

A lightweight deep learning model for tomato ripeness detection based on YOLOv8n enhancements

Wang Ziniu¹, Yan Gaizhen², Fan Xiaoyu^{2*}, Cui Hao³, Bai Lujie¹

¹Anhui Science and Technology University/College of Intelligent Manufacturing – Dept. of Agricultural Engineering and Information Technology, 9 Donghua Road – 233100 – Anhui – China.

²Anhui Science and Technology University/College of Electrical and Electronic Engineering – Dept. of Electronic Engineering, 1501 Huangshan Avenue – 233030 – Anhui – China.

³National University of Malaysia/Information Science and Technology College, Lingkungan Ilmu-Bangi – 43600 – Selangor – Malaysia.

*Corresponding author <fanxy@ahstu.edu.cn>

Edited by: Sebastião Martins Filho

Received August 31, 2025

Accepted October 30, 2025

ABSTRACT: Tomatoes (*Solanum lycopersicum* L.) are one of the most popular vegetables across the globe and their maturity directly affects the taste and market value. Accurate maturity recognition methods ensure that tomatoes are harvested at the optimal time and thus increase yield and quality while reducing post-harvest losses. This paper proposes a lightweight tomato ripeness detection model called GhostNetV2 + Bottleneck Transformer Network + Grouped Shuffle Convolution-You Only Look Once version 8 nano (GBG-YOLOv8n). It is an improved model on YOLOv8n (You Only Look Once version 8 nano) architecture to enhance detection efficiency and accuracy. In the GBG-YOLOv8n model, the original Cross Stage Partial 2-focus (C2f) modules in the YOLOv8n backbone are replaced with GhostNetV2 blocks. In the neck network, both the convolutional layers and the C2f modules are substituted by Grouped Shuffle Convolution (GSConv)-based hybrid convolutions and the Volume-wise Network-GSConv-Cross Stage Partial (VoVGSCSP) cross-stage partial modules. Additionally, the Bottleneck Transformer (BoT) self-attention mechanism is incorporated into the main network. Ablation tests, comparison experiments, and recognition tests were conducted on the tomato dataset. Experiments on the tomato dataset revealed that the optimized GBG-YOLOv8n model achieved 92.6 % mean average precision (mAP)_{@50} and 76.3 % mAP_{@50-95}, with a 21 % reduction in parameters and floating-point operations per second (FLOPs), while maintaining 74 % of the original model's functionality. The optimized GBG-YOLOv8n model demonstrates a high degree of accuracy and low computational cost for tomato ripeness detection.

Keywords: GhostNet V2, BoTNet, VoVGSCSP, GSConv

Introduction

Traditional methods of maturity recognition rely heavily on human visual observation and experience-based judgment, which are characterized by high labor intensity and low operation efficiency, leading to inconsistent and inaccurate results due to the subjective nature of the tester. Rapid development of deep learning technologies have led to widespread application in image processing, offering new solutions for the automated recognition of tomato (*Solanum lycopersicum* L.) maturity, which are expected to improve fruit quality, extend shelf life, and reduce economic losses.

A deep convolutional neural network was trained for greenhouse strawberry (*Fragaria × ananassa* Duch.) detection, with average precision of 88.03 % for mature and 77.21 % for immature classes, and bounding-box overlap values of 0.7394 and 0.7045, respectively (Habaragamuwa et al., 2018). Tomato maturity estimation was implemented via a deep neural network trained with cross-entropy and mean-variance loss, achieving an average F1 score of 91 % (Kim et al., 2023). Snapshot-type hyperspectral imagery combined with a support vector classifier (SVC) has been demonstrated to classify tomato maturity into six stages; the VIS-based SVC achieved levels of 79 % accuracy and 88 % F1 in laboratory tests, as well as 75 % accuracy and

86 % F1 in a small greenhouse field tests (Cho et al., 2022). An ensemble stacking framework integrating support vector machine (SVM), decision tree (DT), random forest (RF), and gradient boosting method (GBM) with handcrafted and InceptionV3- Principal component analysis (PCA) features was developed, achieving a level of 90.35 % accuracy in three-class shelf-life prediction (Goyal et al., 2024). A You Only Look Once version 3-based (YOLOv3) method was proposed for classifying four types of tomato maturity, achieving 90.4 % mean average precision (mAP)_{@0.5} at approximately 41 frames per second (Su et al., 2022). Based on the YOLOv4-Tiny model, blueberry maturity recognition performance was improved through data augmentation and model structure optimization, with reported mAP values of 96.2-97.3 % and a model size of about 24 MB (Wang et al., 2021). An improved object detection algorithm based on YOLOv5n (YOLOv5 nano) was proposed for detecting the maturity of cherry tomatoes, reaching 95.2 % mAP and a weight size of only 4.4 MB (Wang et al., 2023).

Despite the achievement in tomato maturity detection scenarios, this research study tried to promote the work further in the areas of accuracy and efficiency. Based on the YOLOv8 framework, a lightweight tomato ripeness detection model is proposed for identifying the four maturity stages: unripe, half-ripe, ripe, and overripe.

Materials and Methods

Dataset acquisition and processing

The dataset used for training and validation in this study was originally collected from public network platforms. The final dataset consists of 3,250 tomato images, divided into training and validation sets at an 8:2 ratio, each comprising 2,600 and 650 images. To accurately assess tomato maturity stage, four categories (unripe, half-ripe, ripe, and overripe) were used to label the tomato images based on appearance features such as color, size, and skin condition. The maturity classification of tomatoes is shown in Figure 1A-D.

To improve the model's generalization ability and avoid overfitting, various data augmentation operations were performed on the original images. Rotation, noise, cropping, translation, and mirroring operation were to simulate different shooting angles and scene changes. These methods effectively increased the size of the training sample while retaining the semantic information of the images. With the help of data augmentation, the training and validation sets were extended to 13,139 and 3,011 images, respectively, which significantly improved the model's ability to recognize tomato images in different environments.

Model construction

YOLOv8 is an efficient object detection model in the YOLO series, which features fast inference and high detection accuracy. YOLOv8n, the lightest model in the series, has approximately 3 million parameters. This makes it suitable for deployment of edge devices or in resource-constrained environments due to its low computational requirements. Its modular design allows for flexible modifications, and provides excellent scalability and practicality.

This study took YOLOv8n as the base model for recognizing tomato ripeness, and the proposed method GBG-YOLOv8n (GhostNetV2 + Bottleneck Transformer Network + Grouped Shuffle Convolution -YOLOv8n), as shown in Figure 2, improved the original YOLOv8n network in three aspects.

1) Part of the traditional convolutions in the YOLOv8n Backbone are replaced with the Cross Stage Partial 2-focus (C2f)_GhostBottleneckV2 structure. By utilizing GhostConv to generate more feature maps, this modification significantly reduces the number of parameters and computational cost, and thereby improves model lightweightness and inference efficiency.

2) To address the limitation of the original model in capturing global context, a Bottleneck Transformer Network (BoTNet) module is introduced at the end of the Backbone. By incorporating the self-attention mechanism, the model enhances information interaction between different regions of the feature maps, which improves its robustness against complex backgrounds and occlusions, and thereby increases the accuracy of maturity recognition.

3) In the Neck and Head, Volume-wise Network-GSConv-Cross Stage Partial (VoVGSCSP) and Grouped Shuffle Convolution (GSConv) modules are employed to reduce redundant information through shortcut connections and feature compression. Combined with group convolution and attention mechanisms, these modules enable efficient spatial and channel fusion, which enhance the multi-scale feature representation and ultimately improve the detection performance of tomatoes at various maturity stages.

Ghost module for backbone lightweighting

This study replaces the YOLOv8n backbone with GhostNetV2 (Du et al., 2025). Compared with the original YOLOv8n backbone network, GhostNetV2 introduces minor changes in parameter number and computational complexity, but the depth and expressive power of feature extraction is largely improved, particularly when processing subtle color differences between different stages of tomato ripeness.

GhostNetV2 is a lightweight network architecture designed for edge computing. Its core design focuses on generating useful features with fewer computations. It further enhances the information aggregation and structural representation capabilities of GhostNet. The information aggregation process is illustrated in

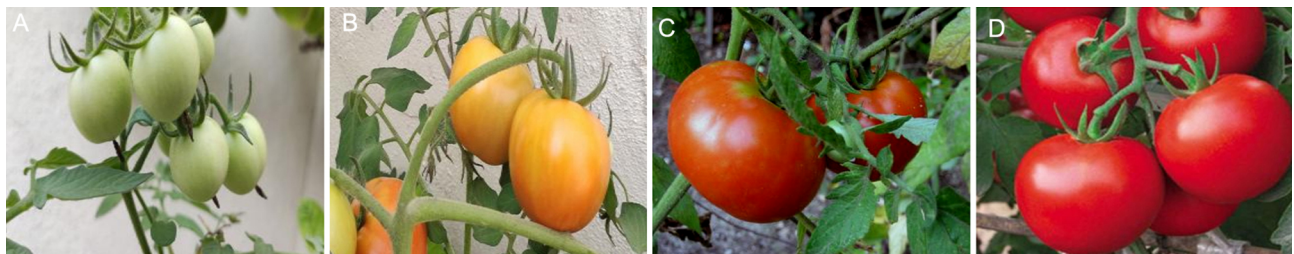


Figure 1 – Four tomato ripeness categories: A) Unripe; B) Half-ripe; C) Ripe; D) Overripe.

Figure 3. The Ghost module first extracts a small number of core features through a single standard convolution. Next, it generates multiple redundant features via a set of linear operations of channel-wise convolution. These redundant features partially

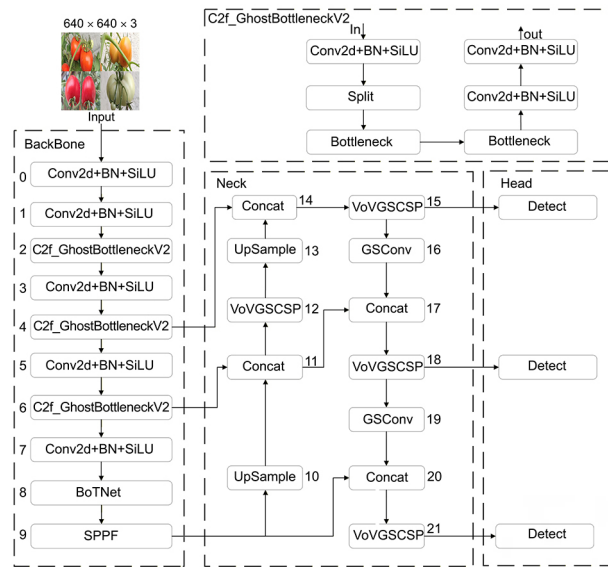


Figure 2 – Architecture of the GhostNetV2 + Bottleneck Transformer Network + Grouped Shuffle Convolution-You Only Look Once version 8 nano model. Conv2d = 2D Convolution; BN = Batch Normalization; SiLU = Sigmoid Linear Unit; BoTNet = Bottleneck Transformer Network; SPPF = Spatial Pyramid Pooling – Fast; VoVGSCSP = Volume-wise Network-GSConv-Cross Stage Partial; GSConv = Grouped Shuffle Convolution.

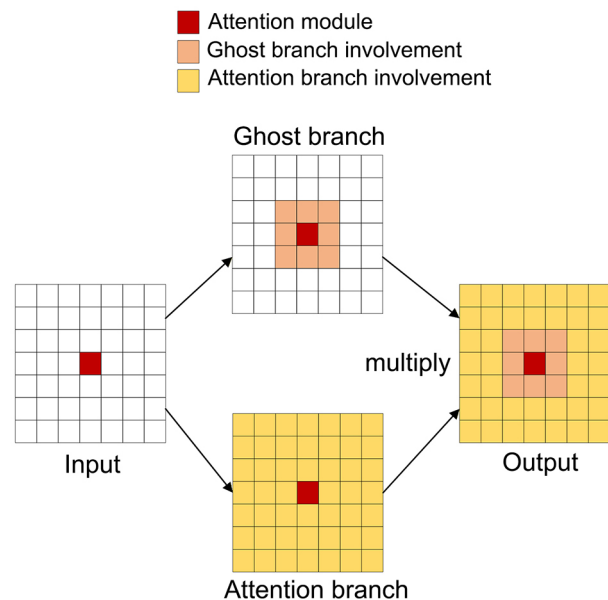


Figure 3 – Information generation and aggregation process of the Ghost module.

mimic the diversity of the original convolution output, and thereby reduce computational costs while retaining sufficient information (Zhang et al., 2024).

To enhance feature expression capabilities further, GhostNetV2 incorporates Ghost re-parameterization mechanism into the Bottleneck structure. During the training process, the bottleneck module creates a multi-branch structure that includes standard, depth-separable, and pixel-wise convolutions. Each type of convolution captures semantic information from different receptive fields and scales. The outputs of these branches are then fused through residual connections. During inference, the multi-branch structure is merged into a single-branch convolution via re-parameterizing. This significantly reduces model complexity and accelerates the inference process.

BoTNet for global feature modeling

Tomato ripeness recognition relies not only on the capture of localized color and texture features, but also on the fruit's overall shape and feature extraction from distant regions. Traditional convolutional neural networks have limited receptive fields and show weakness in understanding relationships between features that are far apart, which limits recognition performance in complex scenarios.

This study introduces the BoTNet attention mechanism (Hussein et al., 2024) into the backbone network to enhance the model's overall information perception and discrimination capabilities, as shown in Figure 4A-B.

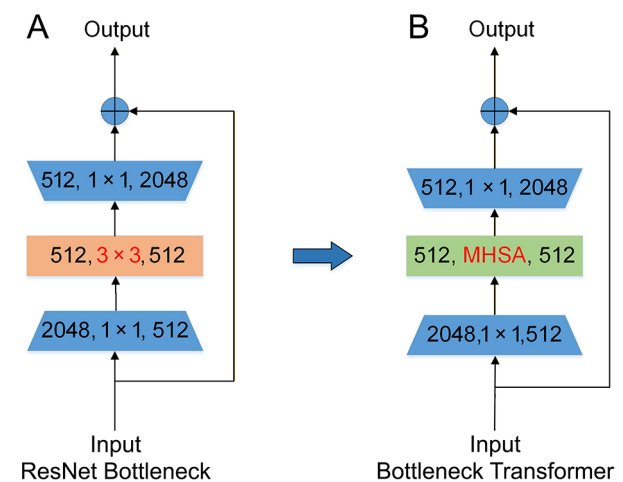


Figure 4 – Structural comparison between Residual Network (ResNet) bottleneck and bottleneck transformer module. A) standard ResNet bottleneck structure; B) bottleneck transformer module formed by embedding the multi-head self-attention (MHSA) module.

BoTNet integrates the transformer's multi-head self-attention (MHSA) module into the traditional Residual Network (ResNet) residual bottleneck structure to replace part of the convolution operations, thereby enhancing the model's capability to capture global dependencies. The core working process of the MHSA mechanism is as follows:

First, the input feature map X is linearly transformed into query (Q), key (K), and value (V), which can be expressed as Eq. (1).

$$Q = W_Q X, K = W_K X, V = W_V X \quad (1)$$

where W_Q , W_K , W_V are learnable weights.

Next, the pixel-to-pixel correlations are calculated through the dot product operation between K and Q . The results are normalized using Softmax to obtain weight matrix A , as shown in Eq. (2).

$$A = \text{Softmax} \left(\frac{QK^T}{\sqrt{d_k}} \right) \quad (2)$$

where K^T and d_k are the transpose of the K matrix and the dimension of the K vectors, respectively.

Next, the weight matrix A is multiplied by the value vector V to obtain the output features, as shown in Eq. (3).

$$\text{Attention}(Q, K, V) = AV \quad (3)$$

Finally, BoTNet performs parallel computations on multiple attention heads, and the outputs of each head are concatenated and linearly mapped to form the final output of the multi-head self-attention module, as shown in Eq. (4).

$$\text{MHSA}(X) = W_0 [\text{head}_1; \text{head}_2; \dots; \text{head}_h] \quad (4)$$

where head_i is the output of the i -th attention head, and W_0 , the output projection weight matrix.

Lightweight convolution and fusion module

Traditional convolution operations process features at all spatial locations uniformly without prioritizing important regions. For tomato ripeness detection, however, features such as color, texture, and shape may be distributed unevenly across images, with certain regions being more critical than others. Color differences and texture changes on tomatoes of different ripeness stage are often subtle, and traditional convolution operations may fail to capture these details adequately.

Inspired by the design philosophy of SlimNeck (Li et al., 2024), this study replaces the Conv and C2f modules in original YOLOv8n's Neck network with the GSConv hybrid convolutional module and the VoVGSCSP cross-stage partial connection module, respectively, which are expected to enhance feature fusion capabilities and improve overall efficiency. The structure of the GSConv module and VoVGSCSP module is shown in Figures 5 and 6, respectively.

GSConv module divides input channels into standard convolution and group convolution and then fuses the outputs through channel mixing. This not only balances the model's feature extraction capabilities and computational efficiency but also improves its ability to utilize inter-channel information. The VoVGSCSP module integrates the GSConv structure with cross-stage partial connection mechanisms. Through multi-layer convolution aggregation and partial residual connections, the VoVGSCSP module achieves deeper semantic information flowing and fusing. This enhances the model's multi-scale perception capabilities and gradient propagation stability.

Results

Experiment setup

In this study, the experiments were conducted on a hardware platform equipped with a GeForce RTX

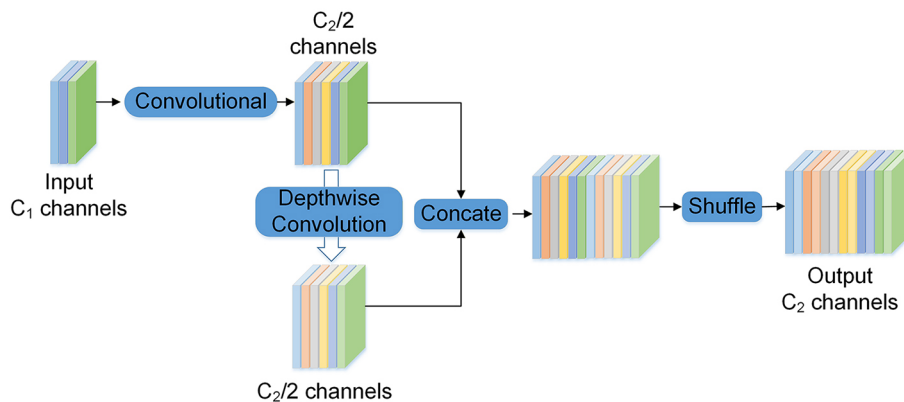


Figure 5 – Structure of the grouped shuffle convolution module.

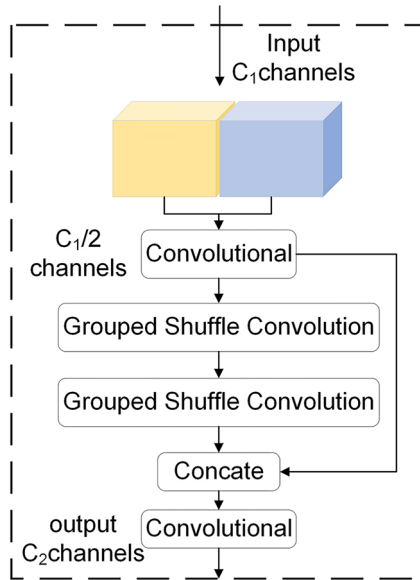


Figure 6 – Architecture of the Volume-wise Network - Grouped Shuffle Convolution - Cross Stage Partial module.

4080 GPU and an Intel Core i9-14900H processor. The software comprised Windows 11, Python 3.9.21, and CUDA 11.8.89. The YOLOv8n model was trained and implemented based on the PyTorch framework.

The input image size was uniformly resized to 640×640 , the batch size set to 32, and the model trained for 50 epochs. The stochastic gradient descent optimizer with a momentum of 0.937 was employed, with an initial learning rate of 0.01 and a weight decay of 0.0005. During training, eight data loading threads were used for parallel processing, and automatic mixed precision was enabled to improve training efficiency. These configurations ensured stable training and superior performance of the model under the experimental environment.

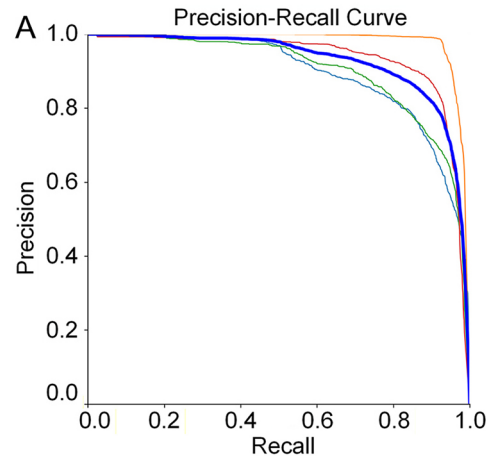
The model's performance was validated using evaluation metrics such as precision (P), recall (R), F1-score, mAP, number of parameters, and (floating-point operations per second) FLOPs (Ren et al., 2025).

Tomato ripeness recognition experiment

To evaluate the overall performance of the model in the tomato ripeness identification task, the precision-recall curves for each ripeness category before and after model improvement are firstly compared and plotted, as shown in Figure 7A-B.

Compared with the original YOLOv8n model, the proposed GBG-YOLOv8n exhibits superior recognition performance across all maturity stages, showing both higher precision and greater robustness. GBG-YOLOv8n achieves mAP@50 of 97.9 %, 93.4 %, 89.8 % and 89.2 % on ripe, unripe, half-ripe and overripe tomatoes, respectively. The model performs

— overripe 0.892 — unripe 0.934
 — ripe 0.979 — all classes 0.926 mAP@0.5
 — halfripe 0.898



— overripe 0.872 — unripe 0.931
 — ripe 0.976 — all classes 0.916 mAP@0.5
 — halfripe 0.884

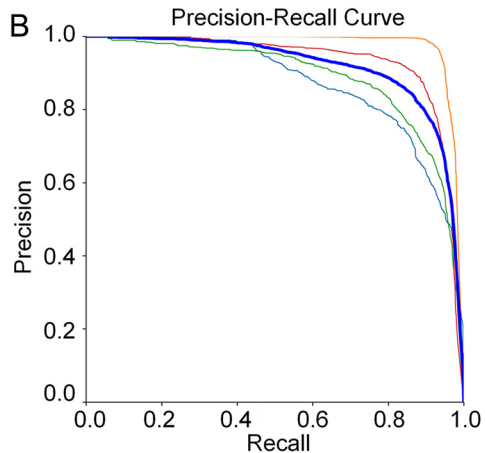


Figure 7 – Precision-recall curves of A) improved model; B) baseline mode.

best on ripe tomatoes, while overripe tomatoes remain comparatively more difficult to distinguish due to their high visual similarity to ripe ones.

Nevertheless, GBG-YOLOv8n consistently surpasses the baseline model in terms of mAP@50, F1-score, precision and recall for all four maturity stages. A more detailed comparison of quantitative metrics is provided in Table 1.

Overall, the proposed model achieves an average mAP@50 of 92.6 %, representing a 1 % improvement over the original model. Moreover, the precision and recall curves demonstrate smooth and stable trends, indicating that the model possesses strong generalization ability and reliable detection performance across different maturity stages. These

results confirm that GBG-YOLOv8n is well-suited to deployment in real-world scenarios such as automated harvesting and intelligent sorting systems.

Ablation test

This experiment introduced the Ghost Bottleneck V2, Bottleneck Transformer (BoT), and GSConv modules based on YOLOv8n to verify the specific impact of each module on the overall performance. Four improved models were designed for ablation tests, where Model 1~3 introduced GhostBottleneckV2, BoT and GSConv to the original YOLOv8n model, respectively and Model 4 introduced the three improved modules. All the models are tested and compared under four metrics, including mAP@50, mAP@50-95, model parameters and giga floating-point operations per second (GFLOPs). The test results are shown in Table 2.

Model 1 shows a slight improvement in mAP@50 and mAP@50-95 compared to the original YOLOv8n while reducing the number of parameters and GFLOPs. This indicates that the GhostBottleneckV2 module enhances performance while maintaining lightweight characteristics. Model 2 shows a slight decrease in mAP@50 and a slight increase in mAP@50-95, which suggests that the BoT module enhances the model's ability to perform fine-grained object recognition. For Model 3, mAP@50 performance is the same to Model 1, and mAP@50-95 performance increases to 76.3 %. This indicates that GSConv significantly improves model

accuracy. Model 4 shows the minimum GFLOPs (6.0) and minimum parameter count (2,376,012), while achieves the highest mAP@50 (92.6 %) and mAP@50-95 (76.3 %). This indicates that using all three modules together effectively improves overall performance while reducing computational complexity.

Ablation studies indicate that each module contributes differently to the overall performance. The Ghost BottleneckV2 primarily reduces computational cost by eliminating redundant feature calculations. The BoT module contributes most significantly to parameter reduction, effectively decreasing model size with only minor impact on detection accuracy. In contrast, the GSConv module provides the largest improvement in mAP@50-95, enhancing small-target detection and overall generalization, while simultaneously offering moderate reductions in parameters and GFLOPs. These results highlight the complementary roles of the modules, with GhostBottleneckV2 and BoT focusing on model efficiency and GSConv optimizing detection precision.

Confusion matrix

The normalized confusion matrix of the improved model on the tomato maturity classification task is shown in Figure 8.

The model demonstrates strong classification performance, particularly in the ripe and unripe categories, with correct classification rates of 93 % and 92 %, respectively. The half-ripe category also achieves a high level of accuracy of 85 %. In contrast,

Table 1 –Detection metrics of GhostNetV2 + Bottleneck Transformer Network + Grouped Shuffle Convolution-You Only Look Once version 8 nano (GBG-YOLOv8n) and YOLOv8n on different tomato maturity stages.

Model	Maturity stages	mAP@50	F1-score	P	R
----- % -----					
GBG-YOLOv8n	ripe	97.9	95.2	98.2	92.6
	unripe	93.4	88.3	89.1	87.5
	half-ripe	89.8	81.2	79.0	83.8
	overripe	89.2	81.5	79.0	84.4
YOLOv8n	ripe	97.6	94.7	98.1	91.5
	unripe	93.1	88.5	90.5	86.6
	half-ripe	88.4	81.4	81.9	80.9
	overripe	87.2	79.5	77.0	82.2

mAP = mean average precision; F1 = balanced F; P = precision; R = recall.

Table 2 – Evaluation results of ablation experiments.

Models	Ghost BottleneckV2	BoT	GSConv	mAP@50	mAP@50-95	Parameters	GFLOPs
----- % -----							
YOLOv8n	— ¹	—	—	91.6	75.6	3,006,428	8.1
Model1	+ ²	—	—	91.9	75.9	2,830,732	7.0
Model2	—	+	—	91.8	76.1	2,761,692	7.9
Model3	—	—	+	91.9	76.3	2,796,444	7.3
Model4	+	+	+	92.6	76.3	2,376,012	6.0

¹Indicates the module is excluded; ²Indicates the module is included. BoT = bottleneck transformer; GSConv = grouped shuffle convolution; mAP = mean average precision; GFLOPs = giga floating-point operations per second; YOLOv8n = You Only Look Once version 8 nano.

the overripe category shows a slightly lower level of accuracy of 84 %, with approximately 10 % of overripe samples being misclassified as half-ripe. Some confusion is observed between the background class and the tomato categories, especially with half-ripe and unripe samples, which are misclassified as background at rates of 3 % and 7 %, respectively. This may be due to edge ambiguity or occlusion in certain images. Overall, the model performs well in identifying clearly distinguishable maturity stages (ripe and unripe), while the intermediate stages (half-ripe and overripe), which have less distinct visual boundaries, remain more challenging and indicate room for further improvement.

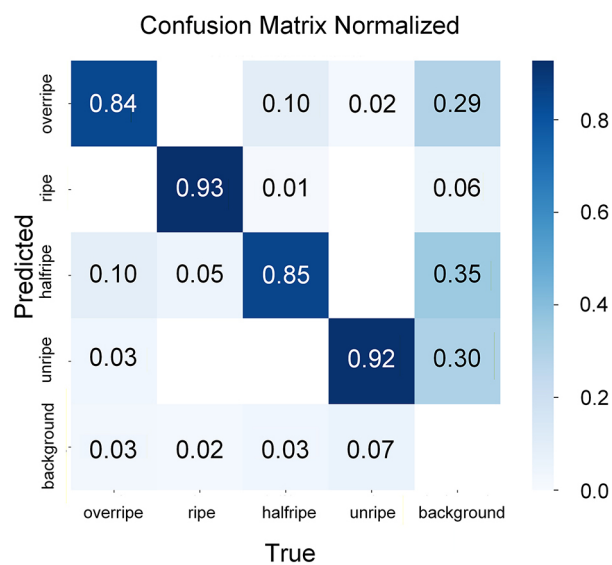


Figure 8 – Confusion matrix.

Backbone replacement experiments

To further validate the superiority of the proposed backbone design in GBG-YOLOv8n, ShuffleNetV2 and MobileNetV3 were introduced as alternative backbones for comparison, as shown in Table 3.

Although both replacements maintain a comparable mAP, neither of them can outperform the original backbone in terms of overall accuracy and stability. Moreover, they introduce a substantial increase in computational cost and parameter size, nearly doubling the GFLOPs. These results clearly demonstrate that the backbone adopted in GBG-YOLOv8n is not only lightweight but also better optimized for tomato maturity detection, making it the most balanced and effective choice among all candidates.

Comparison of GBG-YOLOv8n with other models

Comparative experiments against YOLOv5, YOLOv6, YOLOv8s (YOLOv8 small), and YOLOv10n were firstly conducted to evaluate the overall performance of GBG-YOLOv8n. The results are presented in Table 4.

GBG-YOLOv8n achieved the best results among all the compared models, with mAP@50 and mAP@50-95 reaching 92.6 % and 76.3 %, respectively. Specifically, the mAP@50 is 1.3 % higher than YOLOv10n and 0.7 % higher than YOLOv8s, demonstrating higher detection accuracy. GBG-YOLOv8n has 2.38 million parameters, approximately 21 % of those in YOLOv8s, and significantly fewer than in other models. GFLOPs of the proposed GBG-YOLOv8n is 6.0, the lowest among all models. This indicates lower inference

Table 3 – Performance comparison of different backbone networks.

Models	mAP@50	mAP@50-95	Parameters	GFLOPs	Models size/MB	P	R
	----- % -----					----- % -----	
GBG-YOLOv8n-ShuffleNetV2	92.4	76.1	6,582,684	14.0	13.4	86.7	86.0
GBG-YOLOv8n-MobileNetV3	91.9	76.2	6,478,512	13.5	13.2	86.2	86.4
GBG-YOLOv8n	92.6	76.4	2,376,012	06.0	05.0	86.3	87.1

mAP = mean average precision; GFLOPs = giga floating-point operations per second; GBG-YOLOv8n = GhostNetV2 + Bottleneck Transformer Network + Grouped Shuffle Convolution-You Only Look Once version 8 nano; P = precision; R = recall.

Table 4 – Comparative evaluation results of GhostNetV2 + Bottleneck Transformer Network + Grouped Shuffle Convolution-You Only Look Once version 8 nano (GBG-YOLOv8n) and other models.

Models	mAP@50	mAP@50-95	Parameters	GFLOPs	Models size/MB	Inference time
	----- % -----					ms
YOLOv5	92.0	75.6	2,503,724	7.1	5.2	22.4
YOLOv6	91.6	75.9	4,234,140	11.8	8.6	22.4
YOLOv8s	91.9	76.3	11,127,132	28.4	22.5	57.6
YOLOv10n	91.3	75.3	2,695,976	8.2	5.7	32.0
GBG-YOLOv8n	92.6	76.3	2,376,012	6.0	5.0	25.6

mAP = mean average precision; GFLOPs = giga floating-point operations per second; s = small.

overhead and greater suitability for deployment on resource-constrained devices. In terms of inference time, GBG-YOLOv8n achieves a speed of 25.6 ms, which is significantly faster than YOLOv8s (57.6 ms) and YOLOv10n (32.0 ms), outperforming most other models. GBG-YOLOv8n maintains a small model size and low computational resource consumption while significantly improving detection accuracy.

In comparison with non-YOLO-based tomato maturity estimation methods, GBG-YOLOv8n provides both higher precision and on-device deployability. For example, a deep neural network trained with cross-entropy and mean-variance loss was employed for tomato maturity estimation, achieving an average F1 score of 91 % while supporting both discrete and continuous prediction (Kim et al., 2023). Similarly, an ensemble stacking framework integrating SVM, DT, RF and GBM with handcrafted and Inception-V3-PCA features was developed, achieving 90.35 % accuracy in three-class shelf-life prediction (Goyal et al., 2024). In contrast, our proposed GBG-YOLOv8n simultaneously performs localization and maturity recognition, achieving 92.6 % mAP@50 with significantly lower parameters and faster inference speed, demonstrating superior suitability for deployment under practical agricultural scenarios.

Recognition performance of the GBG-YOLOv8n for tomato ripeness

To validate the GBG-YOLOv8n model's performance in practical tasks, this paper presents a visualization of the model's detection results on tomato images at different stages of maturity, as shown in Figure 9.

The model's recognition of tomato targets under natural lighting and complex backgrounds is demonstrated in the figure.

It covers four stages of tomato maturity: unripe, half-ripe, ripe, and overripe. As illustrated, the improved model accurately identifies and localizes tomatoes at all ripe stages. The bounding boxes closely align with the target contours, and the category classifications are highly accurate, with no obvious false positives or false negatives. These results indicate that GBG-YOLOv8n possesses excellent feature representation and discrimination capabilities. Compared to the original YOLOv8n model, GBG-YOLOv8n shows notable improvements in detection confidence, localization precision, and robustness. For instance, in the case of overlapping or densely clustered ripe tomatoes, GBG-YOLOv8n yields higher confidence scores and more precise bounding boxes. Furthermore, GBG-YOLOv8n maintains stable detection performance under challenging conditions, including varying lighting, occlusion, and low color contrast with the background, outperforming the original model in both accuracy and reliability.



Figure 9 – Tomato ripeness identification results of GhostNetV2 + Bottleneck Transformer Network + Grouped Shuffle Convolution-You Only Look Once version 8 nano (GBG-YOLOv8n) and YOLOv8n.

Discussion

This paper addresses the issue of recognizing tomato ripeness intelligently and proposes a lightweight object detection model. The model's structure and performance have been progressively optimized by introducing Ghost Bottleneck V2, BoT modules, and GSConv structures into the YOLOv8n framework. Ablation experiments demonstrate that the improved model effectively reduces the number of parameters and computational complexity while enhancing prediction accuracy. The improved model achieved 92.6 % mAP@50 and 76.3 % mAP@50-95 in the validation set, while reducing the number of parameters by approximately 21 % and GFLOPs by 26 %. Further comparison with alternative backbones such as ShuffleNetV2 and MobileNetV3 confirms the superiority of the proposed backbone design, as none of the replacement schemes achieved a better balance between accuracy and efficiency.

Despite the overall improvement, the recognition of overripe tomatoes remains less reliable. This limitation primarily arises from the close visual resemblance between ripe and overripe stages, where color distributions largely overlap. Consequently, the

model relies predominantly on chromatic cues while exhibiting limited sensitivity to fine-grained texture degradation such as wrinkling or gloss variation. Furthermore, the relatively small proportion of overripe samples restricts sufficient feature learning for this class.

Future advancements may incorporate texture-aware attention mechanisms, integrate auxiliary physiological indicators (e.g., firmness or spectral reflectance), or employ targeted augmentation strategies to reinforce boundary discrimination between adjacent maturity stages. These observations suggest that, beyond achieving lightweight efficiency, next-generation models should transition from uni-modal color-based classification toward multi-dimensional maturity perception to enhance robustness under real deployment conditions.

Acknowledgments

This research was financially supported by the Natural Science Foundation of Education Department of Anhui Province through the protocols 2022AH051633/2022-12 and 2024AH050293/2024-6, by Anhui Support Project for University Young Talents Program through the protocols gxyq2021202/2021-11, by the key Discipline Construction Project of Anhui Science and Technology University through the protocols XK-XJGY002.

Authors' Contributions

Conceptualization: Fan X. **Data curation:** Wang Z. **Methodology:** Fan X. **Validation:** Wang Z. **Visualization:** Bai L. **Writing-original draft:** Wang Z. **Writing-review & editing:** Yan G, Cui H.

Conflict of interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability statement

The datasets generated and analyzed during the current study are available from the corresponding author on reasonable request.

Declaration of use of AI technologies

The authors declare that no artificial intelligence tools have been used in the writing of this paper.

References

- Cho BH, Kim YH, Lee KB, Hong YK, Kim KC. 2022. Potential of snapshot-type hyperspectral imagery using support vector classifier for the classification of tomato maturity. *Sensors* 22: 4378. <https://doi.org/10.3390/s22124378>
- Du Y, Gao CX, Chen X, Cui MT, Xu LL, Ning AJ. 2025. Mobile malware detection method using improved GhostNetV2 with image enhancement technique. *Scientific Reports* 15: 25019. <https://doi.org/10.1038/s41598-025-07742-8>
- Goyal K, Kumar P, Verma K. 2024. Tomato ripeness and shelf-life prediction system using machine learning. *Journal of Food Measurement and Characterization* 18: 2715-2730. <https://doi.org/10.1007/s11694-023-02349-x>
- Habaragamuwa H, Ogawa Y, Suzuki T, Shiigi T, Ono M, Kondo N. 2018. Detecting greenhouse strawberries (mature and immature) using deep convolutional neural network. *Engineering in Agriculture, Environment and Food* 11: 127-138. <https://doi.org/10.1016/j.eaef.2018.03.001>
- Hussein A, Youssef S, Ahmed MA, Ghatwary N. 2024. MGB-Unet: an improved multiscale unet with bottleneck transformer for myositis segmentation from ultrasound images. *Journal of Imaging Informatics in Medicine* 38: 217-228. <https://doi.org/10.1007/s10278-024-01168-w>
- Kim T, Lee DH, Kim KC, Choi T, Yu JM. 2023. Tomato maturity estimation using deep neural network. *Applied Sciences* 13: 412. <https://doi.org/10.3390/app13010412>
- Li H, Li J, Wei H, Liu Z, Zhang Z, Ren Q. 2024. Slim-neck by GSConv: a lightweight-design for real-time detector architectures. *Journal of Real-Time Image Processing* 21: 62. <https://doi.org/10.1007/s11554-024-01436-6>
- Ren Z, Liang K, Zhang Y, Song J, Wu X, Zhang C, et al. 2025. An intelligent method for detection of small target fungal wheat spores based on an improved YOLOv5 with microscopic images. *Plant Methods* 21: 117. <https://doi.org/10.1186/s13007-025-01436-y>
- Su F, Zhao Y, Wang G, Liu P, Yan Y, Zu L. 2022. Tomato maturity classification based on SE-YOLOv3-MobileNetV1 network under nature greenhouse environment. *Agronomy* 12: 1638. <https://doi.org/10.3390/agronomy12071638>
- Wang L, Qin M, Lei J, Wang X, Tan K. 2021. Blueberry maturity recognition method based on improved YOLOv4-Tiny. *Transactions of the Chinese Society of Agricultural Engineering* 37: 170-178 (in Chinese, with abstract in English). <https://doi.org/10.11975/j.issn.1002-6819.2021.18.020>
- Wang C, Wang C, Wang L, Wang J, Liao J, Li Y, et al. 2023. A lightweight cherry tomato maturity real-time detection algorithm based on improved YOLOV5n. *Agronomy* 13: 2106. <https://doi.org/10.3390/agronomy13082106>
- Zhang H, Dai C, Ren J, Wang G, Teng F, Wang D. 2024. Grape disease identification method based on YOLO v8-GSGF model. *Transactions of the Chinese Society for Agricultural Machinery* 55: 75-83 (in Chinese, with abstract in English). <https://doi.org/10.6041/j.issn.1000-1298.2024.11.008>