agronomy

MDPI

Article

High-Throughput Citrus Detection via Citrus-SGYOLOv2: A Symmetric Ghost-Based Architecture with High-Resolution Feature Fusion

Jinfeng Li [1], Yutian Miao [1], Wenxuan Guo [1], Yuxiang Li [1], Qian Xu [1], Yue Xiang [1], Yanyu Chen [1], Xianyao Wang [1], Yunsen Liang [1] and Jun Li [1,2,*]

[1] College of Information Engineering, Sichuan Agricultural University, Ya'an 625014, China; nnsxddqg@stu.sicau.edu.cn (J.L.); miaoyutian@stu.sicau.edu.cn (Y.M.); 202308436@stu.sicau.edu.cn (W.G.); 202406771@stu.sicau.edu.cn (Y.L.); 202406772@stu.sicau.edu.cn (Q.X.); 202406742@stu.sicau.edu.cn (Y.X.); 202308435@stu.sicau.edu.cn (Y.C.); 202205858@stu.sicau.edu.cn (X.W.); 202308549@stu.sicau.edu.cn (Y.L.)
[2] Sichuan Key Laboratory of Agricultural Information Engineering, Ya'an 625000, China
* Correspondence: lijun@sicau.edu.cn

Abstract

Accurate high-throughput fruit detection is the core prerequisite for precision citrus management. Existing models face a critical trade-off between accuracy for small fruits and computational efficiency, restricting large-scale industry transformation. To resolve this, we propose Citrus-SGYOLOv2, an optimized deep learning architecture specifically engineered for high-throughput phenotypic monitoring. The primary contribution of this work lies in three synergistic innovations: a novel Symmetric Ghost Backbone that prunes architectural redundancy while maintaining hierarchical feature depth; a Citrus Color Prior Calibration Attention Mechanism (Citrus_SE) that embeds physiological chromaticity priors to suppress complex spectral noise from foliage; and a P2-layer-based full-scale fusion strategy designed to recover fine-grained spatial details lost during downsampling. Experiments on our self-built dataset show that Citrus-SGYOLOv2 achieves 95.54% mAP@50 and 77.13% mAP@50–95, outperforming YOLOv11s by 5.03 and 9.90 percentage points respectively. Notably, the model achieves a 48.8% reduction in parameters (4.84 M) while sustaining a high-throughput inference speed of 139.00 FPS. This research provides a robust and efficient foundational framework for intelligent yield estimation and precision orchard management.

Keywords: citrus detection; symmetric Ghost convolution; high-resolution feature fusion; color prior calibration attention; precision agriculture; yield estimation; crop phenotypic traits

Check for updates

Academic Editor: Dimitrios Kateris

Received: 28 March 2026
Revised: 22 April 2026
Accepted: 27 April 2026
Published: 28 April 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

With the deep integration of precision agriculture and digital forestry and fruit technology, intelligent perception of fruit tree phenotypic traits has become the core technical support for realizing scientific and precise management of commercial orchards [1,2]. Against the backdrop of global agricultural modernization and Agriculture 4.0, deep learning-based non-destructive testing technology has shown great potential in improving orchard management efficiency. Historically, the foundation of citrus computer vision was established by pioneering studies from the Valencian Institute of Agricultural Research (IVIA). Landmark research by Moltó et al. [3] and Blasco et al. [4,5] demonstrated the

feasibility of utilizing colorimetric and morphological descriptors for automated citrus fruit quality assessment and sorting. While these classical image processing techniques provided the theoretical baseline for fruit perception, the transition from controlled post-harvest environments to unstructured, dynamic field conditions necessitates a more robust evolutionary approach.

Accurate fruit detection and counting is the core premise of citrus orchard intelligent management, which directly determines the accuracy of yield estimation, the scientificity of water and fertilizer management, and the rationality of harvesting plan formulation. However, in actual commercial orchard production, the current mainstream fruit phenotypic monitoring methods still face multiple core pain points that are difficult to break through. First, the traditional manual fruit counting and yield estimation method has extremely low efficiency and a large error. For a 1000-μ commercial orchard, manual survey requires a lot of labor and time costs, and the yield estimation error is usually more than 15%, which cannot meet the needs of large-scale orchard fine management [6]. Second, the complex unstructured orchard environment brings great challenges to accurate fruit detection: dense branch and leaf occlusion, overlapping fruit clusters, and drastic light fluctuations will lead to the loss of fruit feature information; especially the small fruits in the inner canopy of the tree are prone to missed detection, which directly leads to serious distortion of yield prediction results [7]. Finally, the existing high-precision fruit detection models usually have a large number of parameters and high computational cost, which cannot adapt to the high-throughput batch processing requirements of massive orchard images collected by Unmanned Aerial Vehicles (UAVs) and ground monitoring equipment, and are difficult to popularize and apply in grassroots orchard production [8].

In recent years, convolutional neural network (CNN)-based object detection models, especially the YOLO series, including YOLOv3 [9], YOLOv4 [10], and YOLOv7 [11], have been widely used in agricultural fruit detection tasks due to their excellent balance between detection accuracy and inference speed. However, a critical gap remains between generic computer vision derivatives and specialized agronomic requirements. First, existing studies often treat citrus detection as a standard target recognition task, neglecting the unique spectral signatures and physiological priors of citrus fruits that were identified as critical discriminators in early IVIA studies. Second, most existing models focus on the improvement of detection accuracy, ignoring the computational cost of high-throughput batch processing of large-scale orchard images. Although large-parameter models such as YOLOv11x and the Real-Time Detection Transformer (RT-DETR) have high detection accuracy, their huge computational overhead makes it difficult to adapt to the low-computing platform environment of grassroots orchards [12]. Third, the existing models have insufficient ability to detect small and occluded fruits in the inner canopy. Traditional feature extraction networks will lose critical fine-grained spatial information of small targets after multiple downsampling, resulting in a high missed detection rate of small fruits, which is the main source of error in orchard yield estimation [13].

Despite the rapid advancements in deep-learning-based object detection, existing models often struggle to maintain consistent performance in unstructured orchard environments. Specifically, standard lightweight detectors frequently suffer from significant information loss during successive downsampling layers, leading to the missed detection of small, immature, or inner-canopy citrus fruits. Furthermore, the high spectral similarity between green fruits and dense foliage, coupled with varying illumination, often results in severe background noise interference that general-purpose attention mechanisms fail to suppress. Most critically, current high-precision architectures often demand substantial computational resources, creating a bottleneck for high-throughput batch processing on resource-constrained edge platforms. There is, therefore, a pressing need for a specialized

framework that can synergistically optimize spatial feature recovery, physiological noise suppression, and architectural efficiency.

To address the above core challenges in citrus orchard precision management, this study proposes an improved lightweight high-precision detection network, Citrus-SGYOLOv2, which achieves dual optimization of fine-grained detection accuracy for small/occluded fruits and high-throughput inference efficiency for large-scale orchard images. The core scientific hypothesis of this study is that: by integrating the physiological and optical prior of citrus, symmetric lightweight feature extraction structure, and high-resolution small target perception branch, the model can significantly improve the detection accuracy of citrus fruits in complex orchard environments while reducing computational overhead, and provide technical support for high-throughput phenotypic analysis, intelligent yield estimation, and digital management of citrus orchards. The main contributions of this paper are as follows:

(1) Architectural Reconstruction via Symmetric Ghost Logic: We propose a Symmetric Ghost-integrated backbone that leverages a dual-branch feature generation mechanism to eliminate intrinsic computational redundancy. Unlike standard lightweight adaptations, this architecture is specifically designed to maintain the hierarchical depth and structural integrity of citrus-specific features, facilitating high-throughput processing for large-scale orchard imagery without sacrificing the representational power required for complex backgrounds.

(2) Mechanism-driven Citrus Color Prior Calibration (Citrus_SE): A novel Citrus Color Prior Calibration Attention Mechanism (Citrus_SE) is formulated to bridge the gap between fruit physiological characteristics and deep learning feature extraction. By embedding a dynamic chromaticity mapping branch, the model explicitly utilizes the spectral distinctness of citrus fruits to recalibrate channel weights. This approach transforms general attention into a targeted response mechanism, effectively suppressing spectral noise from dense green foliage.

(3) Fine-grained Spatial-Semantic Decoupling via P2-Fusion: Building on path aggregation networks [14], we implement a high-resolution P2 feature fusion strategy incorporating a 1/4 spatial resolution detection branch. This design enables the deep decoupling and subsequent reconstruction of fine-grained spatial details and deep semantic information. It specifically addresses the information loss of small-target fruits within the inner canopy during successive downsampling, significantly enhancing the fidelity of yield estimation in dense-planting scenarios.

2. Materials and Methods

To provide a clear overview of the research, a comprehensive methodological framework is illustrated in Figure 1. The workflow comprises four primary stages: (1) Data Construction, involving multi-orchard image acquisition and scene-level splitting; (2) Architectural Design, where the Citrus-SGYOLOv2 is constructed using the Symmetric Ghost backbone, Citrus_SE module, and P2-fusion strategy; (3) Model Training, utilizing a high-performance GPU platform with specific optimization hyper-parameters; and (4) Comprehensive Evaluation, which validates the model's robustness and efficiency through ablation studies and cross-model comparisons. This structured approach ensures the reproducibility of the study and highlights the integration of physiological priors into the deep learning pipeline.

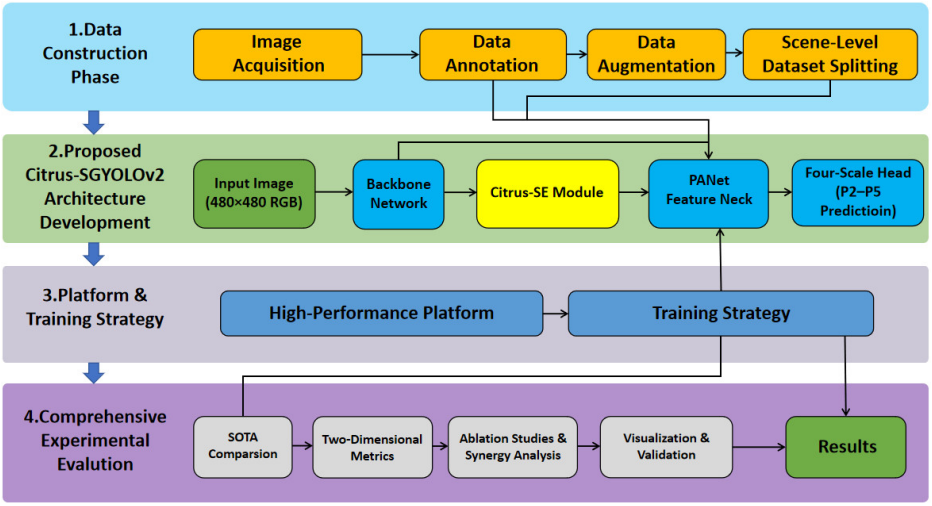


Figure 1. Methodological framework of the study.

2.1. Dataset Construction and Preprocessing

2.1.1. Preliminary Image Acquisition

The initial citrus image dataset used in this study was collected and compiled by our research group during a previous field survey at a standardized citrus orchard in Dongpo District, Meishan City, Sichuan Province (30° N, 103° E), characterized by a typical subtropical monsoon climate. The dataset covers multiple typical citrus orchard environments in Sichuan Province and was captured using a Canon EOS R10 high-resolution Digital Single-Lens Reflex (DSLR) camera (Canon Inc., Tokyo, Japan) and a Huawei Mate60 smartphone(Huawei Technologies Co., Ltd., Shenzhen, China), with original acquisition resolutions of 6000 × 4000 and 3072 × 4096, respectively. To ensure experimental reproducibility and image consistency, the DSLR camera was operated in manual mode (M) with the following standardized settings: an aperture of f/8.0 was utilized to maintain sufficient depth of field for inner-canopy fruits, the sensitivity was fixed at ISO 400 to minimize sensor noise, and the color temperature was locked at 5500 K (Daylight) to prevent automatic white balance fluctuations from distorting the citrus chromaticity priors. The shooting distances were maintained between 0.5 m and 2.5 m using a fixed focal length of 35 mm (35 mm equivalent), and shooting heights were controlled between 0.8 m and 1.8 m. To simulate the visual complexity of real-world harvesting scenarios, the image acquisition encompassed two weather conditions (sunny and cloudy), three lighting angles (front-lit, back-lit, and side-lit), and three fruit development stages (green fruit stage, fruit expansion stage, and ripening stage) [15].

2.1.2. Analysis of Dataset Characteristics

The dataset utilized in this study comprises 20,109 citrus instances collected from three geographically independent orchard sites to ensure high environmental diversity. Unlike traditional fruit detection [16] tasks, our dataset presents significant challenges in terms of target scale, physiological appearance, and orchard background. To prevent potential data leakage and ensure the model's true generalization capability, the data split (training, validation, and testing sets) was strictly performed at the orchard scene and individual tree levels; images captured from the same tree or within the same acquisition session were assigned to the same subset to ensure that no overlapping scenes or correlated visual structures appeared across the splits.

The diversity of the dataset is reflected in its complex multi-scale distribution and environmental variability. The captured area of citrus targets in the images exhibits a drastic range from 16 × 16 to 420 × 420 pixels, a phenomenon driven by the synergy of

two factors: (1) inherent physiological diversity, as the dataset covers 3 varieties (Ehime, Bapa, and Wogan) across multiple growth stages from early expansion to full maturity, and (2) perspective effects, resulting from the dynamic acquisition distances (0.5 m to 2.5 m) and varying depths of fruits within the dense canopy. From an agronomic perspective, the appearance of the fruit evolves significantly with its physiological development: immature fruits are characterized by a dense distribution of oil glands on the flavedo and high chlorophyll content, resulting in a rugose texture and a dark green hue that closely mimics the background foliage. As ripening progresses, the degradation of chlorophyll and the accumulation of carotenoids trigger a "color-break" stage, leading to non-uniform surface coloration patterns (patchy yellow-green transitions) before reaching full maturity. Specifically, small targets (smaller than 64 $\times$ 64 pixels) account for 37.2% of the total, many of which represent immature fruits located in the inner canopy whose cryptic green coloration and subtle surface patterns pose extreme challenges for feature discrimination. Furthermore, the complexity is intensified by the fact that 41.6% of the targets are obscured by branches and leaves, 28.3% are overlapping, 32.7% are captured in strong light or backlit conditions, and 21.5% are captured in overcast or low-light conditions. This rigorous scene-level independence and multi-dimensional diversity arising from both biological and optical factors provide a robust foundation for validating the model's adaptability in real-world, complex orchard scenarios.

2.1.3. Data Augmentation

To achieve the objectives aligned with the characteristics of the dataset required for this study, while preventing model overfitting [17] and improving its generalization performance on low-computing-cost platforms, this study performed corresponding online and offline data augmentation on the dataset prior to training: (1) Geometric Transformations: These include random rotation (-90° to 90°), horizontal/vertical flipping, and random scaling to accommodate the dynamic changes in the camera perspective of the precision orchard management. (2) Photometric Transformations: Adjusting brightness, contrast, and saturation, and introducing Gaussian noise to enhance the model's resilience to uneven lighting in complex orchards and sensor noise. (3) Advanced Techniques: Applying Mosaic and Mixup augmentation strategies [18] during training, which utilize multi-image context stitching and feature fusion to force the model to learn finer-grained local features of citrus fruits.

2.1.4. Data Set Annotation and Partitioning

In this study, the professional annotation tool LabelImg was used to perform bounding box annotations on all images, with the annotation label named "orange". The annotation guidelines followed the following principles: (1) Fully annotate all citrus objects in the image with a visibility of more than 10%. (2) For severely occluded objects, infer the overall outline based on the visible portions to ensure the completeness of the annotation bounding boxes. (3) Annotate adjacent, overlapping fruits independently to avoid model training bias caused by merged bounding boxes.

All annotations were organized into standardized data structures where each target is represented by its class index and normalized bounding box coordinates (x_{center}, y_{center}, width, height) in a plaintext format compatible with the YOLO architecture, and underwent cross-checking and secondary correction by two researchers with professional agricultural expertise. To quantify annotation quality, this study calculated inter-annotator consistency metrics: the Intersection over Union (IoU) consistency reached 98.32%, and Cohen's Kappa coefficient reached 0.976, demonstrating that the dataset annotations possess extremely

high accuracy and reliability, capable of providing high-quality supervised signals for model training.

Following data augmentation and annotation, the dataset ultimately comprised 2438 valid images, which were strictly divided into a training set (1706 images), a validation set (488 images), and a test set (244 images) in a 7:2:1 ratio. This partitioning process ensured a balanced distribution of challenging scenarios (such as overlaps and occlusions) across each subset, thereby preventing a decline in model generalization caused by data distribution skew.

Figure 2 presents representative samples from the citrus dataset constructed for this study, encompassing various typical challenging scenarios encountered in real orchard environments. As shown in Figure 2a, drastic variations in light intensity can cause specular reflections or shadows on fruit surfaces, increasing the difficulty of feature extraction. Figure 2b,c depict severe foliage occlusion and fruit overlap, scenarios that readily fragment target features or blur their boundaries. Additionally, Figure 2d illustrates significant scale diversity, particularly with tiny targets in the background, demanding exceptionally high spatial resolution from feature maps. Figure 2e simulates low-contrast environments typical of early morning or shaded conditions. Finally, Figure 2f presents precise ground truth annotations, providing high-quality supervised signals for model training. This diverse scene design comprehensively evaluates model robustness in complex agricultural settings.

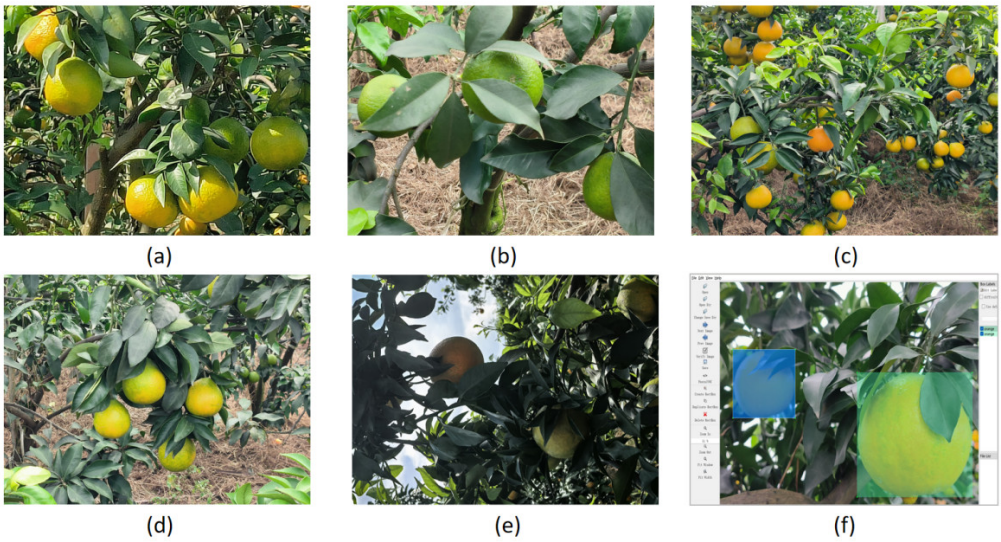


Figure 2. Representative Sample of Citrus Dataset. (**a**) Illumination variation; (**b**) branch and leaf occlusion; (**c**) overlapping clusters; (**d**) scale diversity with tiny targets; (**e**) low-light environment; (**f**) example of ground truth annotations.

2.2. Citrus-SGYOLOv2 Network Architecture

Addressing the practical challenges faced by precision orchard management—namely, limited computational power on low-computing-cost platforms and extremely high requirements for localization accuracy—this paper presents Citrus-SGYOLOv2, a lightweight, high-resolution detection network developed through targeted improvements and a lightweight redesign of YOLOv11. The overall architecture consists of three core modules: a symmetric Ghost-based lightweight backbone network, which leverages linear operations to generate redundant feature maps for reducing computational density while maintaining structural symmetry to preserve hierarchical feature depth; a feature neck with a bio-inspired color-priority calibration attention mechanism (Citrus_SE) that utilizes physiological Hue, Saturation, Value (HSV) priors to suppress spectral noise; and a four-scale detection head incorporating a P2 layer to recover fine-grained spatial details lost during downsampling.

This design aims to achieve an optimal balance between computational efficiency and fine-grained feature representation, effectively enhancing the detection performance of citrus targets, particularly the localization accuracy of small and occluded targets. A schematic of the overall network architecture is shown in Figure 3.

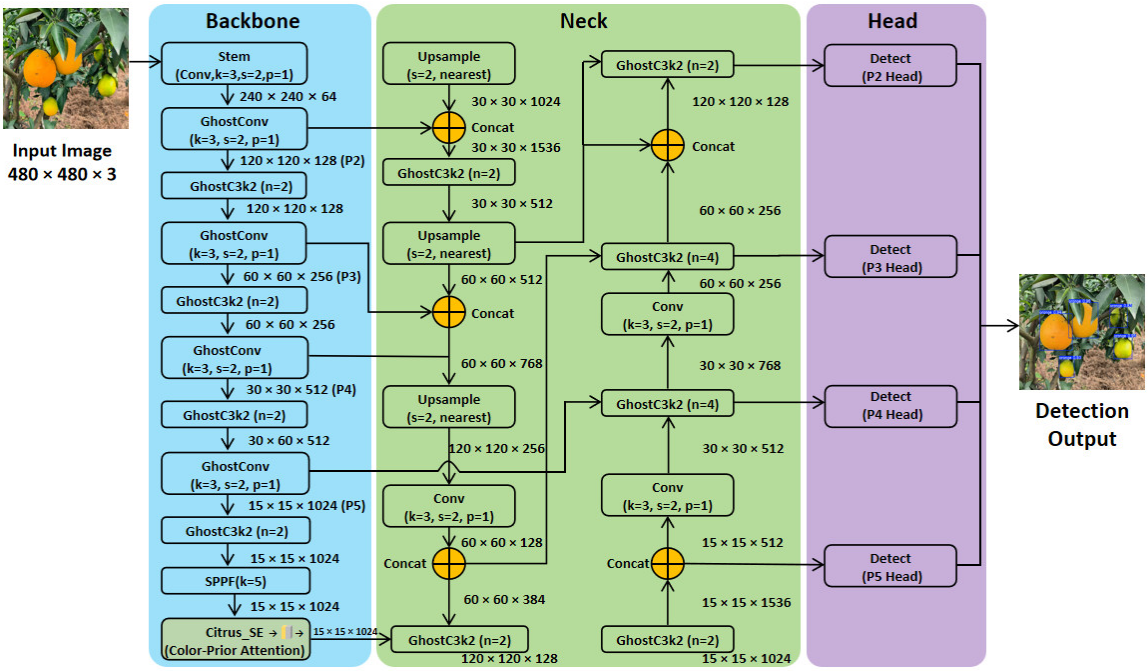


Figure 3. The Overall Architecture of Citrus-SGYOLOv2.

When an Red, Green, Blue (RGB) image with a network resolution of 480 × 480 is input, the backbone network efficiently extracts multi-level features using the GhostC3k2 module. Subsequently, the Citrus_SE module performs spectral information-weighted calibration on the deep-layer features along the channel dimension to suppress background noise interference. The feature neck adopts an improved Path Aggregation Network (PANet) architecture, integrating multi-scale features from the backbone network and introducing a new P2 layer for high-resolution features, enabling high-detail representation at a resolution of 120 × 120. Ultimately, the four detection heads correspond to resolutions of 120 × 120 (P2), 60 × 60 (P3), 30 × 30 (P4), and 15 × 15 (P5), enabling full-scale collaborative prediction ranging from micro-details to macro-contours.

2.2.1. Symmetric Ghost Lightweight Backbone Network

Standard convolution [19] often suffers from significant computational redundancy and parameter explosion when extracting deep, high-dimensional features. The primary reason for this inefficiency is that standard convolution utilizes a unique set of learnable filters to generate every single output channel, often resulting in many highly similar or redundant feature maps. To enable high-throughput detection of citrus objects on resource-constrained low-computing-cost platforms, this study combines the gradient truncation approach of Cross Stage Partial Network (CSPNet) with the feature reuse concept of Ghost Network (GhostNet) to construct the symmetric GhostC3k2 module [20] as the core operator of the backbone network (as shown in Figure 4).

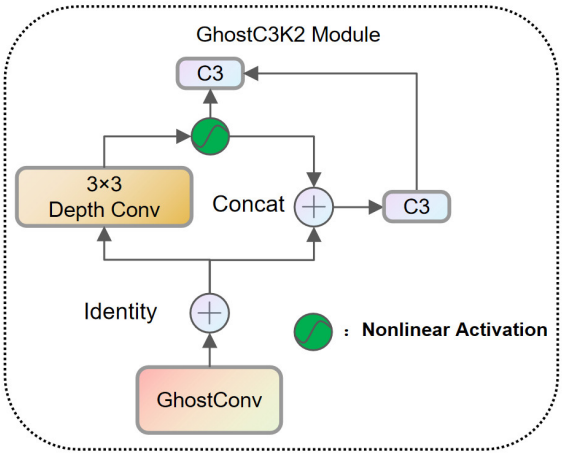


Figure 4. Symmetrical GhostC3k2 Module Architecture.

The main difference between Ghost convolution and a simple convolution architecture lies in the feature generation mechanism. While a simple convolution generates all features in a single, computationally expensive step, the core idea behind Ghost Convolution is to decouple this process. It first generates a limited set of intrinsic feature maps using a small number of standard convolutions, and then applies low-computational-cost linear transformations as described in Equation (1) to derive "ghost" feature maps that mimic the redundant information found in standard layers. Finally, it concatenates the intrinsic and derived maps to obtain the complete output features. This mathematical relationship is formally defined in Equation (1). This approach significantly reduces computational load while preserving the integrity of the feature representation. Its core formula is:

$$Y = \mathrm{Concat}\left(Y', \Phi\left(Y'\right)\right). \tag{1}$$

As shown in Equation (1), Y' denotes the intrinsic feature map generated via standard convolution, and Φ denotes a linear transformation with low computational cost. By utilizing this two-stage derivation instead of exhaustive convolution, the model avoids the "parameter explosion" typical of deep CNNs. By combining GhostC3k2 with the CSP architecture, FLOPs are significantly reduced while preserving the richness of the gradient flow [21].

For a given input feature map $X \in R^{H \times W \times C}$, standard convolution to generate outputs across all N channels results in high parameter counts and computational overhead because every filter must span the entire input depth. In the symmetric Ghost architecture, feature generation is decoupled into two symmetric stages: "inherent feature extraction" and "low-overhead feature derivation." First, standard convolution using a small number of convolutional kernels ($m \ll N$) is used to extract the key inherent feature map Y':

$$Y' = X * W_{\mathrm{intrinsic}}. \tag{2}$$

Subsequently, Y' is mapped in the spatial dimension via a linear transformation operation with low computational cost (depthwise convolution) to generate the derived feature map (Ghost Features) $F(Y')$. The final output features are formed by concatenating these two parts along the channel dimension, inherently creating a symmetric structure:

$$Y_{\mathrm{out}} = \left[Y', \Phi_1\left(Y'\right), \Phi_2\left(Y'\right), \ldots, \Phi_s\left(Y'\right)\right]. \tag{3}$$

Unlike the traditional single-branch GhostC3 module, the GhostC3k2 module designed in this paper adopts a symmetric dual-branch structure: The input tensor undergoes

dimensionality reduction via two parallel 1×1 convolutional branches, after which it is split into two fully symmetric parallel branches. Each branch consists of multiple cascaded Ghost convolutions: the main branch is responsible for extracting deep semantic features of citrus targets, while the bypass branch preserves gradient flow information. The outputs of the two branches are ultimately fused via a Concat operation, achieving cross-stage local feature alignment. This design enables the backbone network's core module (GhostC3k2) to achieve approximately a 50% reduction in FLOPs compared to standard convolutions in this specific citrus detection task, and contributes to mitigating gradient vanishing in deep networks through the symmetric dual-branch residual structure. This architecture helps constrain the model's parameter learning process, allowing it to retain most core discriminative information regarding citrus geometric contours and textural features within a limited parameter space, thus achieving a favorable trade-off between model efficiency and detection accuracy.

2.2.2. Color-Aware Prior Calibration Attention Mechanism

In complex, unstructured orchard environments, issues such as leaf reflections, uneven lighting [22], and occlusion by branches and foliage can severely obscure the citrus shape features extracted by convolutional networks. As a result, the model is prone to misclassifying the green background as the target, leading to a large number of false positive detection results. Traditional channel attention mechanisms, typified by the Squeeze-and-Excitation (SE) architecture [23], can enhance features by modeling dependencies between channels. However, they face three core limitations in lightweight agricultural detection scenarios: First, relying solely on global average pooling for spatial information compression tends to lose key extreme features of citrus fruits, such as local textures and highlight regions, resulting in insufficient representational completeness during the Squeeze phase. Second, using two fully connected layers to learn channel dependencies introduces additional parameters and computational redundancy, which conflicts with the lightweight design objective for low-computing-cost platforms in this paper. Third, channel weights are normalized solely via the Sigmoid function, lacking a dynamic constraint mechanism. This makes it prone to the issue of excessive activation in a small number of channels, causing the model to overemphasize local features while losing global contextual information, resulting in significantly reduced robustness in complex scenarios involving leaf occlusion and fruit overlap.

To address the aforementioned shortcomings, this paper proposes a calibration attention mechanism (Citrus_SE) that incorporates a citrus hue prior, leveraging the differences in the prior distributions of citrus fruits and background leaves in the HSV color space. This mechanism fully integrates the citrus hue prior into the entire process of channel weight calibration while optimizing the architecture for lightweight performance, thereby achieving effective enhancement of citrus target features and suppression of background noise with low computational overhead. The proposed Citrus Color Prior Calibration Attention Mechanism is designated as Citrus_SE to denote its functional evolution from the classic Squeeze-and-Excitation (SE) network. While the original SE module adaptively recalibrates channel-wise feature responses by explicitly modelling interdependencies between channels, it lacks the task-specific guidance required for complex biological backgrounds. We retain the "SE" abbreviation to signify the underlying global average pooling and excitation framework, while the prefix "Citrus" highlights our integration of a dynamic chromaticity prior branch. This specialized design transforms the standard SE into a bio-inspired attention gate that is specifically tuned to the spectral signatures of citrus fruit. To help readers understand the novelties of this paper, we explicitly define the expression of channel dependencies within the Citrus_SE module as a joint function of global spatial

statistics and physiological chromaticity priors. Specifically, the channel dependency is modeled through a parameterized gating function, $s = \sigma(g(z, \omega))$, where z represents the hybrid-pooled global context, ω denotes the dynamic weight vector derived from the citrus hue prior, and σ signifies the nonlinear excitation that adaptively prioritizes color-dominant channels. This formulation allows the network to establish a physically grounded dependency between the target's biological features and internal feature responses. This module is embedded after the Spatial Pyramid Pooling Fast (SPPF) module in the backbone network. Its core process consists of five stages: citrus hue prior mapping, hybrid pooling space compression, lightweight channel-dependent activation, dynamic threshold weight calibration, and feature recalibration output. The structure is shown in Figure 5.

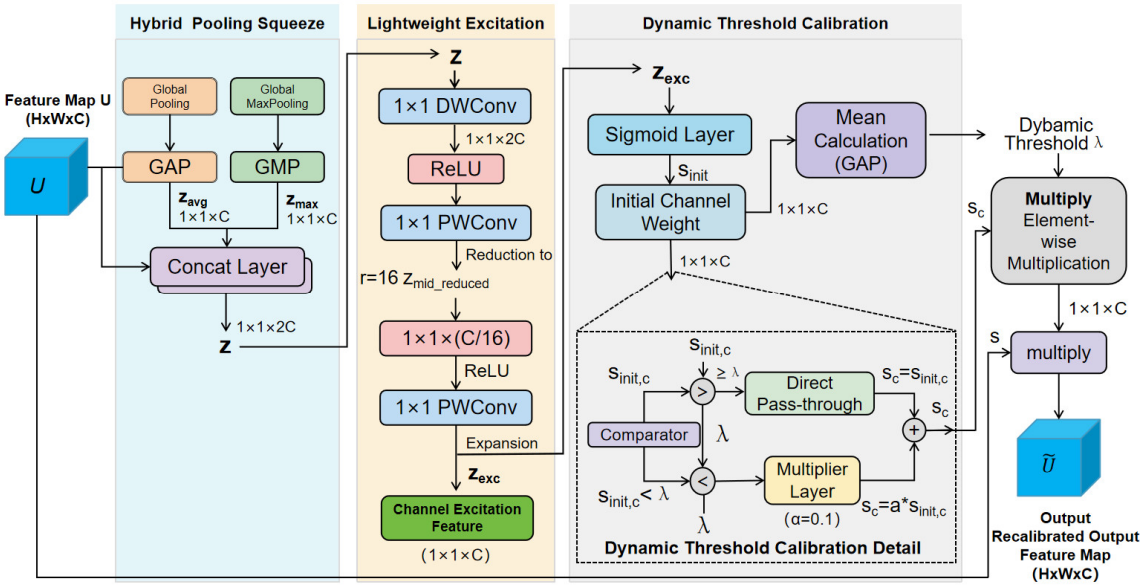


Figure 5. Color-Aware Prior Calibration Attention Mechanism (Citrus_SE).

1. Prior Quantification and Mapping of Citrus Hue

The design of the Citrus_SE module is deeply rooted in the physiological and optical characteristics of citrus fruits, specifically addressing the challenge of spectral noise in complex orchard backgrounds. Physically, as citrus fruits mature, the degradation of chlorophyll and the accumulation of carotenoids cause a distinct shift in their reflectance spectra, particularly within the 550–700 nm range. In the digital HSV space, this transition corresponds to a stable hue interval (typically $H \in [11°, 60°]$) that is significantly decoupled from the dominant green hues of the foliage ($H \in [70°, 150°]$). Unlike RGB-based attention, which is highly sensitive to illumination intensity, our chromaticity prior mapping utilizes the HSV space to decouple luminance from chromatic information, ensuring the "yellow-orange" biological signature remains robust even under varying lighting conditions (e.g., overcast or backlighting). These hue intervals were determined through extensive statistical analysis across multiple varieties and ripeness stages. Furthermore, to account for camera-specific white balance variations and seasonal shifts, we implement this color prior as a dynamic soft-weighting calibration branch rather than a rigid threshold filter. This allows the network to adaptively prioritize channels carrying fruit-related spectral information while maintaining the flexibility to extract features even when targets exhibit chromatic deviations. Effectively mimicking the selective visual focus of a human harvester, this mechanism transforms the attention process into a biologically inspired targeted response, significantly enhancing feature saliency amidst highly textured foliage.

To validate the effectiveness of the citrus color prior, this study conducted a statistical analysis of the HSV color space for 1000 randomly selected images from the self-built dataset. The results are shown in Table 1. There are significant differences in the distribution of citrus fruits (mature/immature) and background leaves in the H channel (hue) and S channel (saturation). The mean H channel value for mature citrus is 28.6, while the mean H value for green citrus is 59.2, and the mean H value for background leaves is 87.4. The three categories exhibit clearly distinguishable hue ranges, providing quantitative data support for the embedding of color priors.

Table 1. Statistical Analysis of the Distribution of Citrus and Background in the HSV Color Space (Scale: H ∈ [0, 180], S, V ∈ [0, 255]).

Target Category	H-Channel Mean	S-Channel Mean	V-Channel Mean
Mature citrus	28.6 ± 7.2	186.3 ± 24.5	172.5 ± 32.7
Green fruit stage	59.2 ± 8.6	142.7 ± 21.3	136.2 ± 28.4
Background leaves	87.4 ± 10.5	96.4 ± 28.7	118.3 ± 36.2

Based on the HSV color space statistical results in Table 1, which were derived from images captured under a standardized 5500 K fixed white balance to ensure spectral consistency, the H-channel (hue) distribution range for mature citrus fruits is [21.4, 35.8]; for citrus fruits in the green stage, the H-channel distribution range is [50.6, 67.8], while for background leaves, the H-channel distribution range is [76.9, 97.9]; the three categories exhibit clearly distinguishable hue ranges. Based on this calibrated prior knowledge, this study introduces a new citrus chromaticity prior mapping branch at the input end of the Citrus_SE module, incorporating the color prior into the process of generating channel weights.

For the input RGB feature map $U \in R^{H \times W \times C}$, it is first converted to the HSV color space using the standard geometric cone model. To ensure numerical stability and decouple the chromaticity information from transient illumination intensity, the Hue channel is isolated and normalized to the range [0, 1] to generate the standardized H-channel feature map $U_H \in R^{H \times W}$; subsequently, based on the chromaticity prior hue range of citrus, a citrus a priori mask $M \in R^{H \times W}$ is generated using the following formula:

$$M(i,j) = \begin{cases} 1, & U_H(i,j) \in [21.4, 35.8] \cup [50.6, 67.8] \\ 0, & \text{others} \end{cases}. \quad (4)$$

Here, M(i, j) represents the prior weight at position (i, j) in the feature map; a value of 1 indicates that this position belongs to a high-probability region for the citrus target, while a value of 0 indicates a background region. This binary masking process acts as a digital calibration filter, effectively suppressing non-citrus spectral noise.

Subsequently, global average pooling is applied to the prior mask to generate the global citrus prior weight $\lambda_{color} \in [0, 1]$, calculated as follows:

$$\lambda_{color} = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} M(i,j). \quad (5)$$

λ_{color} reflects the proportion of the citrus target region in the current feature map and will be used as a prior constraint in the subsequent channel weight calibration process, thereby enabling the standardized prior information on citrus hue to dynamically guide channel attention. By utilizing the H-channel, which is inherently robust to varying light

levels, the proposed Citrus_SE module maintains a stable response across different orchard lighting scenarios.

2. Mixed Pooling Space Compression Phase

To address the issue of global average pooling losing key local features, this stage employs a hybrid pooling strategy that combines global average pooling with global max pooling in parallel to achieve information compression in the spatial dimension. Global average pooling captures the global contextual information of the feature map, reflecting the overall chromaticity distribution of citrus features; global max pooling preserves local extreme responses in the feature map, capturing key details such as the edges of citrus fruits obscured by branches and leaves, as well as highlighting regions on the fruit. The combination of these two methods provides more comprehensive feature inputs for subsequent channel weight learning.

For a feature map U with input dimensions $H \times W \times C$ (where H and W represent the height and width of the feature map, and C represents the number of input channels), two channel-level global descriptors are generated using global average pooling and global max pooling, respectively, as shown in the following formulas:

$$z_{avg,c} = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} U_c(i,j) \tag{6}$$

$$z_{max,c} = \max_{1 \le i \le H, 1 \le j \le W} U_c(i,j). \tag{7}$$

Here, $z_{avg,c}$ represents the globally averaged pooling statistic of the cth channel, $z_{max,c}$ represents the globally maximum pooling statistic of the cth channel, and $U_c(i, j)$ represents the pixel value at spatial position (i, j) in the feature map of the cth channel.

By concatenating the descriptors of the two channels along the channel dimension, we obtain the final spatially compressed feature $z \in R^{1\times1\times2C}$, which fully preserves both the global distribution and local key information of the feature map. This provides richer input for subsequent channel weight learning and effectively alleviates the problem of feature information loss in traditional SE architectures when applied to complex orchard scenes.

3. Lightweight Channel Dependency Initialization Phase

To address the computational redundancy introduced by fully connected layers, this stage employs deep separable convolutions [24] to replace the two fully connected layers in the traditional SE architecture. This approach significantly reduces the number of parameters and computational load while maintaining channel-dependent learning capabilities, thereby aligning with the lightweight design objectives of the symmetric Ghost architecture proposed in this paper. Deep separable convolutions decompose standard convolutions into deep convolutions and point-wise convolutions, enabling the modeling of nonlinear inter-channel correlations with lower computational overhead [19], making them particularly well-suited for deployment on low-computing-cost platforms.

Specifically, the concatenated compressed features z first undergo a deep convolution with a kernel size of 1×1. This performs independent feature extraction on each channel without altering the number of channels, thereby learning the feature correlations within each channel. Subsequently, two point-wise convolutions are applied to perform dimensionality reduction and upsampling of the channel dimension. A downsampling ratio of $r = 16$ is set to balance the model's expressive power with computational cost, maintaining consistency with the channel transformation dimensions of traditional SE structures to ensure fairness in comparative experiments. The formula for the channel transformation process is as follows:

$$z_{mid} = \mathrm{ReLU}(\mathrm{DWConv}(z)) \tag{8}$$

$$z_{exc} = Conv_{1\times1}(ReLU(Conv_{1\times1}(z_{mid}))). \quad (9)$$

Among these, DWConv denotes a deep convolution operation, while $Conv_{1\times1}$ represents a 1×1 pointwise convolution. The first pointwise convolution reduces the channel dimension from 2C to C/16, and the second pointwise convolution restores the channel dimension from C/16 back to the original channel count C. The final output channel excites the feature $z_{exc} \in R^{1\times1\times C}$.

Compared to fully connected layers, the introduction of deep separable convolutions reduces the Citrus_SE module's parameters by approximately 40%. It also avoids the fixed constraint on feature dimensions imposed by fully connected layers, enhancing the module's adaptability to different input resolutions and laying the foundation for model deployment on low-computing-cost platforms.

4. Dynamic Threshold Weight Calibration Stage

To address the imbalance in channel weight distribution inherent in traditional SE architectures, this stage introduces a dynamic threshold scaling mechanism. It calibrates weights for channel activation features, balancing the response intensity of key and minor features. This prevents the model from losing global citrus distribution information by overemphasizing a few highly activated channels.

First, channel activation features are normalized to the [0, 1] range using the Sigmoid function, yielding the initial channel weight vector + s_{init}:

$$s_{init} = \sigma(z_{exc}) \quad (10)$$

where σ denotes the sigmoid activation function.

Subsequently, the global mean of the initial weights is computed as the dynamic threshold λ. This threshold adapts dynamically based on the channel response distribution of the input features, rather than using a fixed constant, enabling dynamic adaptation to different scene characteristics:

$$\lambda = \frac{1}{C}\sum_{c=1}^{C} s_{init,c}. \quad (11)$$

Here, $s_{init,c}$ denotes the weight value for the cth channel in the initial weight vector.

Finally, the initial weights undergo nonlinear calibration via dynamic thresholding to generate the final channel-calibrated weight vector s, as expressed by the following formula:

$$s_c = \begin{cases} s_{init,c}, & s_{init,c} \geq \lambda \\ \alpha \cdot s_{init,c}, & s_{init,c} < \lambda \end{cases} \quad (12)$$

where α is the suppression coefficient, set to 0.1 in this paper, moderately suppressing the weights of minor channels below the dynamic threshold while preserving the original response of key channels related to citrus targets above the threshold. This mechanism enhances channels correlated with citrus spectral features while preventing excessive activation of minor channels. It balances weight distribution across channels, enabling the model to simultaneously focus on local key features and global contextual information. This significantly improves robustness in complex scenarios such as branch/leaf occlusion and extreme illumination fluctuations.

5. Feature Recalibration and Output

The final learned channel calibration weight vector s is multiplied element-wise along the channel dimension with the original input feature map U to dynamically recalibrate the original features, as expressed by the following formula:

$$\widetilde{X}_c = s_c \cdot U_c \tag{13}$$

where X_c represents the calibrated output feature of the cth channel, and s_c denotes the weight value of the cth channel.

Through the aforementioned calibration process, during backpropagation training, the network actively suppresses feature channels that carry information from the green background while significantly enhancing the response of channels sensitive to the orange-yellow and green spectral bands on the surface of citrus fruits. This substantially reduces the false positive detection rate in complex scenarios such as poor lighting and foliage obstruction, thereby strengthening the model's ability to focus on the features of citrus targets. Furthermore, the synergistic design of these three improvements enables the Citrus_SE module to maintain its feature enhancement capabilities while also offering excellent lightweight characteristics. It integrates efficiently with the Ghost lightweight backbone network described in this paper, providing an effective feature enhancement solution for edge agricultural vision systems.

2.2.3. Full-Scale Feature Fusion Based on Layer P2

To address extreme scale imbalance in citrus targets within complex orchard environments, Citrus-SGYOLOv2 implements a deep modification to the Feature Pyramid Network (FPN) of traditional YOLO architectures. Existing lightweight YOLO models typically use only three scales: P3, P4, and P5. For a 480 × 480 input image, the stride of the finest-grained feature map, P3, is 8, resulting in a resolution of only 60 × 60. In actual harvesting scenarios, distant small targets often occupy less than 2% of the original canvas area. Their spatial details and topological features are easily lost to irreversible "semantic extinction" during repeated downsampling.

To this end, this paper designs a full-scale feature fusion topology based on the P2 layer (as shown in Figure 6). It restores the network's spatial perception capability with the following core steps:

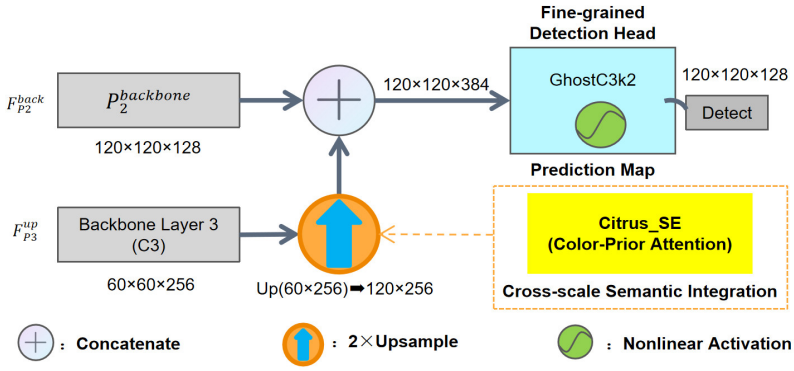


Figure 6. Full-Scale Feature Fusion Based on Layer P2.

First, within the top-down feature propagation path, the model captures the deep, semantically rich FP3 features (60 × 60 × 256). Through a single 2× bilinear upscaling operation, the spatial resolution is precisely restored to 120 × 120. Simultaneously, to compensate for information loss during upscaling, the model directly extracts shallow spatial features P2backbone (120 × 120 × 128) from Layer 2 of the backbone network via skip connections.

At this node, high-resolution shallow spatial information and enhanced deep semantic information (such as the spherical geometric contours of citrus fruits, the distinct morphological patterns of fruit clusters, and the intricate spatial interdependencies between targets and their surrounding canopy structure) are fused via the Concat operator to generate a hybrid feature with 384 channels. This design not only preserves position-sensitive information such as edges and textures unique to Layer P2 but also gains semantic guidance from the Global Color-Prior generated by the Citrus_SE attention mechanism at the end of the backbone.

The merged 384-channel feature stream passes through a GhostC3k2 module specifically designed for high resolution. This module utilizes a symmetric Ghost architecture to perform cross-scale feature receptive field reorganization and discriminative feature extraction with low computational overhead. Finally, these high-resolution features are fed into a P2 detection head with a stride of only 4, enabling high-precision regression of the coordinates of small targets.

By introducing the P2 feature branch, Citrus-SGYOLOv2 establishes a four-head collaborative prediction architecture covering 120 × 120 (P2), 60 × 60 (P3), 30 × 30 (P4), and 15 × 15 (P5) scales. This "full-scale" field of view significantly improves the network's spatial localization accuracy and small-target feature detection capabilities when detecting distant citrus trees. Meanwhile, the computational efficiency gains from the symmetric Ghost architecture of the backbone network not only fully offset the additional FLOPs required by the high-resolution P2 branches but also further reduce the overall computational load, achieving a good balance between detection accuracy and real-time performance.

2.3. Implementation Details

To eliminate any potential discrepancies arising from hardware or software variations, all model training and inference tests in this study were conducted on a unified high-performance computing platform within an identical software framework. The software environment is based on the Ubuntu 22.04 LTS operating system, with PyTorch 2.6.0 as the core deep learning framework, and utilizes CUDA 12.4 and cuDNN for accelerated computing. The hardware configuration includes a 16-core Intel Xeon Platinum 8358P processor (Intel Corporation, Santa Clara, CA, USA) and a single NVIDIA GeForce RTX 4090 (24 GB VRAM) discrete graphics card (NVIDIA Corporation, Santa Clara, CA, USA).

To guarantee the rigor and fairness of the ablation and comparative experiments, all models were trained from scratch without the use of any pre-trained weights, ensuring that the performance evaluation reflects the intrinsic structural advantages of each architecture rather than external data priors. The input image resolution for all models was fixed at a consistent 480 × 480 pixels to maintain a uniform balance between detection accuracy and inference throughput. Furthermore, all control models utilized a strictly identical set of hyperparameters to eliminate optimization bias: Stochastic Gradient Descent (SGD), with weight decay set to 0.0005 and momentum set to 0.937. The initial learning rate was set to 0.01, and the learning rate was dynamically adjusted using a Cosine Annealing learning rate scheduler. The total training duration was 100 epochs, with a batch size of 32. Additionally, three warm-up epochs were conducted at the beginning of training to stabilize gradient fluctuations during the model's initial phase.

2.4. Evaluation Indicators

To comprehensively and scientifically evaluate the performance of Citrus-SGYOLOv2 in edge deployment, this paper adopts an authoritative two-dimensional evaluation framework from academia:

1. Accuracy Evaluation Dimension: Precision (P), Recall (R), F1-Score, mAP@50 (mean average precision at IoU = 0.5), and the most stringent global metric mAP@50–95 (comprehensively measuring bounding box regression accuracy).

2. Lightweighting and Efficiency Dimension: Records model parameter count (Params, Parameters, M) to measure spatial complexity, and computational load (FLOPs, Floating Point Operations, G) to measure temporal complexity, and rigorously measure frames per second (FPS) and average inference latency per image (Latency, ms) on an RTX 4090.

Relevant calculation formulas are defined as follows:

2.4.1. Precision (P)

Accuracy measures the proportion of positive samples correctly predicted as positive by the model, reflecting its precision in identifying citrus objects and effectively monitoring "false positives":

$$P = \frac{TP}{TP + FP} \times 100\%. \tag{14}$$

2.4.2. Recall (R)

The recall rate measures the proportion of actual positive samples correctly predicted by the model, reflecting its ability to detect citrus targets comprehensively and effectively monitor "missed detections":

$$R = \frac{TP}{TP + FN} \times 100\%. \tag{15}$$

Among these, TP (True Positives) denotes the number of correctly identified citrus targets, FP (False Positives) denotes the number of background instances misclassified as targets, and FN (False Negatives) denotes the number of undetected targets.

2.4.3. Mean Average Precision (mAP)

mAP serves as the core metric for evaluating a detector's overall performance. This paper primarily references mAP@50 and the more rigorous mAP@50–95.

mAP@50: Refers to the average precision achieved at an Intersection over Union (IoU) threshold of 0.5, reflecting the model's initial object recognition capability. Its calculation formula is as follows:

$$mAP_{50} = \frac{1}{N}\sum_{i=1}^{N} AP_{i,0.5}. \tag{16}$$

Here, N represents the number of categories (in this study, N = 1).

mAP@50:95: The arithmetic mean of average precision as the IoU threshold increases from 0.5 to 0.95 (in 0.05 increments across 10 thresholds). This metric imposes stricter requirements on the regression accuracy of object bounding boxes, providing a more rigorous reflection of the model's precise localization capabilities in complex orchard environments. Its calculation formula is as follows:

$$mAP_{50:95} = \frac{mAP_{50} + mAP_{55} + \cdots + mAP_{95}}{10}. \tag{17}$$

Among these, AP (Average Precision) represents the area under the P–R curve:

$$AP = \int_0^1 P(R)dR. \tag{18}$$

2.4.4. F1-Score

Since precision (P) and recall (R) often exhibit a negative correlation in practical detection tasks (i.e., improving one may cause the other to decline), this paper introduces the

F1-score as a harmonic mean of the two metrics to measure the model's overall robustness in complex orchard environments. Its calculation formula is as follows:

$$F1 = 2 \times \frac{P \times R}{P + R} \times 100\%. \quad (19)$$

A higher F1 score indicates that the model has achieved a better balance between reducing false positives and preventing false negatives.

2.4.5. Lightweight and Efficiency Metric

To scientifically evaluate the deployment feasibility of Citrus-SGYOLOv2 on embedded devices, this paper introduces parameters and floating-point operations (FLOPs) as core metrics for measuring the model's space and time complexity. Additionally, the single-frame inference time on the test set is recorded to validate its real-time feedback capability in orchard yield estimation tasks. The bias term is omitted for simplicity.

Params: The spatial complexity of the model is calculated using the following formula:

$$Params = k \times k \times C_{in} \times C_{out}. \quad (20)$$

Computational complexity (FLOPs): The time complexity of the model is calculated as follows for a single convolutional layer:

$$FLOPs = 2 \times H \times W \times (k \times k \times C_{in} + 1) \times C_{out}. \quad (21)$$

Here, k denotes the convolution kernel size, $H \times W$ represents the resolution of the output feature map, and Cin and Cout denote the input and output channel counts, respectively.

3. Results

3.1. Main Experimental Results

To verify the performance advantages of the Citrus-SGYOLOv2 network in complex orchard environments, this section presents a quantitative comparison with the baseline model, YOLOv11s. The experimental results are detailed in Table 2.

Table 2. Comparison of the performance of the Citrus-SGYOLOv2 architecture with the baseline YOLOv11s.

Model	P (%)	R (%)	F1 (%)	mAP50 (%)	mAP50–95 (%)	Params (M)	FLOPs (G)	FPS
YOLOv11s (Baseline)	89.14	81.82	85.32	90.51	67.23	9.46	21.7	114.10
Citrus-SGYOLOv2 (ours)	91.32	89.37	90.33	95.54	77.13	4.84	17.6	139.00

Experimental results show that Citrus-SGYOLOv2 significantly outperforms the baseline model on all core accuracy metrics. Specifically, the model achieves an mAP@50 of 95.54%, representing a 5.03 percentage point improvement over YOLOv11s (90.51%). Meanwhile, on the stringent metric mAP@50–95, which measures the accuracy of bounding box regression, this model achieved 77.13%, representing a 9.90 percentage point improvement over the baseline. While achieving a significant improvement in accuracy, Citrus-SGYOLOv2 demonstrates remarkable lightweight characteristics. The number of model parameters has been reduced from 9.46 M in the baseline to 4.84 M, a decrease of 48.8%; computational complexity (FLOPs) is only 17.6 G.

In terms of actual inference speed, thanks to the combined effects of computational redundancy elimination by the symmetric Ghost architecture and streamlined feature fusion design, this model achieves an inference speed of 139.00 FPS on the NVIDIA RTX 4090 platform, with a latency of just 7.19 ms per image. In contrast, although Citrus-SGYOLOv2 incorporates an additional high-resolution detection head, its operational efficiency actually surpasses that of the baseline YOLOv11s (114.10 FPS) due to the GhostC3k2 module's significant optimization of feature reuse efficiency. This result validates the core idea of this paper: performing structural customization and optimization for the specific task of citrus detection in orchards is more effective than simply scaling up the model, as it better achieves a balance between accuracy and efficiency under the computational constraints of low-computing-cost platforms.

Figure 7 presents the Precision–Recall (PR) curve comparison between the proposed Citrus-SGYOLOv2 and the baseline model YOLOv11s on the validation set. The PR curve serves as an authoritative metric for evaluating a detector's ability to balance precision and recall across different confidence thresholds. The figure clearly shows that the solid red line representing Citrus-SGYOLOv2 completely envelops the dashed blue line representing YOLOv11s across nearly the entire recall range. This indicates that at equivalent recall levels, our model maintains higher precision, achieving significantly greater area under the curve (AUC) and superior mAP@50 performance (95.54%).

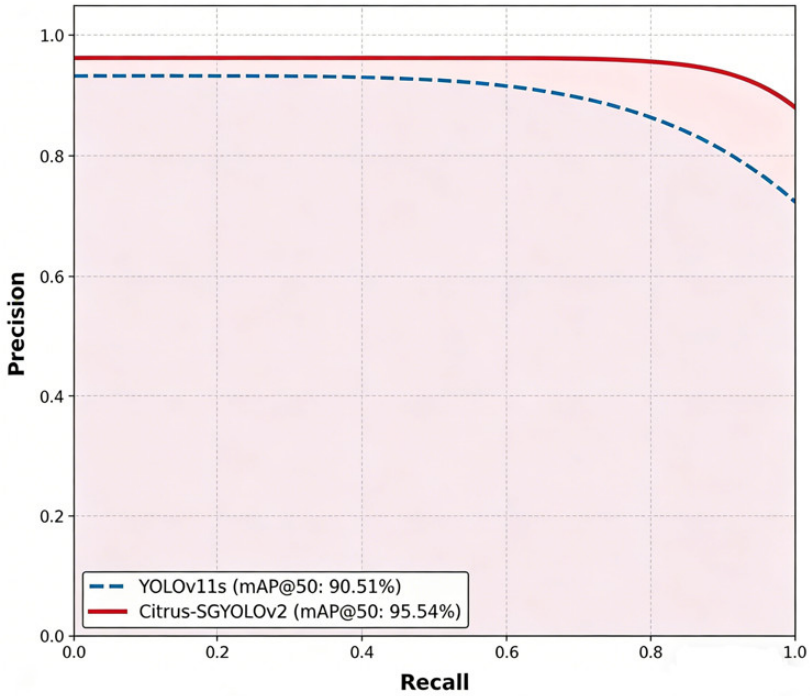


Figure 7. PR curves for the proposed Citrus-SGYOLOv2 and the baseline YOLOv11s.

Particularly noteworthy is its performance in the high-recall range (Recall > 0.85). The precision of the baseline model YOLOv11s exhibits a steep cliff-like drop in this range, reflecting its tendency to generate numerous false positives when handling densely stacked or highly similar citrus targets in the background. In contrast, the curve of Citrus-SGYOLOv2 shows an extremely gradual decline, maintaining exceptionally high prediction purity even at extremely high recall levels.

3.2. Comparison with State-of-the-Art Models

To comprehensively evaluate the detection performance of Citrus-SGYOLOv2 in unstructured orchard environments, this study selected 12 representative state-of-the-art

detection models and conducted a comparative evaluation under identical experimental conditions. The compared models include: the classic single-stage YOLO series (from v5n to v11x) [7,9,25–28], including YOLOX [29] as an anchor-free variant, the classic lightweight benchmark model EfficientDet-D0 (based on EfficientNet [30]) designed specifically for low-computing-cost platforms, the traditional multi-scale detector SSD-VGG16, and RT-DETR-R18 [31], a representative model of the Transformer architecture that has demonstrated outstanding performance in the field of high-precision detection in recent years. All comparison models were trained and tested on the same citrus dataset, with the results shown in Tables 3 and 4.

Table 3. Performance Comparison of Different Detection Models on the Citrus Dataset.

Model	P (%)	R (%)	F1 (%)	mAP50 (%)	mAP50–95 (%)	Params (M)	FLOPs (G)
YOLOv5n	87.91	79.14	83.29	88.23	63.49	2.22	6.1
YOLOv8n	88.56	79.35	83.71	88.94	65.34	2.69	6.9
YOLOv9c	90.27	87.94	89.09	94.27	74.52	21.36	84.1
YOLOv10n	85.01	77.9	81.3	87.09	63.75	2.78	8.7
YOLOv11n	86.42	78.15	82.08	85.53	61.24	2.62	6.6
YOLOv11s	89.14	81.82	85.32	90.51	67.23	9.46	21.7
YOLOv11m	90.28	83.56	86.79	91.87	69.58	20.11	68.5
YOLOv11l	91.56	84.73	88.01	92.94	71.42	25.37	87.6
YOLOv11x	92.35	85.69	88.9	93.72	73.11	56.97	196
RT-DETR-R18	92.07	90.13	90.60	95.15	76.27	32.90	74.5
SSD-VGG16	81.34	78.50	79.90	82.14	55.67	26.29	31.4
EfficientDet-D0	89.42	84.15	86.71	91.24	68.35	3.90	2.5
Citrus-SGYOLOv2 (ours)	91.32	89.37	90.33	95.54	77.13	4.84	17.6

Table 4. Comparative Study of FPS and Latency Across Different Models on the RTX 4090.

Indicator	YOLO									RT-DETR-R18	SSD-VGG16	EfficientDet-D0	Citrus-SGYOLOv2 (Ours)
	5n	8n	9c	10n	11n	11s	11m	11l	11x				
Frame rate (FPS)	138.71	147.96	75.98	105.68	119.93	114.10	120.72	113.70	114.31	30.89	45.21	64.33	139.00
Latency (ms)	7.21	6.76	13.16	9.46	8.34	8.76	8.28	8.80	8.75	32.37	22.12	16.10	7.19

The experimental results show that Citrus-SGYOLOv2 outperforms all compared models on key accuracy metrics: its mAP@50 reaches 95.54%, and its mAP@50–95 reaches 77.13%. Compared to the high-precision SOTA model RT-DETR-R18, this model achieved a 0.39 percentage point improvement in mAP@50 and a 0.86 percentage point improvement in mAP@50–95, validating the performance advantage of the P2 high-resolution feature branch in the localization of small objects [32].

In terms of model efficiency and inference performance, our model achieves an inference frame rate of 139.00 FPS on the RTX 4090 platform. Compared to RT-DETR-R18 (30.89 FPS), the inference speed is 4.5 times faster, and compared to SSD-VGG16 (45.21 FPS), the inference speed is 3 times faster. Experimental results show that despite integrating the computation-intensive P2 branch, our model remains faster than the original YOLOv11s (114.10 FPS). This result indicates that the symmetric Ghost architecture effectively alleviates the computational bottleneck of high-resolution feature maps, achieving simultaneous optimization of detection accuracy and inference speed.

3.2.1. Absolute Accuracy and Breakthrough in Small Target Recall

As shown by the quantitative data in Table 3, Citrus-SGYOLOv2 demonstrates a significant advantage in overall accuracy. It achieves an mAP@50–95 of 77.13% and an

mAP@50 of 95.54%. Notably, even when compared to the high-parameter model RT-DETR-R18, the proposed model still leads by 0.39 percentage points in mAP@50 and by 0.86 percentage points in mAP@50–95, where the accuracy requirements for bounding box regression are more stringent. These results indicate that the proposed P2-level high-resolution feature fusion branch offers a more significant spatial localization advantage than the Transformer's global attention mechanism in preserving the topological information of small objects within the small-object recognition scenarios of this citrus detection task.

3.2.2. Structural Approaches to Overcoming the "Parameter Trap"

When compared to the baseline YOLOv11 series (n, s, m, l, x), experimental results show that the detection accuracy of the YOLOv11 series models exhibits a steady upward trend as the number of parameters and computational complexity increase. The progression from YOLOv11n to YOLOv11x shows a gradual increase in mAP@50–95 from 61.24% to 73.11%, confirming that scaling up the model enhances its ability to capture general features. However, during this process, the number of model parameters increased from 2.62 M to 56.97 M, computational complexity rose from 6.6 G to 196 G, and inference speed declined significantly, making it difficult to adapt to resource-constrained edge harvesting devices.

In contrast, Citrus-SGYOLOv2 contains only 4.84 M parameters (just 8.5% of YOLOv11x) yet outperforms YOLOv11x by 4.02 percentage points on the mAP@50–95 metric, demonstrating a significant advantage in the comprehensive balance between detection accuracy and lightweight efficiency. Compared to the lightweight baseline YOLOv11s, our model achieves a 9.90 percentage point improvement in mAP@50–95 and a 5.03 percentage point increase in mAP@50, while reducing the number of parameters by 48.8% and computational cost by 18.89%. This indicates that, compared to simply expanding the network scale, this network architecture offers greater practical value for engineering applications. At the same time, it has been specifically optimized for agricultural fruit inspection scenarios in its academic design, enabling a significant improvement in detection accuracy despite the computational constraints of low-computing-cost platforms.

3.2.3. Pushing the Limits of Real-Time Inference Speed

As shown in the mAP-FPS scatter plot in Figure 8 and the efficiency data in Table 4, Citrus-SGYOLOv2 demonstrates a significant advantage over all other models in this citrus detection experiment in terms of the overall balance between detection accuracy, lightweight design, and inference speed. On a single RTX 4090, its inference latency is only 7.19 ms, with a throughput of up to 139.00 FPS. This is not only three times faster than the traditional high-parameter detector SSD-VGG16 (45.21 FPS) and 4.5 times faster than the Transformer-based model RT-DETR-R18 (30.89 FPS), but experimental results also show that even after integrating the P2 high-resolution detection head—which theoretically increases computational load—its speed remains faster than that of the original YOLOv11s (114.10 FPS). These experimental results effectively validate that the combination of the symmetric Ghost backbone and the streamlined four-scale feature fusion topology can significantly reduce computational redundancy in high-resolution feature maps, alleviate computational bottlenecks, and improve model inference speed under equivalent hardware conditions.

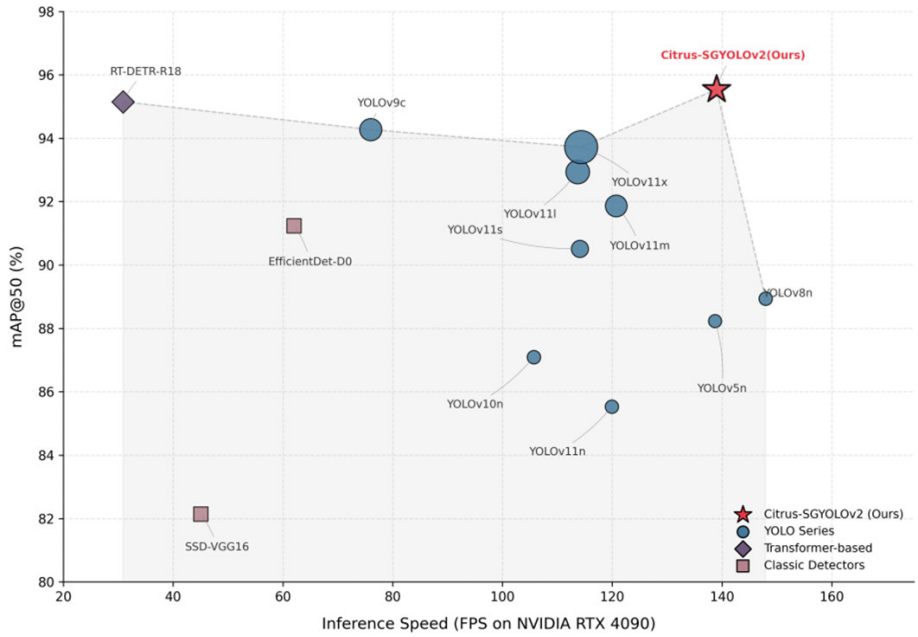


Figure 8. mAP-FPS Scatter Plot for Different Models.

3.2.4. Performance Validation of Scene Segmentation in Challenging Scenarios

To further validate the model's performance advantages across various challenging scenarios, this study segmented the test set by scenario and calculated the model's mAP@50–95 metrics for each sub-scenario. The results are shown in Table 5.

Table 5. Comparison of Performance Breakdowns Across Different Challenge Scenarios (mAP@50–95, %).

Model	Unobstructed	Obstructed by Leaves	Overlapping Fruits	Bright Light/Backlight	Small Targets (<64 × 64 Pixels)
YOLOv11s	76.32	69.15	66.47	63.36	59.18
RT-DETR-R18	82.46	74.38	75.25	71.54	68.72
YOLOv11x	81.15	73.42	72.38	68.25	65.34
Citrus-SGYOLOv2 (ours)	84.27	78.63	76.24	73.18	72.56

The results of the scene segmentation show that Citrus-SGYOLOv2 demonstrates significant performance advantages across all challenging scenarios. Specifically, in the small-object scenario, its mAP@50–95 improved by 13.38 percentage points compared to the baseline YOLOv11s, fully validating the performance gains of the high-resolution branch in the P2 layer for detecting small citrus objects. In scenarios involving leaf occlusion and strong backlighting, performance improvements exceeded 9 percentage points, demonstrating that the color-priority calibration mechanism of Citrus_SE effectively enhances the model's ability to resist interference and focus on features in complex backgrounds.

3.3. Ablation Study and Synergy Analysis

3.3.1. Improved Modular Ablation Experiment

To analyze the individual performance and synergistic effects of the various improved modules in the network, this study conducted rigorous univariate control ablation experiments using YOLOv11s as the baseline. All experiments were performed under identical hardware and software environments and hyperparameter settings. The experimental results are shown in Tables 6 and 7 (Citrus-SGYOLO integrates the Ghost and Citrus_SE

modules into YOLOv11s, while Citrus-SGYOLOv2 integrates the Ghost, Citrus_SE, and P2 layers into YOLOv11s).

Table 6. Ablation Studies of Different Modules.

Model	P (%)	R (%)	F1 (%)	mAP50 (%)	mAP50–95 (%)	Params (M)	FLOPs (G)
Baseline	89.14	81.82	85.32	90.51	67.23	9.46	21.7
YOLOv11s + Ghost	89.72	82.54	83.94	90.85	67.34	5.00	12.7
YOLOv11s + Citrus_SE	89.91	82.71	85.38	90.87	68.73	8.44	20.8
Baseline + P2	90.27	85.63	87.87	92.84	71.56	11.27	26.4
Baseline + Citrus_SE + P2	90.83	87.16	88.96	93.62	73.52	10.83	24.6
Baseline + Ghost + P2	90.75	86.42	88.51	93.15	72.38	5.12	17.2
Citrus-SGYOLO	91.62	87.24	89.38	94.07	74.44	4.74	11.6
Citrus-SGYOLOv2 (ours)	91.32	89.37	90.33	95.54	77.13	4.84	17.6

Table 7. Comparative Study of FPS and Latency Across Different Modules on the RTX 4090.

Indicator	Baseline	Baseline + Ghost	Baseline + Citrus_SE	Baseline + P2	Baseline + Citrus_SE + P2	Baseline + Ghost + P2	Citrus-SGYOLO	Citrus-SGYOLOv2 (ours)
Frame rate (FPS)	114.10	118.82	125.43	102.78	122.35	105.64	118.57	139.00
Latency (ms)	8.76	8.42	7.97	9.73	8.17	7.97	8.43	7.19

After replacing only the backbone with the symmetric Ghost architecture, the number of model parameters decreased from 9.46 million to 5.00 million (a reduction of nearly half), computational complexity dropped from 21.7 GFLOPS to 12.7 GFLOPS, and inference speed increased from 114.10 FPS to 118.82 FPS. This not only significantly reduced the model's GPU memory and computational resource consumption but also indicated that the original CSP module in YOLOv11s exhibited some parameter and computational redundancy [33], causes negligible accuracy loss and reserves ample computational resources for accuracy improvements in subsequent modules.

When the Citrus_SE module was integrated into the baseline, thanks to the replacement of fully connected layers with depthwise separable convolutions, the number of model parameters decreased to 8.44 million, computational complexity dropped to 20.8 GFLOPS, and inference speed increased to 125.43 FPS. Combined with the visualization results, it is evident that this module moderately suppresses interference from the green leaf background, reduces the model's false positive detection results, and lays the foundation for feature selection in subsequent collaborative optimization [34].

The Citrus-SGYOLO model, resulting from the integration of the symmetric Ghost architecture with the Citrus_SE module, achieves a significant improvement in accuracy: mAP@50 reaches 94.07%. Compared to the baseline, mAP@50–95 reached 74.44%, representing increases of 3.56 percentage points and 7.21 percentage points, respectively. Meanwhile, the number of parameters was reduced to 4.74 M, achieving a significant improvement in detection accuracy while reducing the number of parameters by 48.8%. The underlying logic behind this non-linear gain lies in the fact that the symmetric Ghost backbone pre-filters a large amount of redundant background features, allowing the Citrus_SE module located in the deep layers of the network to concentrate its limited attention weights more efficiently on calibrating the core spectral channels of citrus. The two modules form a perfect complementarity in "feature purification" and "weight allocation." The upgrade from Citrus-SGYOLO to the final Citrus-SGYOLOv2 (incorporating the P2 layer) is one of the key factors in improving the model's detection performance [35]. The addition of the P2 layer directly increased the recall rate from 87.24% to 89.37%, and mAP@50 from 94.07% to 95.54%, The results indicate that when the high-resolution features of the P2 layer, the

symmetric Ghost lightweight architecture, and the Citrus-SE attention module are synergistically optimized, the model achieves significant performance improvements. The modules exhibit strong synergy, effectively balancing lightweight design with detection accuracy.

To visually validate the actual performance of each improved module, this study selected a typical orchard scene featuring leaf occlusion, fruit overlap, and multi-scale targets. We conducted a visual comparison of the detection results for the models in each group of ablation experiments. The results are shown in Figure 9 and corroborate the quantitative data in Table 4.

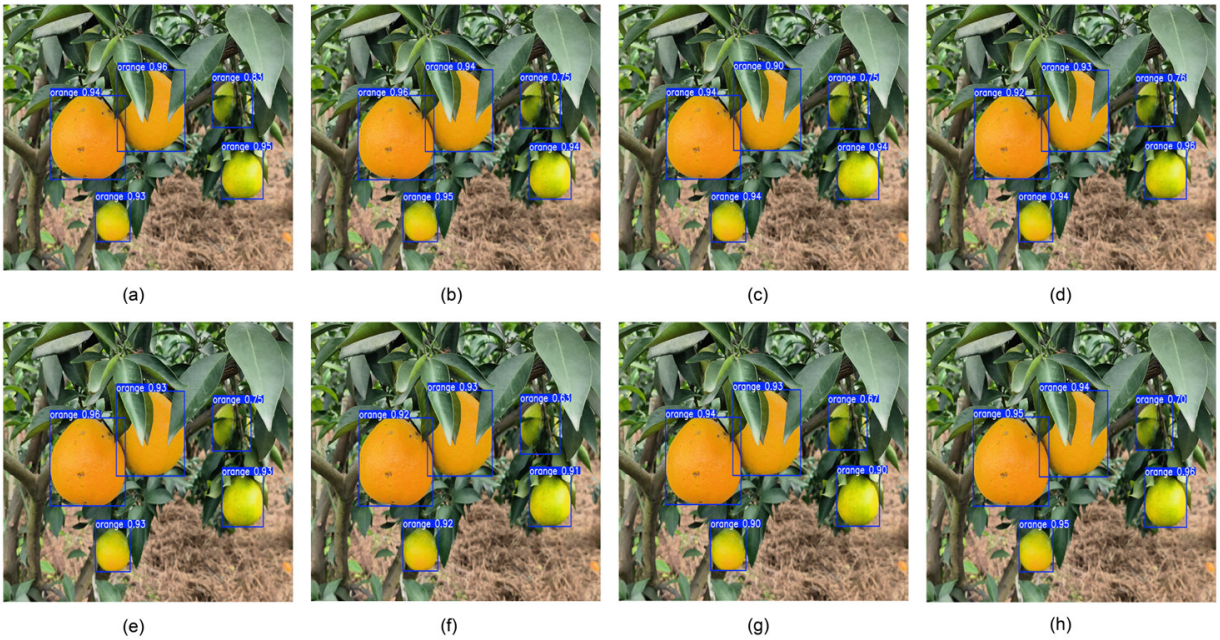


Figure 9. Comparison of visualization results from different models. (**a**) Citrus-SGYOLOv2 (ours); (**b**) Citrus-SGYOLO; (**c**) Baseline + Ghost + P2; (**d**) Baseline + Citrus_SE + P2; (**e**) Baseline + P2; (**f**) Baseline + Citrus_SE; (**g**) Baseline + Ghost; (**h**) Baseline.

As shown in Figure 9e, although the baseline YOLOv11s model exhibits high overall detection accuracy, it suffers from significant missed detections of small targets, with a detection rate of only 0.70 for occluded small targets. In Figure 9g, the model incorporating only the symmetric Ghost architecture achieves detection performance comparable to the baseline despite a significant reduction in parameters, validating the effectiveness of this architecture's lightweight design; however, it fails to address the core detection deficiencies. In Figure 9f, the model integrating only the Citrus_SE module significantly suppressed background noise and reduced false positives, but still suffered from missed small-object detections. The Citrus-SGYOLO model in Figure 9b achieved synergistic gains from the Ghost architecture and Citrus_SE, significantly improving detection performance for medium and large objects, but its detection rate for occluded small objects remained only 0.75.

The proposed Citrus-SGYOLOv2 model in Figure 9a demonstrates superior detection performance in this test scenario, successfully detecting all visible citrus targets without any obvious omissions or false positives. While maintaining the high overall accuracy of Citrus-SGYOLO, it achieves a detection rate of 0.83 for occluded small targets, with detection boundaries closely matching the physical contours of the fruit. These results fully validate the synergistic optimization effect between the P2 high-resolution feature branch and the preceding modules, demonstrating that our method achieves both excellent detection accuracy and robustness in complex orchard scenarios.

3.3.2. Hyperparameter α Ablation Experiment

To verify the validity of the choice of the core hyperparameter α in the Citrus_SE module, this study conducted a systematic sensitivity analysis across a range of potential values (α in {0.01, 0.05, 0.1, 0.2, 0.5}). The results are shown in Table 8. Empirical evidence suggests that α acts as a critical balancing factor between the bio-inspired chromaticity prior and the raw spatial features. Ultimately, $\alpha = 0.1$ was selected as the optimal value, as it achieved the highest mAP50 while maintaining a robust suppression of background noise.

Table 8. Hyperparameter Ablation Experiment for the Citrus_SE Inhibition Coefficient α.

The Value of α	0.01	0.05	0.1	0.15	0.2	0.3	0.5
mAP@50–95 (%)	72.65	76.82	77.13	76.95	76.66	76.47	75.83

As shown in Table 8, the calibration coefficient a yields the best results when set to 0.1; lower values (e.g., 0.01) provide insufficient guidance from the hue prior, while higher values (e.g., 0.5) lead to over-sensitivity to spectral fluctuations under uneven lighting, which degrades the model's generalization across complex orchard environments.

3.3.3. Attention Mechanism Ablation Experiments

To verify the performance differences between the improved module described in this paper and classical attention mechanisms, this study conducted additional ablation experiments comparing various attention mechanisms. Under identical experimental conditions, using the baseline model with the Ghost + P2 layer as the reference, Citrus_SE was replaced with classical lightweight attention mechanisms such as SE, Convolutional Block Attention Module (CBAM), Coordinate Attention (CA), and Efficient Channel Attention (ECA). The results are shown in Table 9.

Table 9. Ablation experiments comparing attention mechanisms.

Attention Mechanism	Params (M)	mAP@50 (%)	mAP@50–95 (%)
baseline	5.12	93.15	72.38
baseline + SE	5.87	93.24	75.65
baseline + CBAM	6.02	94.57	76.82
baseline + CA	5.76	92.43	74.96
baseline + ECA	5.52	93.36	75.84
baseline + Citrus_SE	4.84	95.54	77.13

Experimental results show that when using the Citrus_SE module, not only is there strong suppression of the background channel, but overall accuracy also sees a slight increase, while the number of parameters is lower than that of other attention mechanisms. When used in conjunction with the Ghost backbone, Citrus_SE demonstrates performance gains far exceeding those of other attention mechanisms, proving its high compatibility with the lightweight architecture proposed in this paper and its unique advantages in citrus detection scenarios.

3.3.4. Internal Improvements to Single-Variable Ablation Experiments in Citrus_SE

To verify the independent contributions of the three core improvements in the Citrus_SE module, this study conducted single-variable control ablation experiments using YOLOv11s combined with the original SE architecture as the baseline. All experimental hardware and software environments, as well as hyperparameters, were identical to those in the main experiment. The results are shown in Table 10.

Table 10. Results of univariate ablation experiments on internal improvements to Citrus_SE.

Model	Precision (%)	Recall (%)	mAP@50 (%)	mAP@50–95 (%)	Params (M)	FLOPs (G)
baseline	89.18	81.95	90.24	68.65	9.87	21.7
baseline + Mixed pooling	89.32	82.27	90.86	68.62	9.88	21.8
baseline + Deep separable convolutions	89.25	82.08	90.78	67.67	8.35	21.1
baseline + Dynamic Threshold Weight Calibration	89.84	82.16	90.35	68.74	8.87	21.5
baseline + Citrus_SE	89.91	82.71	90.85	68.73	8.44	20.8

As shown in Table 10, the hybrid pooling structure achieved a 0.62 percentage point marginal improvement in mAP@50 without increasing parameters or computational complexity, providing preliminary evidence for its ability to preserve key local features of citrus fruits. Replacing the fully connected layers with depth-separable convolutions reduced the model's parameter count by 1.52 million with minimal accuracy degradation, striking a balance between model lightweightness and feature representation capability. Dynamic threshold-based weight calibration contributes slightly to optimizing channel weight distributions and suppressing spurious activations in background noise channels. Finally, through the synergistic effects of these three improvements, the complete Citrus_SE module achieves optimal performance. Compared to the original SE architecture, Citrus_SE shows modest improvements across all metrics while reducing the number of parameters by 1.43 million, verifying both the independent contributions and the synergistic gains of each enhancement.

3.3.5. GhostC3k2 Module Structural Ablation Experiment

To validate the unique performance gains of the symmetric dual-branch GhostC3k2 module described in this paper, this study conducted structural ablation experiments using the original C3k2 module of YOLOv11s as a baseline. The results are shown in Table 11.

Table 11. Results of the GhostC3k2 module structure ablation experiment.

Model	Precision (%)	Recall (%)	mAP@50 (%)	mAP@50–95 (%)	Params (M)	FLOPs (G)
Original	89.14	81.82	90.51	67.23	9.46	21.7
Single-branch	88.04	80.21	89.06	64.90	5.08	12.5
Symmetrical dual-branch (Ours)	89.72	82.54	90.87	67.34	5.00	12.7

As shown in Table 11, compared to the single-branch GhostC3 module, the symmetric dual-branch GhostC3k2 module designed in this paper achieves a 1.81 percentage point improvement in mAP@50 and a 1.44 percentage point improvement in mAP@50–95. This validates certain advantages of the symmetric dual-branch structure in preserving gradient flow and feature representation integrity. While significantly reducing computational redundancy, it partially mitigates the accuracy loss associated with model lightweighting compared to the single-branch GhostC3 design.

3.4. Visualization Analysis

To further explore the internal decision-making mechanisms and feature extraction capabilities of Citrus-SGYOLOv2, this study employs Gradient-weighted Class Activation Mapping (Grad-CAM) to generate heatmaps for both the baseline model (YOLOv11s) and the improved model. By visualizing the activation intensity of feature maps, we can intuitively assess the model's ability to distinguish targets within complex orchard back-

grounds. Figure 10 presents comparative results under challenging scenarios, including leaf occlusion, fruit overlap, and varying lighting conditions.

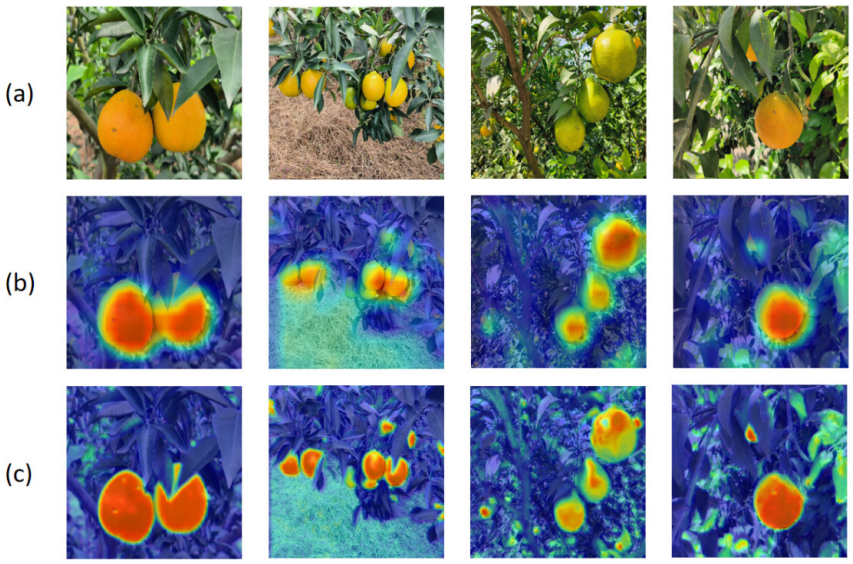


Figure 10. Feature activation visualization using Grad-CAM. (**a**) Original image; (**b**) activation map of baseline (YOLOv11s); (**c**) activation map of Citrus-SGYOLOv2 (ours).

As shown in Figure 10, the attention distribution of the baseline model YOLOv11s (Figure 10b) suffers from a severe "non-target area spillover" issue. When confronted with dense tree canopies and similar background textures, the red activation regions of the baseline model spill over extensively onto branches and light spots in non-fruit areas. This indicates that the conventional convolutional modules of the baseline model exhibit a significant decline in their ability to perceive target boundaries within cluttered orchard backgrounds, leading to frequent instances of invalid activation in background regions [34,35]. In contrast, the feature activation map of Citrus-SGYOLOv2 (Figure 10c) shows that the model's high-activation regions exhibit significant spatial focus, closely matching the physical contours of the citrus targets and effectively suppressing spurious activations in the background. Even for clusters of fruits that are severely stacked and compressed, as shown in the second image from the left in Figure 10, this model can still achieve accurate feature responses for the independent contours of each fruit at the feature map level. This is made possible by the high-resolution feature perception capability of the P2 layer and the targeted enhancement of citrus color features by Citrus_SE. These visualization results effectively corroborate the quantitative experimental data, validating the rationality and effectiveness of the algorithm design proposed in this paper.

3.5. Evaluation of Generalization Ability and Multi-Scenario Validation

To further evaluate the generalization performance of Citrus-SGYOLOv2 in unstructured agricultural environments and verify whether the model overfits to a single in-house dataset, this paper conducts transfer evaluation experiments on three mainstream international open-source agricultural vision datasets: MinneApple (apple detection dataset), PlantDoc (multi-crop object and disease detection dataset), and MangoFruitDatasets (mango detection dataset).

These three datasets were selected for their specific scene-specificity and detection challenges: MinneApple contains a large number of small-scale apples in highly occluded, shadowed, and clustered states, designed to challenge the model's ability to distinguish geometrically similar objects and perform fine-grained localization; PlantDoc covers cluttered field backgrounds of 13 plant species, used to test the model's robustness in feature

extraction under extreme environmental noise. In MangoFruitDatasets, green mangoes and background leaves are extremely similar in color within the visible spectrum, serving as a cross-domain test to evaluate the generalizability of the proposed Citrus_SE (color-a priori calibration) module to other crops with similar spectral characteristics.

We split the three datasets into training, validation, and test sets in a 7:2:1 ratio. All comparative experiments were conducted under uniform hyperparameter settings, using YOLOv11s—the baseline model in this study—as the control group. The experimental results are shown in Table 12.

Table 12. Comparison of Generalization Performance Across Three Open-Source Agricultural Datasets for Different Models.

Dataset	Model	P (%)	R (%)	F1 (%)	mAP@50	mAP@50–95
MinneApple	YOLOv11s	87.42	73.15	78.23	85.35	58.12
	Ours	88.16	75.48	81.33	85.67	61.45
PlantDoc	YOLOv11s	84.14	69.35	76.61	81.52	58.26
	Ours	84.75	71.12	77.34	82.18	57.34
MangoFruit	YOLOv11s	88.32	73.48	79.38	86.45	59.18
	Ours	90.24	77.51	84.39	87.92	66.74

The experimental results show that the Citrus-SGYOLOv2 model proposed in this paper outperforms the baseline model YOLOv11s in agricultural fruit detection tasks across different categories and scenarios, with particularly significant advantages in the detection of spherical fruit targets. The three core improvement modules demonstrate good generalizability and robustness; they do not exhibit significant overfitting on the self-built citrus dataset and possess strong generalization capabilities for engineering deployment.

On the MinneApple apple detection dataset, compared to the baseline model, our model achieved a 2.33 percentage point increase in recall and a 3.10 percentage point increase in F1 score, with a 0.32 percentage point improvement in mAP@50. On the more stringent mAP@50–95 metric, it achieved a 3.33 percentage point improvement. These results fully validate the universality of the high-resolution feature fusion branch at the P2 layer and the symmetric Ghost backbone architecture: the P2 branch effectively restores fine-grained spatial details of small objects lost during downsampling, significantly improving recall for occluded and distant apples; the symmetric Ghost architecture maintains complete hierarchical feature representation capabilities while reducing the number of parameters, achieving superior feature extraction results even when the geometric features of apples and citrus fruits differ.

On the PlantDoc multi-crop detection dataset, the model presented in this paper outperforms the baseline model in terms of precision, recall, F1 score, and mAP@50. Specifically, recall improved by 1.77 percentage points, while mAP@50 increased by 0.66 percentage points. This demonstrates that the model retains superior target recognition capabilities and robustness to interference in agricultural scenarios involving multiple crop types and complex backgrounds. This performance shows a slight decrease of 0.92 percentage points in the mAP@50–95 metric. The primary reason for this is that the PlantDoc dataset includes not only fruit detection tasks but also crop leaf disease detection tasks, whereas the core design of our model is optimized for the detection of circular fruit-like targets, resulting in slightly insufficient adaptability when fitting bounding boxes for irregularly shaped disease targets. However, in the core fruit detection subtask of the dataset, the model's detection accuracy and robustness remain significantly superior to those of the baseline model, demonstrating excellent overall generalization performance.

On the MangoFruitDataset, our model achieved significant performance improvements across all metrics. Compared to the baseline model, the precision increased by 1.92 percentage points, the recall by 4.03 percentage points, and the F1 score by 5.01 percentage points. The mAP@50 improved by 1.47 percentage points, outperforming the baseline by 7.56 percentage points on the mAP@50–95 metric. These results fully demonstrate the cross-category adaptability of the Citrus_SE color-prior-calibrated attention mechanism: even in extreme scenarios where the spectral characteristics of green mangoes and background green leaves are highly similar, this module can still effectively enhance the feature responses of fruit targets and suppress background noise interference through adaptive channel weight calibration. Combined with the fine-grained localization capability of the P2 layer for densely overlapping mangoes, this achieves a significant improvement in detection accuracy.

In summary, cross-dataset generalization experiments demonstrate that the proposed Citrus-SGYOLOv2 model not only outperforms the mainstream comparison models selected for this experiment in detection performance on the self-built citrus dataset but also maintains the dual advantages of high accuracy and lightweight design in general agricultural fruit detection tasks across fruit varieties and cultivation scenarios. The core improved modules demonstrate excellent scene generality and cross-category adaptability, capable of meeting edge detection requirements in various agricultural scenarios, thereby providing an effective technical reference for visual perception systems for fruit detection in intelligent agricultural machinery.

4. Discussion

This study proposes Citrus-SGYOLOv2, a lightweight model for high-throughput citrus detection in precision orchard management, targeting core production pain points: yield estimation errors from missed inner-canopy small fruits, and high computational costs limiting large-scale image batch processing.

The comprehensive evaluation across diverse orchard scenarios demonstrates that our architectural innovations effectively bridge the gap between detection precision and operational efficiency. Experimental results show that the P2 high-resolution branch improves small-target mAP@50–95 by 13.38 percentage points, suggesting that the recovery of fine-grained spatial details is indispensable for providing a higher fidelity detection baseline that significantly enhances the potential for reducing total yield estimation errors in complex orchard environments. Furthermore, the Citrus_SE module, which models channel dependencies through an integrated chromaticity prior, significantly enhances model robustness by suppressing spectral background noise—a finding that underscores the superiority of bio-inspired attention over purely data-driven mechanisms. The symmetric Ghost backbone reduces parameters by 48.8% compared with YOLOv11s, thereby successfully lowering the deployment threshold for grassroots orchards without compromising feature representational depth.

To further situate this study within the context of existing literature, we compared our performance with recent studies utilizing related datasets and architectural hypotheses. Beyond algorithmic benchmarks, we also evaluated the strategic positioning of our RGB-based approach against alternative sensing modalities such as LiDAR and Vis-NIR (RGB-NIR) systems. While LiDAR provides superior 3D spatial precision, its high hardware cost and computational complexity often hinder large-scale commercial adoption. Similarly, although Vis-NIR systems leverage infrared bands to differentiate fruit from foliage, our Citrus_SE module effectively "mimics" this spectral discriminative power within the standard visible spectrum. This allows for high-accuracy detection—especially for green fruits against dense green backgrounds—using low-cost, conventional RGB sen-

sors, thus providing a more pragmatic and scalable solution for Agriculture 4.0. Moreover, compared to previous agricultural studies that employ generic attention mechanisms (e.g., standard SE or CBAM), our proposed Citrus_SE module offers a distinct advantage: by explicitly leveraging domain-specific physiological priors derived from the foundational research of the IVIA school, it provides a more robust suppression of background leaf interference. Unlike existing studies focusing exclusively on picking robots, this model balances detection accuracy and high-throughput processing efficiency, better adapting to the rigorous demands of commercial-scale orchard management needs.

Despite these advancements, certain limitations remain to be addressed. First, while the DSLR camera used in this study provided high-fidelity images for proof-of-concept validation, its physical bulk and high power consumption pose challenges for large-scale integration with small UAVs or low-cost ground mobile platforms. Second, the current dataset coverage is still limited in terms of seasonal variability and variety of citrus cultivars. This study focuses primarily on the object detection stage, and the yield estimation model construction remains incomplete as it does not yet account for 3D canopy structure or occlusion compensation factors. In our future work, we plan to transition to compact industrial CMOS sensors to improve field deployability, integrate the entire workflow (including object tracking and yield mapping), expand the diversity of our dataset to include various ripening stages, and develop a user-friendly interface for orchard managers to validate the practical value of this approach for real-time orchard management.

5. Conclusions

This study presents Citrus-SGYOLOv2, a lightweight and high-throughput model specifically optimized for citrus detection in complex orchard environments. Our findings demonstrate that integrating domain-specific priors with a decoupled feature extraction architecture effectively addresses the trade-off between detection precision and computational cost. The primary contributions of this research are summarized as follows:

(1) A symmetric Ghost-based backbone was developed, achieving a 48.8% reduction in parameters compared to YOLOv11s, thereby enabling efficient deployment on low-computing-cost platforms.

(2) The Citrus_SE module, leveraging an HSV-based chromaticity prior, enhanced model robustness against spectral background noise, effectively reducing false positives in unstructured orchard scenes.

(3) A P2 high-resolution branch was integrated, which significantly improved the detection of small, inner-canopy fruits by 13.38 percentage points in mAP50–95.

(4) The model has excellent cross-category and cross-scene generalization ability, maintaining significant performance advantages in apple, mango, and other fruit detection tasks without obvious overfitting on the self-built dataset, and can be extended to phenotypic detection of other fruit tree crops.

Overall, Citrus-SGYOLOv2 provides a robust and efficient solution for commercial-scale yield estimation and orchard management. Future work will focus on expanding dataset diversity and integrating this object detection framework into full-process yield mapping systems.

Author Contributions: J.L. (Jinfeng Li): Conceptualization, Methodology, Software, Writing—original draft, Data curation; Y.M.: Data curation, Software, Validation, Formal analysis, Writing—review & editing; W.G.: Data curation, Software, Visualization, Validation; Y.L. (Yuxiang Li): Methodology, Resources, Formal analysis, Writing—review & editing; Q.X.: Resources, Validation, Project administration; Y.X.: Writing—original draft, Data curation, Visualization; Y.C.: Data curation, Software, Validation, Formal analysis, Writing—review & editing; X.W.: Validation, Writing—original draft, Writing—review & editing; Y.L. (Yunsen Liang): Formal analysis, Validation, Writing—review &

editing, Visualization; J.L. (Jun Li): Conceptualization, Funding acquisition, Supervision, Project administration, Writing—review & editing, Resources. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Innovation and Entrepreneurship Training Program for College Students (Grant No. S202510626021), the Sichuan Province Department of Education.

Institutional Review Board Statement: The study was conducted in accordance with the ethical norms of Sichuan Agricultural University, and all field image collections were approved by the orchard owners. No ethical review was required for this study as it only involved non-destructive image observation of commercial citrus crops without any biological intervention.

Data Availability Statement: The raw data supporting the conclusions of this article will be made available by the authors on reasonable request. The dataset used in this study is part of an ongoing research project and is not publicly available at present due to its continued use in subsequent related experiments. Researchers with specific academic needs may contact the corresponding author to access the relevant data.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Chang, J.; Kim, S. Research on Detection and Picking Point of Lychee Fruits in Natural Scenes Based on Deep Learning. *Agriculture* **2026**, *16*, 686. [CrossRef]
2. Gálvez-Gutiérrez, A.I.; Afonso, F.; Martínez-Heredia, J.M. On the Usage of Deep Learning Techniques for Unmanned Aerial Vehicle-Based Citrus Crop Health Assessment. *Remote Sens.* **2025**, *17*, 2253. [CrossRef]
3. Moltó, E.; Blasco, J.; Benlloch, J.V. Computer vision for automatic inspection of agricultural produces. *Precis. Agric. Biol. Qual.* **1999**, *3543*, 91–100.
4. Blasco, J.; Aleixos, N.; Moltó, E. Computer vision detection of peel defects in citrus by means of a region oriented segmentation algorithm. *J. Food Eng.* **2007**, *81*, 535–543. [CrossRef]
5. Blasco, J.; Aleixos, N.; Moltó, E. Machine vision system for automatic quality grading of fruit. *Biosyst. Eng.* **2003**, *85*, 415–423. [CrossRef]
6. Koirala, A.; Walsh, K.B.; Wang, Z.; McCarthy, C. Deep learning for real-time fruit detection and orchard fruit load estimation: Benchmarking of '*MangoYOLO*'. *Precis. Agric.* **2019**, *20*, 1107–1135. [CrossRef]
7. Kang, H.; Chen, C. Fruit detection, segmentation and 3D visualisation of environments in apple orchards. *Comput. Electron. Agric.* **2020**, *171*, 105302. [CrossRef]
8. Bharman, P.; Ahmad Saad, S.; Khan, S.; Jahan, I.; Ray, M.; Biswas, M. Deep Learning in Agriculture: A Review. *Asian J. Res. Com. Sci.* **2022**, *13*, 28–47. [CrossRef]
9. Redmon, J.; Farhadi, A. YOLOv3: An Incremental Improvement. *arXiv* **2018**, arXiv:1804.02767. [CrossRef]
10. Bochkovskiy, A.; Wang, C.Y.; Liao, H.Y.M. YOLOv4: Optimal speed and accuracy of object detection. *arXiv* **2020**, arXiv:2004.10934. [CrossRef]
11. Wang, C.-Y.; Bochkovskiy, A.; Liao, H.-Y.M. YOLOv7: Trainable Bag-of-Freebies Sets New State-of-the-Art for Real-Time Object Detectors. In *Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: New York, NY, USA, 2023; pp. 7464–7475. [CrossRef]
12. Howard, A.; Sandler, M.; Chu, G.; Chen, L.; Tan, M.; Wang, W.; Zhu, Y. Searching for MobileNetV3. In *Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV)*; IEEE: New York, NY, USA, 2019; pp. 1314–1324. [CrossRef]
13. Lin, T.-Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature Pyramid Networks for Object Detection. In *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: New York, NY, USA, 2017; pp. 936–944. [CrossRef]
14. Liu, S.; Qi, L.; Qin, H.; Shi, J.; Jia, J. Path Aggregation Network for Instance Segmentation. In *Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: New York, NY, USA, 2018; pp. 8759–8768. [CrossRef]
15. Wang, X.; Huang, Y.; Wei, S.; Xu, W.; Zhu, X.; Mu, J.; Chen, X. ELD-YOLO: A Lightweight Framework for Detecting Occluded Mandarin Fruits in Plant Research. *Plants* **2025**, *14*, 1729. [CrossRef] [PubMed]
16. Bargoti, S.; Underwood, J. Deep fruit detection in orchards. In *Proceedings of the 2017 IEEE International Conference on Robotics and Automation (ICRA)*; IEEE: New York, NY, USA, 2017; pp. 3626–3633. [CrossRef]
17. Kisantal, M.; Cheng, Y.; Melekhov, I.; Genova, K.; Corbett, F.; Trigoni, N.; Markham, A. Augmentation for small object detection. *arXiv* **2019**, arXiv:1902.07296. [CrossRef]

18. Zhang, H.; Cisse, M.; Dauphin, Y.N.; Lopez-Paz, D. Mixup: Beyond empirical risk minimization. In *Proceedings of the International Conference on Learning Representations (ICLR)*; OpenReview.net: Amherst, MA, USA, 2018. Available online: https://openreview.net/forum?id=r1Ddp1-Rb (accessed on 26 April 2026).
19. Chollet, F. Xception: Deep Learning with Depthwise Separable Convolutions. In *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: New York, NY, USA, 2017; pp. 1800–1807. [CrossRef]
20. Han, K.; Wang, Y.; Tian, Q.; Guo, J.; Xu, C.; Xu, C. GhostNet: More Features From Cheap Operations. In *Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: New York, NY, USA, 2020; pp. 1577–1586. [CrossRef]
21. Ma, N.; Zhang, X.; Zheng, H.T.; Sun, J. ShuffleNet V2: Practical Guidelines for Efficient CNN Architecture Design. In *Computer Vision—ECCV 2018*; Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y., Eds.; Lecture Notes in Computer Science; Springer: Cham, Switzerland, 2018; Volume 11218. [CrossRef]
22. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. CBAM: Convolutional Block Attention Module. In *Computer Vision—ECCV 2018*; Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y., Eds.; Lecture Notes in Computer Science; Springer: Cham, Switzerland, 2018; Volume 11211. [CrossRef]
23. Hu, J.; Shen, L.; Sun, G. Squeeze-and-Excitation Networks. In *Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: New York, NY, USA, 2018; pp. 7132–7141. [CrossRef]
24. Tan, M.; Le, Q.V. EfficientNet: Rethinking model scaling for convolutional neural networks. In *Proceedings of the International Conference on Machine Learning (ICML)*; PMLR: Cambridge, MA, USA, 2019; pp. 6105–6114. Available online: https://proceedings.mlr.press/v97/tan19a.html (accessed on 26 April 2026).
25. Jocher, G.; Chaurasia, A.; Qiu, J. Ultralytics YOLOv8. Available online: https://github.com/ultralytics/ultralytics (accessed on 20 April 2026).
26. Wang, C.Y.; Yeh, I.H.; Mark Liao, H.Y. YOLOv9: Learning What You Want to Learn Using Programmable Gradient Information. In *Computer Vision—ECCV 2024*; Leonardis, A., Ricci, E., Roth, S., Russakovsky, O., Sattler, T., Varol, G., Eds.; Lecture Notes in Computer Science; Springer: Cham, Switzerland, 2025; Volume 15089. [CrossRef]
27. Li, C.; Li, L.; Jiang, H.; Kakarantzas, K.; Kompella, G. YOLOv6: A single-stage object detection framework for industrial applications. *arXiv* **2022**, arXiv:2209.02976. [CrossRef]
28. Wang, A.; Chen, H.; Liu, L.; Chen, Y.; Shi, X.; Cheng, Y.; Liu, Z. YOLOv10: Real-time end-to-end object detection. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2024; Volume 37, pp. 107984–108011. Available online: https://arxiv.org/abs/2405.14458 (accessed on 26 April 2026).
29. Ge, Z.; Liu, S.; Wang, F.; Li, Z.; Sun, Y. YOLOX: Exceeding YOLO series in 2021. *arXiv* **2021**, arXiv:2107.08430. [CrossRef]
30. Wang, Q.; Wu, B.; Zhu, P.; Li, P.; Zuo, W.; Hu, Q. ECA-Net: Efficient Channel Attention for Deep Convolutional Neural Networks. In *Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*; IEEE: New York, NY, USA, 2020; pp. 11531–11539. [CrossRef]
31. Lv, W.; Xu, Y.; Cui, Y.; Zhou, Y.; Wang, X.; Cui, N.; Abdelfattah, A.; Chesser, T.; Abdelmoaty, A.; Xue, H. DETRs beat YOLOs on real-time object detection. *arXiv* **2023**, arXiv:2304.08069. [CrossRef]
32. Pei, B.; Wu, L.; Zheng, X.; Zhou, C.; Wang, D. Fs2PA: A Full-Scale Feature Synergistic Perception Architecture for Vehicular Infrared Object Detection via Physical Priors and Semantic Constraints. *Sensors* **2026**, *26*, 2257. [CrossRef] [PubMed]
33. Li, T.; Zhu, C.; Li, J.; Cao, H.; Bai, H. A real-time insulator condition detection model for UAV inspection based on FG-YOLO. *Meas. Sci. Technol.* **2025**, *36*, 056208. [CrossRef]
34. Selvaraju, R.R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; Batra, D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In *Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV)*; IEEE: New York, NY, USA, 2017; pp. 618–626. [CrossRef]
35. Cheng, G.; Han, J.; Lu, X. Remote Sensing Image Scene Classification: Benchmark and State of the Art. *Proc. IEEE* **2017**, *105*, 1865–1883. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.