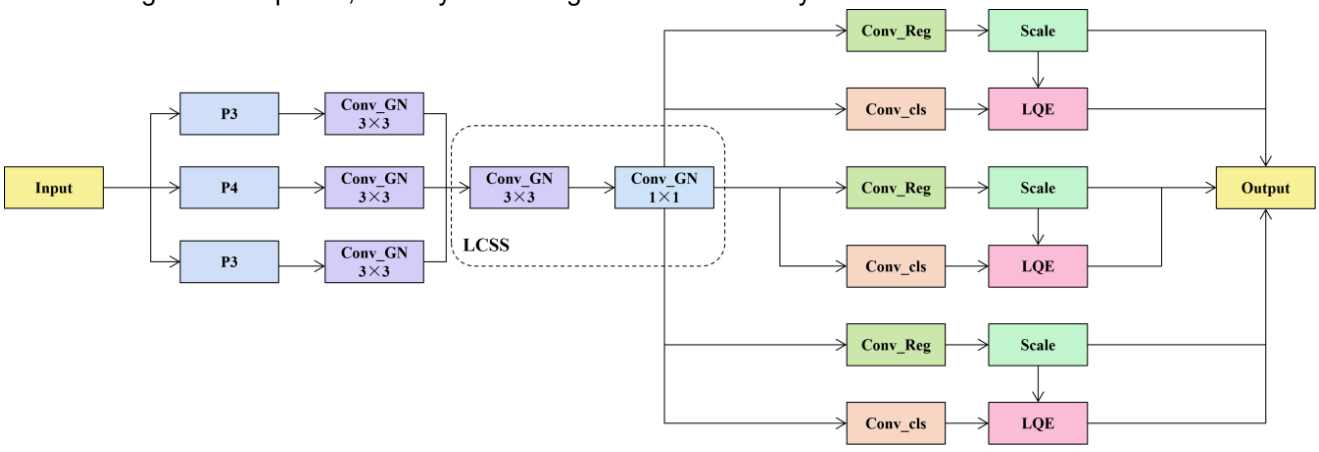

Fig. 5 - LQAD Schematic Diagram

Test Environment and Parameter Settings

The test platform and environment are shown in Table 1.

Table 1

Experimental Environment Configuration

Configuration Name	Version & Model
GPU Model & Quantity	NVIDIA GeForce RTX 4080
CPU Model	Intel64 Family 6 Model 183 Stepping 1 GenuineIntel
System RAM	65,349 MB

Configuration Name	Version & Model
Operating System	Microsoft Windows 11 Professional
Deep Learning Framework	PyTorch 2.x (Ultralytics YOLOv11)
CUDA with cuDNN version	CUDA 12.9 + cuDNN

Training Parameters

Key parameters for training CMLE-YOLO are shown in Table 2, with primary settings as follows: training epochs set to 200, batch size set to 16, learning rate set to 0.01, SGD optimizer employed, input image size set to 640, and IoU threshold set to 0.7.

The training strategy employs a learning rate warm-up mechanism. During the initial 3 epochs, the learning rate linearly increases from 10% of the base learning rate to the preset value of 0.01. This avoids gradient oscillations caused by random weight initialization in the early training stages, stabilizing the model's initial learning of fine-grained features related to tomato ripeness.

Table 2

CMLE-YOLO Key Training Parameters

Parameter	Value	Parameter	Value
Epochs	200	Patience	100
Batch size	16	AMP	False
LRF	0.01	IoU	0.7
Optimizer	SGD	lr0	0.01
Input size	640		

Evaluation Criteria

To quantitatively evaluate the model's performance in tomato ripeness detection tasks, this paper adopts Precision, Recall, and mAP@50 as the primary metrics for assessing detection accuracy.

Precision measures the proportion of samples predicted as targets that are actually targets, defined by Formula (1):

$$P = \frac{TP}{TP + FP} \times 100\% \quad (1)$$

where TP denotes the number of samples correctly detected as targets, while FP denotes the number of samples incorrectly detected as targets.

The recall rate measures the proportion of actual targets successfully detected by the model, defined by Formula (2):

$$R = \frac{TP}{TP + FN} \times 100\% \quad (2)$$

where FN denotes the number of targets not detected by the model.

mAP@0.5 is calculated by averaging the average precision (AP) across all categories at an IoU threshold of 0.5. It serves as a comprehensive metric reflecting the model's overall detection performance across categories of varying maturity levels.

RESULTS AND DISCUSSIONS

Ablation Experiment

To evaluate the contribution of each structural component in the proposed model, ablation experiments were conducted by progressively introducing the CFMA (Cross-Fusion and Multi-scale Attention) module and the LQAD (Lightweight Quality-Aware Detection Head) into the baseline YOLOv11n framework. The purpose was to quantify the effect of each module on tomato ripeness detection performance under the same experimental settings. The results of the ablation study are presented in Table 3.

As shown in Table 3, adding the CFMA module to YOLOv11n improved recall from 0.7787 to 0.7820 and increased mAP@50 from 0.8219 to 0.8320. At the same time, the number of parameters decreased from 2.58 M to 2.29 M, and GFLOPs decreased from 6.2 to 6.0.

When only the LQAD module was introduced, recall increased to 0.7860 and mAP@50 increased to 0.8410. The number of parameters became 2.42 M, while GFLOPs decreased to 5.6.

When both CFMA and LQAD were integrated, the proposed CMLE-YOLO achieved the best overall performance. Specifically, mAP@50 increased to 0.8508 and recall reached 0.7893, while the number of

parameters was reduced to 2.13 M and GFLOPs decreased to 5.2. These results indicate that the combined design provides the most favorable balance between detection accuracy and computational efficiency.

Overall, the ablation results demonstrate that both CFMA and LQAD contribute positively to tomato ripeness detection. CFMA mainly improves feature representation, while LQAD enhances prediction quality at the detection head. Their combination further improves performance and supports the effectiveness of the proposed model design.

Table 3

Ablation Experiment Results

YOLOv11n	CFMA	LQAD	P	R	mAP@50	Params(M)	GFLOPs(G)
√			0.7979	0.7787	0.8219	2.58	6.2
√	√		0.7925	0.7820	0.8320	2.29	6.0
√		√	0.7890	0.7860	0.8410	2.42	5.6
√	√	√	0.7851	0.7893	0.8508	2.13	5.2

Model Comparison

To validate the overall detection performance and deployment feasibility of the proposed CMLE-YOLO, this section conducts comparative experiments against mainstream YOLO series models (YOLOv5, YOLOv6, YOLOv8, YOLOv10, YOLOv11) under unified training and testing settings. Evaluation metrics encompass localization accuracy (mAP), classification and detection reliability (P, R, F1), inference computational complexity (GFLOPs), and inference speed (FPS) to systematically characterize the model's performance-efficiency trade-off.

Table 4 summarizes the comparison results on the validation set. Regarding core accuracy metrics, CMLE-YOLO achieves an mAP@50 of 0.8508, outperforming YOLOv5 (0.8291), YOLOv6 (0.8279), YOLOv8 (0.8172), YOLOv10 (0.8225), and YOLOv11 (0.8219). This demonstrates that the proposed structural improvements consistently enhance object detection accuracy. Under the stricter mAP50–95 metric, CMLE-YOLO achieves 0.6905, matching YOLOv6 (0.6905) and outperforming all other models (YOLOv5: 0.6808; YOLOv8: 0.6692; YOLOv10: 0.6672; YOLOv11: 0.6762), demonstrating the method's ability to maintain fine-grained localization capabilities across different IoU thresholds.

Regarding detection reliability, CMLE-YOLO exhibits stable overall performance with P=0.7851, R=0.7893, and F1=0.7872. Its recall outperforms YOLOv6 (0.7571), YOLOv10 (0.7396), and YOLOv11 (0.7787), indicating stronger object coverage capabilities in complex backgrounds and occlusion scenarios, thereby reducing the risk of missed detections.

Regarding computational complexity, CMLE-YOLO achieves 5.2 GFLOPs, lower than YOLOv5 (5.8), YOLOv8 (6.8), YOLOv10 (8.2), and YOLOv11 (6.2), and significantly lower than YOLOv6 (11.5). This result indicates that CMLE-YOLO achieves higher detection accuracy without introducing additional inference overhead, demonstrating deployment potential for real-time tomato ripeness detection and automated counting tasks.

Additionally, CMLE-YOLO demonstrates strong inference speed, achieving 213.5 FPS, which is superior to YOLOv5 (169.8 FPS), YOLOv6 (185.5 FPS), YOLOv8 (190.7 FPS), YOLOv10 (202.5 FPS), and YOLOv11 (195.2 FPS). This highlights the model's efficiency and suitability for real-time applications, particularly in scenarios where high-speed processing is critical.

In summary, CMLE-YOLO achieves a more advantageous trade-off between performance and efficiency: it significantly improves mAP@50 while maintaining low computational overhead, and attains the best performance among the compared models across mAP50–95. The inclusion of inference speed (FPS) further demonstrates the model's potential for deployment in real-time tomato ripeness detection and automated counting tasks.

Table 4

Comparison of Experimental Results

Model	P	R	F1 Score	mAP@50	mAP50-95	GFLOPs	FPS
YOLOv5	0.7806	0.7987	0.7895	0.8291	0.6808	5.8	169.8
YOLOv6	0.8178	0.7571	0.7861	0.8279	0.6905	11.5	185.5
YOLOv8	0.7821	0.7837	0.7829	0.8172	0.6692	6.8	190.7
YOLOv10	0.8201	0.7396	0.7776	0.8225	0.6672	8.2	202.5
YOLOv11	0.7979	0.7787	0.7879	0.8219	0.6762	6.2	195.2
CMLE-YOLO	0.7851	0.7893	0.7872	0.8508	0.6905	5.2	213.5

Visualization of Test Results

To visually validate the suitability of the improved CMLE-YOLO model for tomato ripeness detection scenarios, Fig. 6 presents a comparison of actual detection results between CMLE-YOLO and YOLOv3, YOLOv5, YOLOv8, YOLOv10, and YOLOv11 under tomato cultivation conditions (occlusion confusion, mixed ripeness levels, and simple unobstructed scenarios). The visual results reveal that CMLE-YOLO maintains optimal maturity detection performance across all three scenarios while effectively mitigating common detection flaws observed in the comparison models.

CMLE-YOLO

Fig. 6 - Different models yield different environmental effects

In simple, unobstructed scenarios, while all models achieve basic fruit ripeness recognition, CMLE-YOLO demonstrates higher detection coverage. It avoids the occasional missed detection of small, unripe fruits observed in YOLOv3 and YOLOv10, and its detection confidence is significantly improved compared to the reference models. In occlusion-confusion scenarios, models like YOLOv5 and YOLOv8 frequently missed fruits in leaf-obscured areas and confused boundaries of clustered fruits. CMLE-YOLO, however, precisely segmented the contour boundaries of clustered fruits and stably identified fruits in obscured regions, achieving a significantly lower miss rate than the comparison models. In mixed-ripeness scenarios, YOLOv8 and YOLOv11 often misclassify tomatoes with different ripeness levels (e.g., half-ripe and fully ripe). CMLE-YOLO clearly distinguishes categories like “green”, “half_ripened” and “fully_ripened” achieving superior classification accuracy and confidence compared to the comparison models.

Heatmap

This paper generates Grad-CAM heatmaps based on the corresponding feature layers upstream of the detection head. It compares the same batch of tomato samples across three scenarios: occlusion confusion, mixed maturity levels, and simple unobstructed conditions, as shown in Fig. 7.

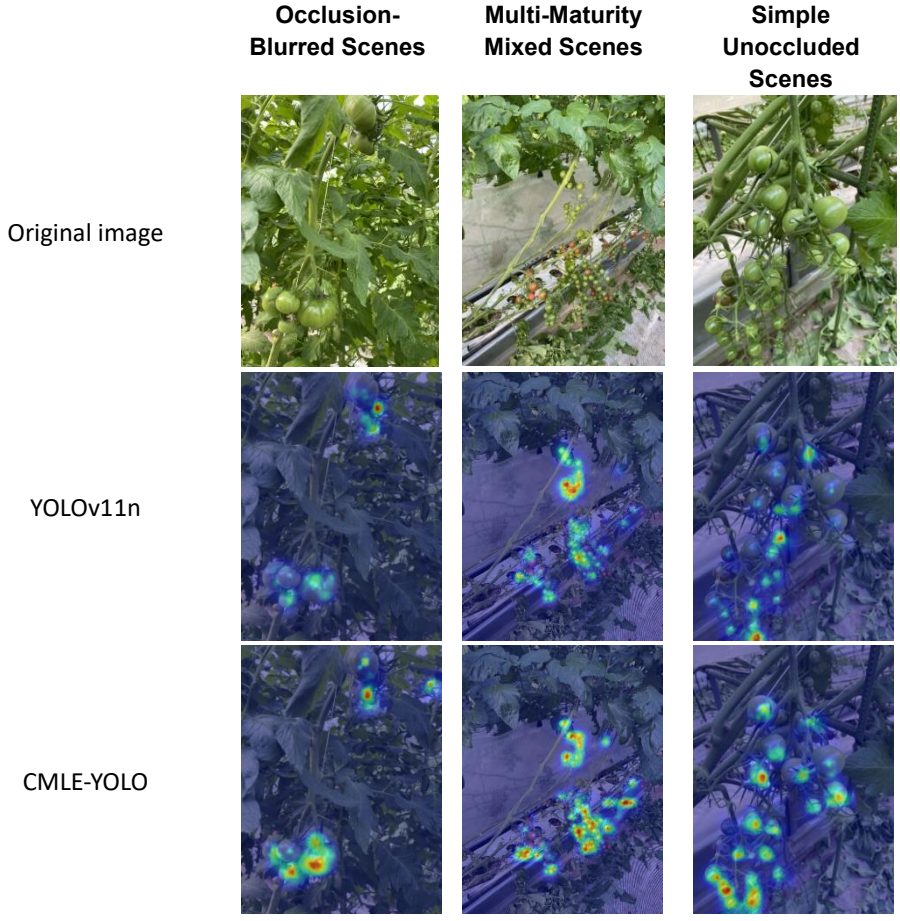

Fig. 7 - Heatmaps for Different Models

In occlusion-cluttered scenes, compared to YOLOv11, CMLE-YOLO's high-response regions more precisely focus on the main fruit bodies obscured by leaves, while the thermal activation levels of background leaves and branches are significantly reduced. This demonstrates its superior background interference resistance and target feature focusing capabilities in occluded environments, helping to reduce missed detections of fruits in obscured areas.

In mixed-maturity scenarios, CMLE-YOLO exhibits clearly classified thermal hotspot distributions, forming distinct high-response regions for tomatoes at different maturity levels. This avoids the hotspot confusion between fruits of varying maturity observed in YOLOv11. In simple, unobstructed scenarios, CMLE-YOLO achieves more comprehensive thermal coverage. Even small, immature fruits generate distinct hotspots, effectively mitigating the occasional weak activation of small targets observed in YOLOv11.

Threshold and Performance Analysis Visualization

Fig. 8 presents the precision-recall performance of the model on the validation set, along with the rationale for threshold selection.

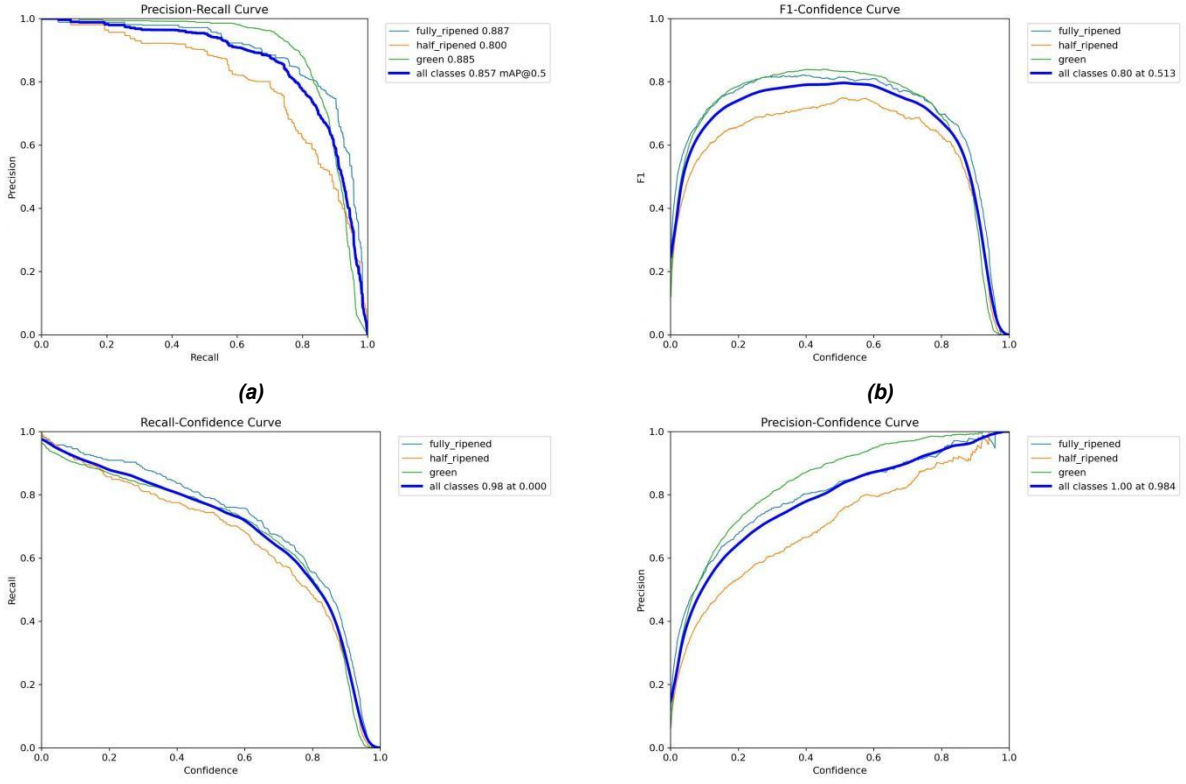

Fig. 8 - Performance Curves and Threshold Selection: (a) PR curve; (b) F1–Confidence; (c) Recall–Confidence; (d) Precision–Confidence

The PR curve results indicate that the AP values for the fully_ripened, half_ripened, and green categories are 0.887, 0.800, and 0.885 respectively, with an overallmAP@0.5of 0.85. The relatively low AP for half_ripened stems primarily from its position in a continuous transition zone between maturity stages. Its visual characteristics exhibit significant overlap with fully ripe and unripe categories, and it is more susceptible to occlusion and light fluctuations, thereby increasing misclassifications and missed detections. As the confidence threshold increases, Precision gradually rises while Recall correspondingly decreases, reflecting the trade-off between the two metrics. Comprehensive F1 curve analysis indicates that the best overall performance across all categories (F1=0.80) is achieved at conf=0.513. Therefore, subsequent experiments and inference stages in this paper uniformly adopt conf=0.513 as the default confidence threshold.

Visualization of the Training Process

Fig. 9 illustrates the trends in various loss curves and core performance metrics during the training process of the CMLE-YOLO model. During both training and validation phases, box_loss, cls_loss, and dfl_loss steadily decreased with each iteration. Concurrently, core performance metrics such as precision, recall, mAP50, and mAP50-95 consistently improved and stabilized in the later stages of training.

This indicates continuous optimization throughout the training process, ultimately achieving effective convergence. The curve patterns between the training and validation sets exhibit high consistency, with minimal numerical gaps and no significant divergence. This confirms the model avoids substantial overfitting and demonstrates strong generalization capabilities. Combined with the high stability of performance metrics in the later training stages, the current training configuration converges the model to an optimal solution. This provides reliable performance support for tomato ripeness detection tasks across diverse scenarios.

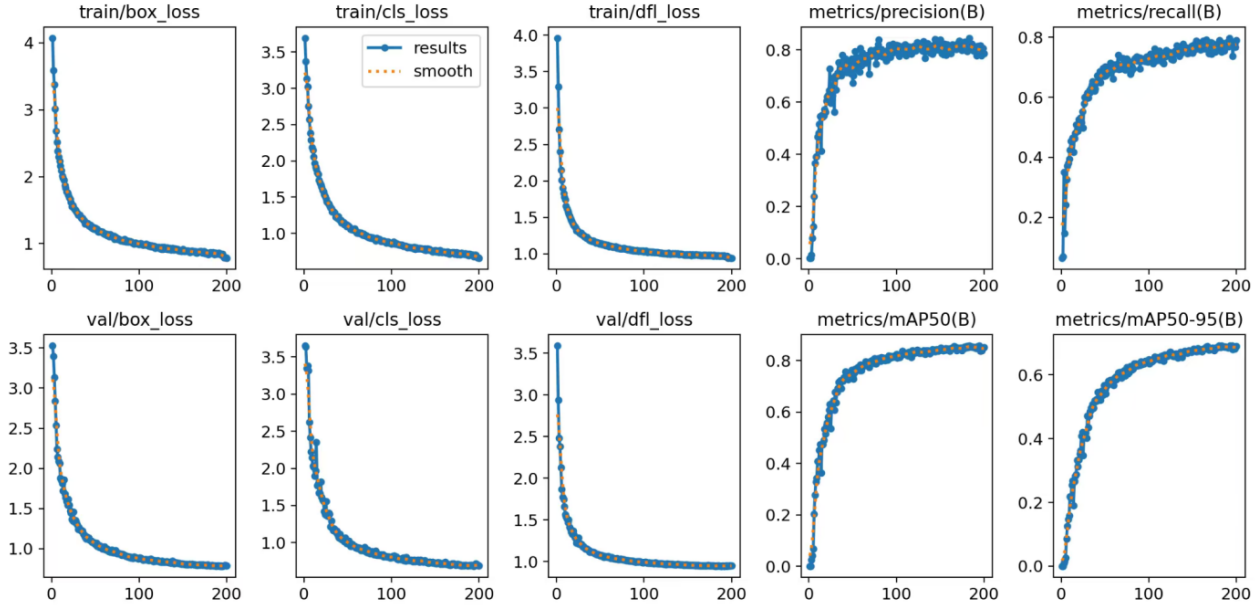

Fig. 9 - Training Process Loss and Metric Change Curve

The standardized confusion matrix results are shown in Fig. 10, indicating that the model demonstrates good discrimination capabilities across different maturity categories. Overall, the proportions of each maturity category along the diagonal are significantly higher than those of the off-diagonal elements, suggesting that most samples can be correctly identified. Misclassifications primarily occur between adjacent transitional maturity categories, such as confusion between fully_ripened and half_ripened, or half_ripened and green. These errors typically stem from the gradual color and texture changes during maturity transitions, coupled with unclear inter-class boundaries. Such issues are further amplified under occlusion and fluctuating lighting conditions. Meanwhile, confusion between the background and fruit categories remained low, indicating the model's strong background suppression and target separation capabilities. Overall, the confusion matrix validates the model's classification effectiveness for tomato ripeness detection, though further optimization is needed for fine-grained differentiation during transitional ripeness stages.

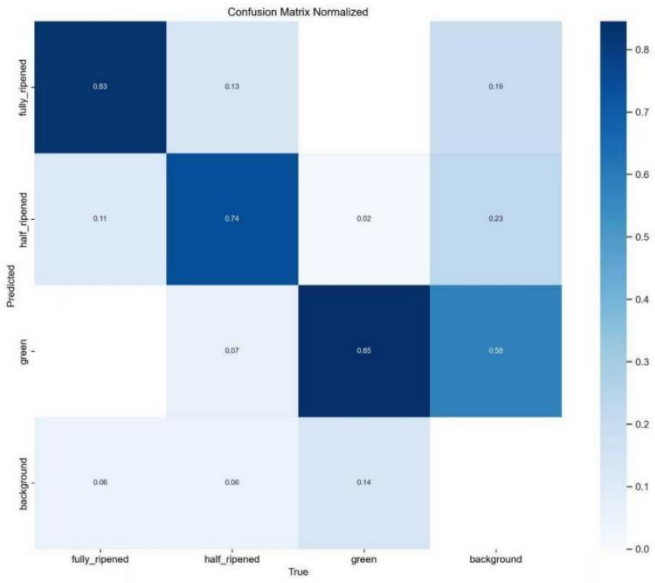

Fig. 10 - Confusion Matrix Normalized

To evaluate the model's detection capability on tomatoes at different maturity levels, this study further conducted quantitative statistics based on the test set. Specifically, the trained CMLE-YOLO model was applied to test set images, and the quantities of unripe(green), semi-ripe(half_ripened), and ripe(fully_ripened) tomatoes were counted based on the detection results, as shown in Fig.11. This statistical process not only reflects the model's detection performance across different maturity stages but also provides direct evidence for quantitatively assessing maturity distribution. This validates the model's effectiveness in both maturity classification and quantity estimation tasks.

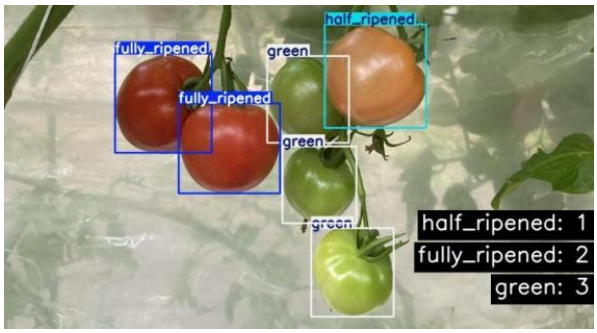

Fig. 11 - Detection and Automatic Counting Results for Tomatoes at Different Maturity Stages

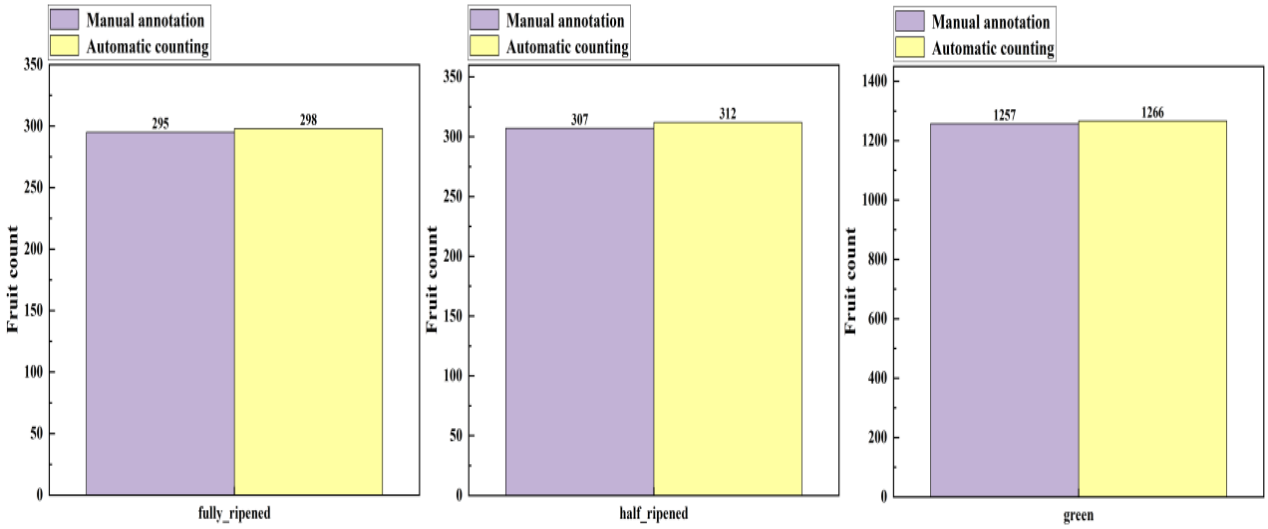

Fig. 12 - Automatic counting of tomatoes at different maturity stages compared with manual annotations

Figure 12 shows a comparison between the automatic counting results and manual annotations for three categories of tomatoes: fully_ripened, half_ripened, and green. In the fully_ripened and half_ripened categories, the differences between the automatic counts and manual annotations were 3 and 5, respectively. In the green category, although the number of targets was higher than in the other categories, the automatic counting results still showed a high level of consistency with the manual annotations, with a difference of only 9. These results indicate that the proposed CMLE-YOLO model demonstrates strong stability and reliability in estimating the number of tomatoes at different maturity stages. In summary, CMLE-YOLO not only provides high-accuracy object detection, but also effectively estimates the number of tomatoes at different maturity stages, offering a feasible automated strategy for fruit monitoring and yield assessment.

CONCLUSIONS

This study developed CMLE-YOLO, a lightweight framework for tomato ripeness detection and automatic counting under complex agricultural conditions. Using a dataset of 2,000 images with 8,593 annotated tomato instances, the method achieved end-to-end recognition of three ripeness stages, namely green, half-ripened, and fully ripened, while simultaneously providing corresponding counting results. The experimental results showed that CMLE-YOLO achieved an mAP@50 of 0.8508 and an mAP50–95 of 0.6905, with a recall of 0.7893.

In addition, the model maintained low computational complexity at 5.2 GFLOPs and reached an inference speed of 213.5 FPS, indicating a favorable balance between detection accuracy and deployment efficiency for real-time agricultural applications.

For the counting task, the automatic counting results were highly consistent with manual annotations across the three ripeness categories. The differences between predicted and manually labeled counts were 3 for fully ripened tomatoes, 5 for half-ripened tomatoes, and 9 for green tomatoes. These results confirm that the proposed method is not only effective for ripeness detection, but also capable of providing reliable quantity statistics for yield estimation and field management.

Despite these advantages, several limitations remain. Misclassifications still mainly occurred between adjacent ripeness stages, especially between half-ripened and fully ripened or green tomatoes, due to gradual transitions in color and texture. Detection performance was also affected by abrupt illumination changes, severe occlusion, and high similarity between fruits and background objects. These findings indicate that relying solely on RGB appearance information remains insufficient in some challenging scenarios.

Future work should therefore focus on improving generalization and long-term deployment reliability by expanding data diversity across tomato varieties, seasons, lighting conditions, and background environments, and by introducing more robust strategies such as illumination-invariant representation and domain-adaptive training. Additional cross-domain evaluation and long-term hardware deployment tests would further support the practical application of this approach in agricultural harvesting, grading, and yield assessment.

ACKNOWLEDGEMENT

This research is supported by the Research on Incremental Multi-Fish Disease Recognition Method for Model Intelligent Growth (Shanxi Provincial Basic Research Program Youth Scientific Research Project, NO.202303021222039; 2024.01—2026.12; Funding: 50,000 CNY; In progress; Principal Investigator).

REFERENCES

[1] Al-Dairi, M., Pathare, P. B., Al-Yahyai, R. (2021). Effect of postharvest transport and storage on color and firmness quality of tomato. *Horticulturae*, Vol. 7, No. 7, 163.

[2] Badgujar, C. M., Poulose, A., Gan, H. (2024). Agricultural object detection with You Only Look Once (YOLO) algorithm: A bibliometric and systematic literature review. *Computers and Electronics in Agriculture*, Vol. 223, 109090.

[3] Feng, C., Zhong, Y., Gao, Y., Scott, M. R., Huang, W. (2021). TOOD: Task-aligned one-stage object detection. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, Montreal, QC, Canada, pp. 3510-3519.

[4] Food and Agriculture Organization of the United Nations. (2019). The State of Food and Agriculture 2019: *Moving Forward on Food Loss and Waste Reduction*. FAO, Rome, Italy.

[5] Fu, Y., Li, W., Li, G., Dong, Y., Wang, S., Zhang, Q., Li, Y., Dai, Z. (2024). Multi-stage tomato fruit recognition method based on improved YOLOv8. *Frontiers in Plant Science*, Vol. 15, 1447263.

[6] Hu, J., Shen, L., Sun, G. (2018). Squeeze-and-Excitation Networks. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Salt Lake City, UT, USA, pp. 7132-7141.

[7] Huang, W., Hu, N., Xiao, Z., Qiu, Y., Yang, Y., Yang, J., Mao, X., Wang, Y., Li, Z., Guo, H. (2022). A molecular framework of ethylene-mediated fruit growth and ripening processes in tomato. *The Plant Cell*, Vol. 34, No. 9, pp. 3280-3300.

[8] Ioffe, S., Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift. *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, Lille, France, Vol. 37, pp. 448-456.

[9] Koirala, A., Walsh, K. B., Wang, Z., McCarthy, C. (2019). Deep learning - Method overview and review of use for fruit detection and yield estimation. *Computers and Electronics in Agriculture*, Vol. 162, pp. 219-234.

[10] Li, J., Huang, Z., Xia, L., Sun, H., Wang, H. (2025). Tomato maturity detection based on improved YOLOv8n. *INMATEH - Agricultural Engineering*, Vol. 75, No. 1, pp. 619-629. DOI: https://doi.org/10.35633/inmateh-75-53

[11] Li, R., Ji, Z., Hu, S., Huang, X., Yang, J., Li, W. (2023). Tomato maturity recognition model based on improved YOLOv5 in greenhouse. *Agronomy*, Vol. 13, No. 2, 603.

[12] Lin, T.-Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S. (2017). Feature pyramid networks for object detection. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, pp. 2117-2125.

[13] Liu, S., Qi, L., Qin, H., Shi, J., Jia, J. (2018). Path aggregation network for instance segmentation. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Salt Lake City, UT, USA, pp. 8759-8768.

[14] Ma, F., Li, S., Tan, J., Li, Y. (2025). YOLO-TRS: An improved YOLO11 for tomato fruit ripeness and stem detection. *INMATEH - Agricultural Engineering*, Vol. 77, No. 3, pp. 1131-1144. DOI: https://doi.org/10.35633/inmateh-77-91

[15] Tan, M., Pang, R., Le, Q. V. (2020). EfficientDet: Scalable and efficient object detection. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA, pp. 10781-10790.

[16] Wang, X., Girshick, R., Gupta, A., He, K. (2018). Non-local neural networks. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Salt Lake City, UT, USA, pp. 7794-7803.

[17] 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*, Vol. 10, No. 9, 1006.

[18] Xu, Y., Li, J., Dong, Y., Zhang, X. (2024). Review of YOLO series object detection algorithms (YOLO 系列目标检测算法综述). *Computer Science and Exploration*, Vol. 18, No. 9, pp. 2221-2238.

[19] Zhao, C., Liu, Y., Xu, J. (2024). Tomato ripeness detection based on YOLOv5 and feature fusion. *Frontiers in Plant Science*, Vol. 15, 1457236.