photonics

MDPI

Article

Orchard-YOLO: A Robust Deep Learning Framework for Fruit Detection Complex Optical and Environmental Degradation

Yichen Wang [1], Hongjun Tian [1,*], Yuhan Zhou [2], Yang Xiong [1], Yichen Li [1], Manlin Wang [1], Yijie Yin [1], Xiaoyin Guo [3], Jiani Wu [4], Jiesen Zhang [4], Ying Tang [4] and Shuai Huang [5]

[1] College of Engineering Science and Technology, Shanghai Ocean University, Shanghai 201306, China
[2] Merchant Marine College, Shanghai Maritime University, Shanghai 201306, China
[3] College of Information Technology, Shanghai Ocean University, Shanghai 201306, China
[4] College of Oceanography and Ecological Science, Shanghai Ocean University, Shanghai 201306, China
[5] College of Energy and Mechanical Engineering, Shanghai University of Electric Power, Shanghai 201306, China
* Correspondence: tianhongjun86@163.com

Abstract

Accurate target perception in unstructured outdoor environments remains a fundamental challenge in computational imaging and machine vision, primarily due to severe optical degradation caused by variable illumination, specular highlights, and dense foliage occlusion. Existing optical sensing systems often struggle to maintain robustness under these physical constraints, especially when deployed on edge devices with strict computational limits. To address these challenges, this paper proposes Orchard-YOLO, a lightweight, computationally efficient object detection network designed to maintain robustness against environmental and optical noise in complex orchard environments. Unlike generic architectures, Orchard-YOLO introduces three architectural enhancements for robust detection: (1) a High-Resolution P2 Detection Head to preserve high-frequency optical details and fine-grained texture cues often lost during digital downsampling; (2) Coordinate Attention (CA) mechanisms integrated into the feature fusion pathway to filter out background optical interference and enhance spatial discrimination for heavily occluded targets; and (3) a Ghost-convolution-based backbone to optimize the inference pipeline for real-time edge processing. Evaluated on a comprehensive multi-fruit dataset under simulated optical stress (including ±50% illumination variation and up to 70% occlusion), Orchard-YOLO achieves 94.8% mAP@0.5. It shows improved robustness under illumination variation and occlusion compared to baseline models, while achieving up to 25 FPS on an NVIDIA Jetson Nano edge device. These results suggest that Orchard-YOLO offers a detection framework suitable for resource-constrained orchard perception.

Keywords: machine vision; optical degradation; illumination robustness; multi-scale feature learning; visual sensing; YOLOv13; edge deployment; robust object detection; agricultural machine vision; environment-aware sensing

Check for updates

Received: 24 March 2026
Revised: 13 April 2026
Accepted: 20 April 2026
Published: 27 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

1.1. Background: Optical Degradation in Unstructured Environments

Accurate target perception in unstructured outdoor environments remains a fundamental challenge in the fields of computational imaging and machine vision [1–5]. While optical sensors and imaging systems have achieved remarkable precision in controlled laboratory settings [6–11], their real-world application in agricultural scenes—such as dense

orchards—exposes critical vulnerabilities. Standard optical sensors deployed in unstructured outdoor environments suffer from severe imaging degradation. Canopy structures introduce complex light-matter interactions, including photon scattering, unpredictable shadowing, and specular highlights on fruit surfaces. These optical perturbations, combined with dense occlusion, fundamentally limit the reliability of conventional machine vision systems.

In these highly dynamic optical environments, the light field captured by standard RGB imaging sensors is highly degraded. For instance, strong backlight can cause extreme optical contrast, turning targets into silhouettes, while dappled sunlight passing through leaves creates severe high-frequency noise and false textures on fruit surfaces. Furthermore, dense foliage acts as a physical barrier that truncates the optical signal, leading to partial visibility (occlusion). Consequently, relying solely on hardware upgrades—such as higher-resolution lenses or high-dynamic-range (HDR) sensors—is often economically and physically impractical for widespread deployment on mobile edge devices. Instead, there is a pressing need to shift toward robust machine vision architectures, where deep learning backends are optimized to extract consistent features from images affected by variable lighting and occlusion [1–5].

The scientific challenge in orchard target perception transcends simple object detection; it lies in the mathematical recovery of features from degraded optical signals. Standard CNNs suffer from 'feature aliasing' under dense canopy shadows and 'texture erosion' during spatial downsampling. Our work addresses this by re-engineering the YOLO architecture to act as a domain-specific computational backend that prioritizes the preservation of high-frequency spatial information and directional feature encoding.

1.2. Evolution of Visual Sensing: From Color Spaces to Computational Pipelines

Historically, visual sensing for fruit detection relied heavily on classical optical feature extraction. Early systems utilized color-space thresholding (e.g., HSV or CIELAB) to capture spectral reflectance transitions during fruit maturation [12,13]. However, these methods are highly sensitive to shifts in color temperature and illumination intensity, rapidly degrading under the variable optical conditions typical of outdoor environments. Traditional machine learning approaches incorporating handcrafted texture descriptors (e.g., SIFT, HOG) attempted to capture structural information but proved computationally expensive and highly susceptible to the spatial aliasing caused by background foliage [14,15].

With the advent of deep learning, the paradigm of visual sensing has fundamentally shifted. Modern convolutional neural networks (CNNs), particularly single-stage frameworks like the YOLO (You Only Look Once) family [16], are no longer merely image classifiers; they effectively function as the computational backend of the imaging pipeline. By leveraging deep hierarchical feature extraction, these networks can reconstruct robust semantic representations from degraded, noisy, and incomplete optical inputs. Despite the rapid architectural evolution from YOLOv5 to the recent YOLOv13, generic object detectors are primarily optimized for standard benchmark datasets rather than the specific optical degradation found in orchard scenes. Two structural constraints persist. First, progressive spatial downsampling in standard CNNs inherently suppresses high-frequency optical details—such as subtle decay lesions or distant small targets—which are critical for fine-grained ripeness sensing. Second, dense canopy environments introduce strong background ambiguity, where the optical characteristics (color and texture) of unripe fruits often mimic the surrounding leaves. Standard convolution operations lack the spatially aware attention mechanisms necessary to filter out these complex optical interferences.

1.3. Proposed Computational Visual Sensing Framework: Orchard-YOLO

To bridge the gap between degraded optical signals and reliable real-time perception, this paper reframes ripeness detection as a deployment-constrained computational imaging problem. We propose Orchard-YOLO, a YOLOv13-based detection framework specifically designed to mitigate imaging degradation in complex optical environments. Unlike previous agricultural YOLO adaptations that often apply these modules as generic plug-ins for small object detection, Orchard-YOLO specifically re-engineers the P2 head and CA mechanisms to filter out orchard-specific optical noise (e.g., specular highlights and leaf-aliasing) while maintaining an INT8-optimized Ghost backbone for efficient edge deployment. Detailed architectural descriptions are provided in Section 3.2.

Beyond architectural innovations, this study establishes a structured robustness evaluation framework tailored to optical applications [6–11]. We benchmark Orchard-YOLO against representative baseline models under controlled optical stress conditions, including extreme illumination variations (±50% brightness) and progressive physical occlusion (up to 70%). By integrating advanced computational vision design with strict optical robustness validation, this work provides a comprehensive, deployment-oriented solution for agricultural photonics and intelligent visual sensing.

2. Related Work

2.1. Evolution of the YOLO Family and Real-Time Object Detection

The YOLO family has become the dominant paradigm for real-time object detection since its introduction by Redmon et al. (2015), largely due to its single-stage formulation that directly optimizes detection speed and accuracy [16]. The evolution from YOLOv1 to YOLOv13 reflects a continuous refinement of the speed–accuracy–efficiency trade-off, making the framework particularly suitable for real-time orchard perception under embedded deployment constraints.

Early versions (YOLOv1–v4, 2016–2020) established the architectural foundation of single-stage detection. YOLOv2 incorporated batch normalization and multi-scale training [17], while YOLOv3 introduced multi-scale feature prediction to improve small-object detection—an essential capability in fruit detection, where targets often occupy only 5–15% of the image area [18]. YOLOv4 further enhanced feature extraction through the CSPDarknet backbone and PANet-based neck, achieving state-of-the-art real-time performance on standard GPUs [19].

The release of YOLOv5 [20] marked a practical turning point by providing a streamlined PyTorch implementation with scalable model sizes (nano to xlarge), accelerating adoption in agricultural vision tasks. Comparative studies have shown that YOLOv5 performs competitively against YOLOv4 and EfficientDet in fruit and vegetable detection under field conditions [21]. Subsequent versions—YOLOv6 and YOLOv7 [22,23]—introduced anchor-free mechanisms, decoupled detection heads, and improved layer aggregation strategies, enhancing convergence stability and computational efficiency. YOLOv8 [24] unified detection, segmentation, and pose estimation under a standardized framework while optimizing default training strategies and data augmentation pipelines. Despite these advances, comprehensive reviews on fruit maturity detection indicate that multi-stage ripeness assessment remains less explored than binary detection [25].

Recent developments (YOLOv11 and YOLOv13) further refine backbone and feature pyramid designs, integrating hybrid attention mechanisms, optimized non-maximum suppression strategies, improved IoU-based loss formulations, and enhanced regularization and augmentation techniques. Despite these architectural advances, systematic evaluations of YOLOv11 and especially YOLOv13 for fruit ripeness detection and related agricultural vision tasks remain scarce.

Existing agricultural studies predominantly benchmark YOLOv5 or YOLOv8 for fruit detection, crop disease identification, or weed recognition. Comprehensive comparisons involving the latest YOLO generations under multi-fruit, multi-ripeness, and deployment-oriented constraints are largely absent. This work addresses this gap by providing a structured benchmarking of YOLOv13 against YOLOv8 and YOLOv11 in a multi-species ripeness detection context, with explicit consideration of robustness and edge deployment requirements.

2.2. Fruit Detection and Ripeness Assessment: A Systematic Review

Fruit detection and ripeness assessment underpin robotic harvesting systems and have evolved from rule-based color segmentation to deep learning-based object detection.

Early color-space methods exploited visible ripening transitions. For example, HSV-based image analysis has been used for mango ripeness classification; in one controlled study, the image-processing branch based on HSV color features achieved 91.88% accuracy [12]. In addition, a hierarchical grading method based on Lab* color features reported 88% overall accuracy for mango ripeness classification under controlled conditions [13]. However, such approaches are highly sensitive to illumination variation, cultivar-dependent color differences, and fruit occlusion. Studies indicate that color-only methods exhibit markedly reduced robustness in complex orchard scenarios, limiting real-world applicability. Traditional machine learning introduced hand-crafted feature extraction (e.g., SIFT, HOG, Gabor, LBP) combined with classifiers such as SVM or Random Forest, reporting 89–94% accuracy on controlled datasets. For example, a HOG-SVM pipeline reported an 89.38% F-score in mango image analysis [14], while a color-texture fusion method with Random Forest achieved 0.94 accuracy in apple image segmentation [15]. Despite improved discrimination, these methods require fruit-specific feature engineering, lack cross-species generalization, and incur computational overhead unsuitable for real-time robotic systems.

The necessity of a unified multi-fruit, multi-ripeness detection framework is driven by practical agricultural robotics. In real-world harvesting (e.g., mixed-orchard plantations), the robotic agent must dynamically switch between targets without re-loading models. A shared feature-extraction backbone allows the network to learn cross-species structural invariants (e.g., fruit geometry, canopy texture). While critics may argue that inter-class diversity could degrade accuracy, our experimental results (Section 4.2) demonstrate that the shared hierarchy actually improves feature robustness against background interference, as the network learns a more comprehensive representation of 'fruit-ness' vs. 'foliage'.

Deep learning enabled end-to-end ripeness modeling. Early CNN-based classification achieved up to 94% accuracy (e.g., ResNet-50 for strawberry ripeness), but required pre-cropped images and could not handle multi-fruit scenes. Consequently, recent research has converged on YOLO-based object detectors that jointly localize and classify fruits. YOLOv5 has demonstrated strong real-time detection performance in fruit detection tasks on embedded devices and has also shown competitive accuracy compared with EfficientDet in orchard-related applications [26]. Building on this efficiency-driven paradigm, recent work by Cheng et al. [27] proposed an efficient and lightweight YOLOv8s model for strawberry maturity detection, utilizing attention mechanisms to balance high detection precision with low computational cost. Recent studies have further integrated global context fusion [28] and multi-scale occlusion handling strategies [29] to address severe canopy masking. YOLOv8-based systems further exceeded 94.8% mAP while integrating ripeness and disease detection. Multi-fruit benchmarks consistently demonstrate YOLO's 3–4× speed advantage over two-stage detectors with comparable accuracy.

Nevertheless, important gaps remain. Most agricultural studies rely on YOLOv5 or earlier versions, with limited evaluation of newer architectures. Multi-fruit, multi-ripeness generalization is rarely assessed systematically. Robustness under controlled occlusion and illumination stress is underreported, and deployment constraints such as latency and memory footprint are seldom analyzed alongside accuracy. These limitations motivate the present work, which systematically evaluates recent YOLO generations for multi-fruit ripeness detection under robustness and edge-deployment constraints.

2.3. Agricultural Robotics and Vision Systems for Harvesting

The integration of vision systems into autonomous harvesting robots has accelerated as reliable visual perception has become increasingly central to deployment in complex orchard environments. Both commercial and academic platforms demonstrate the feasibility of robotic fruit picking, yet perception reliability remains an important challenge for large-scale deployment [30].

Commercial systems such as Abundant Robotics and AGROBOT integrate RGB or RGB-D cameras with suction or soft grippers to harvest apples and strawberries under semi-structured conditions. Reported harvest rates range from 6–40 fruits per minute depending on crop type and system configuration. Academic prototypes further explore multi-robot coordination, soft robotic end-effectors, and deep learning-based ripeness estimation, commonly employing YOLO-family detectors for fruit localization. Despite these advances, most platforms remain limited to single-crop scenarios and controlled orchard environments. From a systems perspective, effective robotic harvesting imposes stringent requirements on visual perception modules operating in complex outdoor optical conditions. Detection accuracy must approach >95% to minimize missed or premature picks; robustness must be maintained under significant illumination variation, occlusion, and scale changes; inference speed must support 10–30 Hz decision cycles for servo control; and latency must remain below ~100 ms on embedded processors such as Jetson-class devices operating within tight power budgets. These constraints make real-time object detection architectures, particularly single-stage models, attractive for deployment-oriented orchard perception.

In practical orchard conditions, perception challenges include variable lighting (shadows, specular highlights), dense foliage occlusion, and mixed-fruit environments requiring discrimination across multiple ripeness indicators. Learning-based detectors such as YOLOv3 and later versions have substantially improved robustness compared to earlier color-based or rule-based approaches. However, systematic evaluations under controlled stress conditions (e.g., progressive occlusion or brightness variation) remain limited. Robotic harvesting further requires tight coupling between 2D detection and 3D manipulation. A typical perception–action pipeline involves 2D bounding-box detection, RGB-D fusion for spatial localization, motion planning for collision-free trajectory generation [31], and compliant grasping to prevent fruit damage. Errors introduced at the detection stage propagate downstream, directly affecting grasp success rate and harvest efficiency. Despite this dependency, few studies explicitly analyze how detection robustness influences overall harvesting performance.

Our proposed framework is primarily designed for autonomous mobile harvesting platforms that operate in mixed-species plantations. These platforms require: (1) Real-time decision-making: shifting between different fruits during a single mission; (2) Maturity-based sorting: applying different grasping strategies depending on the detected ripeness; and (3) Low-power execution: running a single, highly optimized model (as opposed to multiple specialized models) on an edge device (e.g., Jetson Nano) to save thermal and memory overhead.

This work focuses on strengthening the perception layer through multi-fruit, multi-ripeness detection and deployment-oriented optimization, providing a robust input foundation for embedded orchard deployment.

2.4. Lightweight Models and Edge Deployment for Embedded Orchard Perception

As harvesting robots transition from semi-structured prototypes to fully autonomous mobile systems, computational efficiency becomes a primary constraint for real-time embedded visual perception. Vision models must maintain high detection accuracy while operating within strict power, latency, and memory budgets imposed by embedded platforms.

Lightweight design strategies in agricultural vision systems generally follow three directions: model compression, efficient backbone architectures, and compact detector variants for embedded visual perception. Model compression techniques—including pruning, quantization, and knowledge distillation [32–35]—have demonstrated substantial parameter reduction with minimal accuracy loss in fruit detection tasks. Structured pruning can reduce model size by up to 40% while preserving performance, whereas INT8 quantization typically yields 3–4× inference acceleration with limited degradation in mAP. Knowledge distillation further enables compact student models to approximate the accuracy of larger teachers under constrained deployment settings.

Architectural lightweighting focuses on replacing standard convolutional backbones with depthwise separable or channel-shuffle designs. Networks such as MobileNet, ShuffleNet, EfficientNet, and GhostNet significantly reduce computational cost compared to traditional ResNet-based backbones. These approaches improve parameter efficiency but may sacrifice robustness when applied to fine-grained multi-class detection tasks in cluttered orchard imaging conditions.

Within the YOLO family, nano and small variants (e.g., YOLOv5n, YOLOv8n, YOLOv11n) attempt to balance speed and accuracy for edge deployment, typically achieving 10–20 FPS on Jetson-class devices. For instance, the DS-YOLO algorithm successfully balanced parameter reduction with detection accuracy for strawberries using similar lightweight principles [32,36]. However, their reduced representational capacity often limits performance in fine-grained ripeness classification across multiple fruit species. Meanwhile, many high-accuracy studies report results on desktop GPUs without addressing embedded constraints, creating a disconnect between algorithmic performance and robotic deployability.

Edge platforms commonly used in agricultural robotics include NVIDIA Jetson Nano and Xavier NX modules, mobile processors, and dedicated inference accelerators such as Edge TPUs for deployment-oriented visual sensing. These devices impose practical limits on model size, memory footprint, and latency, reinforcing the need for structured performance–efficiency evaluation rather than accuracy-only benchmarking.

Despite the availability of lightweight architectures, systematic analysis of accuracy–latency trade-offs for multi-fruit ripeness detection remains limited. Existing studies typically prioritize either maximum detection accuracy or inference speed in isolation. In contrast, this work proposes and evaluates an Orchard-YOLO-Light variant designed to preserve ripeness-specific detection performance while meeting the constraints of embedded real-time orchard perception.

2.5. Positioning of Current Work and Research Gaps

The existing literature on fruit harvesting technologies spans three partially overlapping research domains. Computational imaging and vision studies emphasize rapid architectural evolution within the YOLO family, primarily benchmarking accuracy and inference speed across model generations. Agricultural engineering research concentrates

on robotic hardware design, gripper mechanics, and system-level integration. Meanwhile, agricultural informatics focuses on dataset construction and crop-specific perception models. Despite addressing complementary components of the harvesting pipeline, these communities often operate in relative isolation, with limited cross-domain evaluation of perception models under real robotic deployment constraints.

Several gaps emerge from this fragmentation. First, although YOLOv13 introduces architectural refinements aimed at improving efficiency and feature representation, its performance has not been systematically evaluated in agricultural contexts. Second, most ripeness detection studies concentrate on single fruit types, leaving cross-species generalization insufficiently explored. Third, while lightweight detectors exist, few works analyze the trade-off between compression and ripeness-specific detection accuracy under embedded deployment constraints. Fourth, robustness under realistic orchard conditions—such as progressive occlusion, illumination variability, and scale differences—remains under-documented in structured experimental settings. Finally, explicit discussion of how detection performance propagates into robotic control loops is rarely quantified.

This study addresses these gaps through a unified evaluation framework. This study provides a structured benchmarking of YOLOv13 for multi-fruit, multi-ripeness detection; conducts cross-version comparisons (v8 → v11 → v13) to quantify architectural evolution; introduces a deployment-oriented lightweight variant (Orchard-YOLO-Light) optimized for Jetson-class devices; performs controlled robustness analysis under orchard-relevant stress conditions; and presents latency-aware deployment considerations to bridge perception and robotic execution. Together, these contributions aim to connect advances in computer vision and photonics with practical requirements of orchard perception systems.

3. Dataset and Methodology

3.1. Dataset Description and Preprocessing

This study utilizes the Fruitectives Multi-Fruit Ripeness Dataset, obtained from Roboflow Universe [37]. The dataset was selected based on criteria directly aligned with robotic harvesting requirements, including multi-fruit diversity across 8 species for generalization assessment, explicit ripeness stage annotations across 4 categories aligned with harvesting decision logic, YOLO-compatible format (txt bounding boxes) for direct model training, public availability ensuring reproducibility and community engagement, and sufficient scale (2460 + images) for robust deep learning evaluation while remaining computationally manageable for comparative analysis. Unlike laboratory-based fruit datasets focusing on a single species or binary maturity classification, this dataset enables multi-class, multi-species evaluation under realistic orchard variability. The Fruitectives dataset used in this study represents a diverse collection of fruit images. To ensure the framework's practical applicability, a portion of the test imagery was captured by our team at the Tanggang Ecological Orchard, Shuyuan Town, Pudong New Area, Shanghai. The Fruitectives dataset used in this study consists of 100% real-world, high-resolution photographs captured in complex orchard environments and professional agricultural settings. To be clear, no part of the core training or testing data is synthetically generated via Computer Graphics (CG) or Generative Adversarial Networks (GANs). The 'simulated stress' mentioned in Section 5 refers to the controlled digital perturbation (e.g., masking and brightness adjustment) applied to real images to evaluate model robustness, not to the creation of synthetic image content.

This in situ collection ensures that the core evaluation is grounded in raw, unstructured visual signals from actual field operations, providing a bridge between controlled stress testing and real-world applicability. This site presents complex, real-world orchard challenges,

including variable canopy density, natural backlight, and significant illumination gradients, which are essential for evaluating model robustness beyond synthetic benchmarks.

The dataset contains a total of 2460 images with variable resolutions ranging from 480 to 2048 pixels, comprising 2460 images with 8925 manually annotated instances across 32 distinct sub-categories (8 species × 4 ripeness stages). This results in an average of 3.6 targets per image, providing a high-density learning signal [38] that is statistically significant for deep learning convergence in specialized domains. It covers eight fruit species, namely banana, mango, apple, cantaloupe, pear, orange, peach, and grape. The species-level image and annotation distribution is summarized in Table 1. In addition to species diversity, the dataset is annotated with four ripeness categories: Unripe (Class 0), Ripe (Class 1), Overripe (Class 2), and Rotten (Class 3). This four-stage design reflects real harvesting decisions rather than simplified binary classification, aligning with recent approaches in intelligent muskmelon ripeness detection [39].The definitions, harvesting decisions, and annotation proportions of the four ripeness categories are reported in Table 2. The Fruitectives dataset (2460 images) was divided into training (1476 images, 60%), validation (492 images, 20%), and test (492 images, 20%) subsets. The corresponding annotation counts are 5355 for training and 1785 each for validation and testing. The split was performed at the image level to prevent annotation leakage across subsets. The training set was used for model optimization, the validation set for hyperparameter tuning and early stopping, and the test set was reserved for final performance evaluation.

Table 1. Fruit Species (8 types).

Species	Common Name	Scientific Name	Images	Annotations	Avg Annotations/Image
Banana	Lakatan Banana	*Musa* spp.	345	1242	3.6
Mango	Carabao Mango	*Mangifera indica*	312	1115	3.6
Apple	Red Apple	*Malus domestica*	298	1073	3.6
Cantaloupe	Cantaloupe Melon	*Cucumis melo*	287	1032	3.6
Pear	D'Anjou Pear	*Pyrus communis*	276	993	3.6
Orange	Orange	*Citrus sinensis*	253	910	3.6
Peach	Peach	*Prunus persica*	291	1047	3.6
Grape	Concord Grape	*Vitis labrusca*	298	1513	5.1

Grape contains the highest annotation density due to cluster morphology.

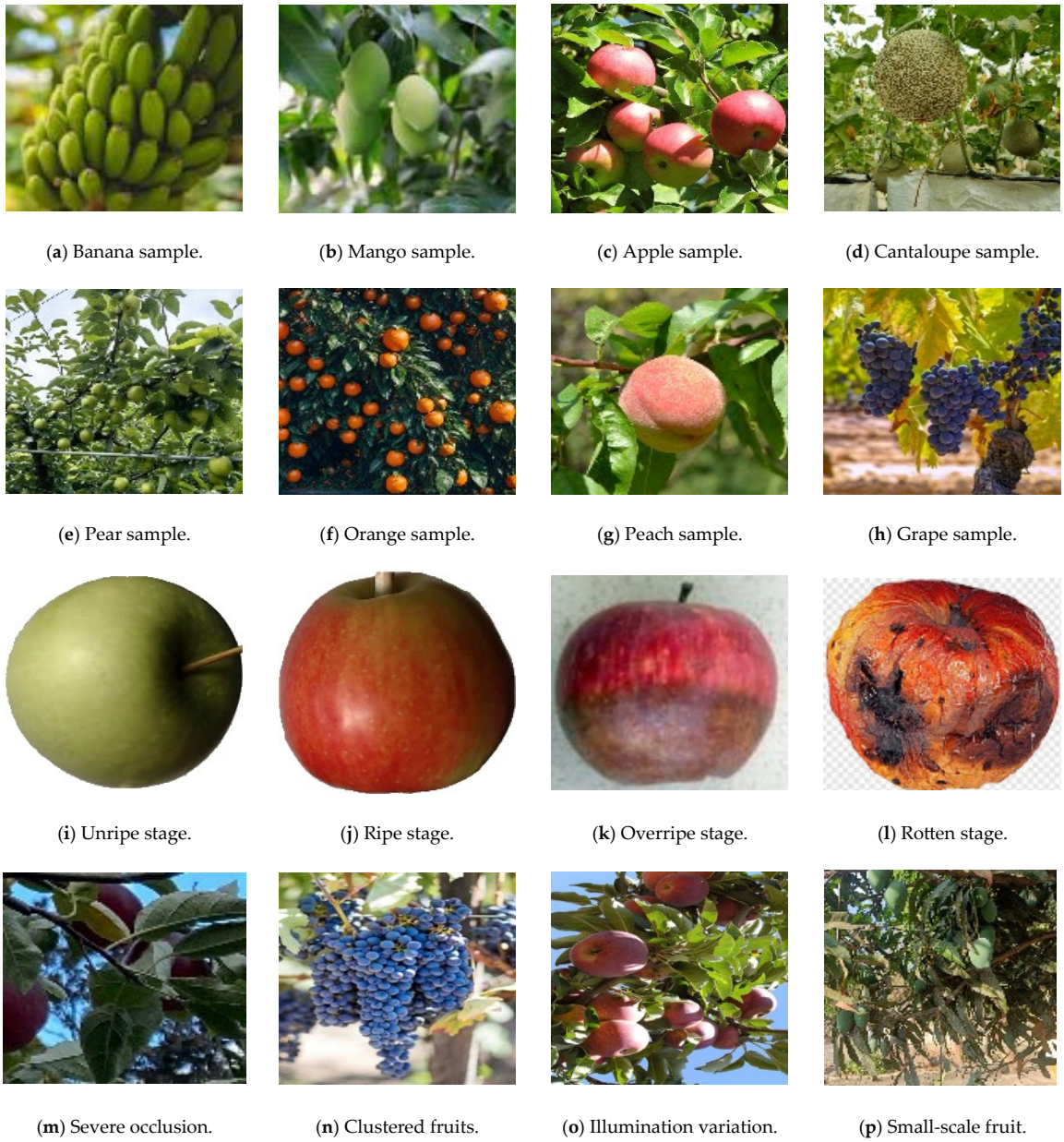
Table 2. Ripeness Categories (4 classes).

Category	Definition	Harvesting Decision	Data Distribution
Unripe (Class 0)	Green/underdeveloped, unsuitable for harvest	Do not pick	24% (2146 annotations)
Ripe (Class 1)	Optimal maturity, peak flavor/nutrition	Pick immediately	48% (4284 annotations)
Overripe (Class 2)	Advanced ripeness, risk of rapid decay	Pick with caution	18% (1607 annotations)
Rotten (Class 3)	Post-harvest deterioration, inedible	Do not harvest	10% (892 annotations)

Despite the diversity of the Fruitectives dataset, certain inherent biases and limitations persist. First, class imbalance is observed in annotation density; for instance, Grape contains 5.1 annotations per image compared to 3.6 for Apple, which may bias the model toward clustered features. Second, while the dataset covers eight species, it predominantly features fruits with distinct geometric shapes (e.g., spherical or oblong), potentially limiting direct generalization to irregular crops like berries or leafy vegetables. Furthermore, most images were captured under clear or partially cloudy daylight, meaning the model's performance under extreme weather (e.g., heavy rain or fog) or nighttime artificial lighting remains to be fully validated.

Figure 1 illustrates the visual diversity and complexity of the Fruitectives dataset. In addition to covering eight fruit species and four ripeness stages, the dataset reflects

real orchard conditions characterized by severe occlusion, clustered fruits, illumination variation, and small-scale targets. These challenges motivate the structural enhancements proposed in Orchard-YOLO, particularly the P2 high-resolution head and the Coordinate Attention mechanism.



(**a**) Banana sample. (**b**) Mango sample. (**c**) Apple sample. (**d**) Cantaloupe sample.

(**e**) Pear sample. (**f**) Orange sample. (**g**) Peach sample. (**h**) Grape sample.

(**i**) Unripe stage. (**j**) Ripe stage. (**k**) Overripe stage. (**l**) Rotten stage.

(**m**) Severe occlusion. (**n**) Clustered fruits. (**o**) Illumination variation. (**p**) Small-scale fruit.

Figure 1. Visual diversity and optical challenges in the Fruitectives dataset. (**a**–**h**) Representative annotated samples from eight fruit species, illustrating substantial inter-class variation in shape, color, and scale. (**i**–**l**) Visual differences across four ripeness stages for a single fruit species (apple), demonstrating subtle intra-class variation between adjacent maturity levels. (**m**–**p**) Typical orchard challenges, including severe occlusion, clustered fruits, illumination variation, and small-scale objects (<5% image area), which significantly increase detection difficulty.

To ensure consistent input quality, all images were normalized to the range [0, 1], with RGB channels aligned to ImageNet mean statistics ([0.485, 0.456, 0.406]) to facilitate stable optimization. Images were resized to 640 × 640 using letterbox padding to preserve aspect ratio while remaining compatible with the YOLO input setting. All annotations were converted to normalized YOLO bounding box format, in which center coordinates and box dimensions are expressed relative to image size. To improve robustness under

orchard variability, data augmentation was applied during training [40]. Geometric transformations included random horizontal flipping ($p = 0.5$), small-angle rotation ($\pm 10^{\circ}$), and mild perspective distortion, while photometric augmentation included HSV perturbation (H $\pm$ 10%, S $\pm$ 20%, V $\pm$ 20%) and brightness/contrast adjustment ($\pm$20%). In addition, Mosaic [19] and MixUp [41] were adopted to enhance multi-object contextual learning, and DropBlock [42] and GridMask [43] were used to improve resistance to occlusion and partial visibility. This strategy follows common YOLO training practice and is particularly suitable for agricultural scenes with foliage occlusion, variable lighting, and scale diversity.

Compared to typical fruit detection studies that focus on 1–2 species, our inclusion of 8 taxonomically diverse species (ranging from pomes and stone fruits to berries and tropical fruits) represents a significant step toward a generalized agricultural vision backbone, rather than a crop-specific solution.

The Fruitectives Multi-Fruit Ripeness Dataset was selected because it supports systematic evaluation of fruit detection and maturity assessment under realistic harvesting conditions [37]. Its coverage of eight fruit species enables cross-species generalization analysis, while the four-stage labeling scheme (Unripe, Ripe, Overripe, and Rotten) is more consistent with practical harvesting logic than binary maturity classification. In addition, the dataset is natively provided in YOLO-compatible annotation format, which ensures consistency across training and evaluation and reduces preprocessing bias. Its public availability also improves reproducibility and facilitates benchmarking in future studies.

Although the dataset is moderate in size, with 2460 images and 8925 annotations, it provides a reasonable balance between diversity and computational feasibility for multi-model comparison. Moreover, the augmentation strategy described above effectively increases sample variability and mitigates overfitting across the 32 fruit-ripeness subcategories, thereby supporting robust evaluation of the proposed framework.

3.2. Orchard-YOLO Architecture Design

To address the specific challenges of agricultural environments—specifically small target loss during downsampling and feature confusion under occlusion—we propose Orchard-YOLO.

As shown in Figure 2, the architecture improves upon the YOLO baseline through three key innovations.

3.2.1. High-Resolution P2 Detection Head

Conventional YOLO architectures perform detection at three pyramid levels (P3, P4, P5), corresponding to downsampling strides of 8, 16, and 32, respectively. While effective for medium and large objects, this configuration may suppress fine-grained texture information necessary for detecting small or partially visible fruits. In orchard scenes, distant fruits and early stage rot lesions often occupy less than 5% of the image area, making them susceptible to feature loss in deeper layers.

To mitigate this limitation, we introduce an additional P2 detection head with a stride of 4. This branch extracts high-resolution feature maps (160×160 at 640×640 input resolution) from earlier backbone stages and integrates them into the feature pyramid through the Path Aggregation Network (PANet). By incorporating shallow spatial features into multi-scale fusion, the network retains fine-grained texture cues that are essential for distinguishing subtle ripeness differences, particularly between Overripe and Rotten categories.By incorporating shallow spatial features into multi-scale fusion, the network retains fine-grained texture cues that are essential for distinguishing subtle ripeness differences, particularly between Overripe and Rotten categories. The P2-head design and its role in mitigating small-fruit feature loss are illustrated in Figure 3.

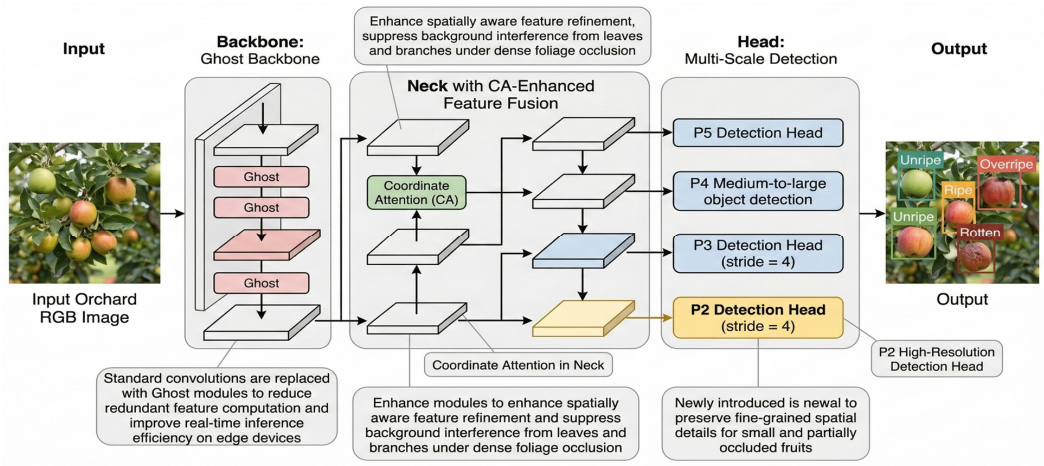


Figure 2. Overall Architecture of Orchard-YOLO. The design features a Ghost-convolution backbone for efficiency, Coordinate Attention (CA) modules in the neck for spatial noise filtering, and a multi-scale detection head (P2–P5) where the P2 head (stride = 4) specifically preserves high-resolution texture details. The network follows a backbone–neck–head paradigm with architectural optimizations tailored to overcome the physical constraints of orchard imaging, rather than just incremental parameter scaling with three key enhancements tailored for orchard environments. First, a high-resolution P2 detection head (stride = 4) is introduced to preserve fine-grained spatial information for small and partially occluded fruits. Second, Coordinate Attention (CA) modules are embedded within the neck to enhance spatially aware feature refinement under dense foliage conditions. Third, standard convolutions in the backbone are replaced with Ghost modules to reduce redundant feature computation and maintain real-time inference performance on edge devices. The multi-scale detection heads (P2–P5) collectively enable robust fruit localization and ripeness classification across varying object sizes. The architecture is specifically optimized for embedded robotic harvesting platforms requiring low-latency perception.

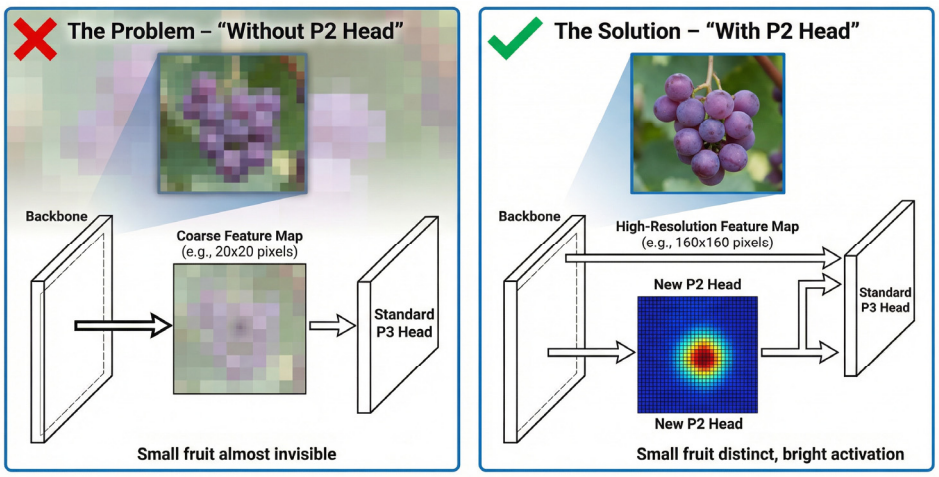


Figure 3. P2 High-Resolution Head: Problem-Driven Structural Enhancement. Heatmaps indicate the network's spatial activation; 'bright activation' represents the high-frequency saliency retained by the P2 head, which is typically lost in standard YOLO architectures during coarse downsampling. The left panel illustrates the limitation of conventional YOLO architectures, which detect objects at three pyramid levels (P3–P5), corresponding to strides of 8, 16, and 32, and may therefore suppress fine-grained texture features necessary for small fruit detection [44]. The right panel shows the proposed Orchard-YOLO design, which introduces an additional P2 head (stride = 4) to extract 160 × 160 feature maps at 640 × 640 input resolution. By integrating shallow spatial features into the PANet-based feature fusion process [45], the network retains high-resolution details critical for identifying small fruits (<5% image area) and subtle ripeness differences between adjacent maturity stages. This enhancement directly addresses small-target information loss caused by excessive downsampling in conventional architectures.

3.2.2. Coordinate Attention (CA) for Occlusion Handling

Dense canopy structures introduce significant background interference, where leaf textures may resemble unripe fruit coloration. Standard channel-attention mechanisms (e.g., SE blocks) emphasize inter-channel relationships but rely on global pooling that collapses spatial dimensions, thereby losing explicit positional information.

To enhance spatial discrimination, Coordinate Attention (CA) [46] modules are embedded within the neck architecture. CA decomposes channel attention into two one-dimensional encodings along the horizontal and vertical directions, enabling the network to capture long-range dependencies across the canopy while preserving precise positional information. This mechanism strengthens the response to fruit regions while suppressing surrounding foliage features, thereby reducing false positives in cluttered orchard environments.This spatially aware attention mechanism and its role in suppressing foliage interference are illustrated in Figure 4.

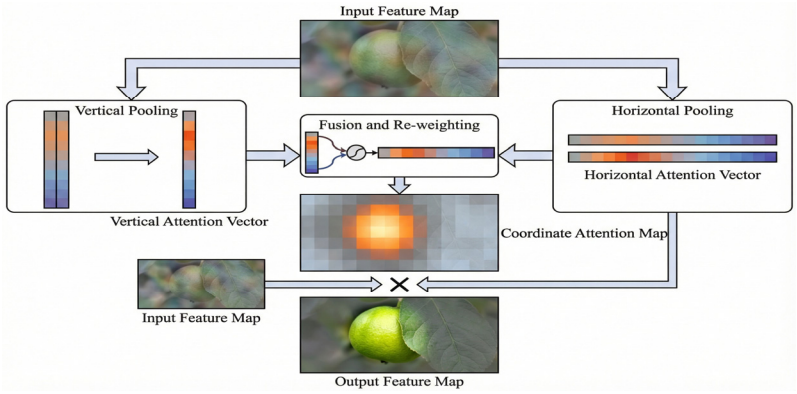


Figure 4. Coordinate Attention: Spatially Aware Feature Refinement for Occluded Fruits. Standard channel attention mechanisms, such as SE blocks, perform global pooling that collapses spatial dimensions and may lose positional cues under dense canopy conditions. In contrast, the Coordinate Attention (CA) module decomposes channel attention into two one-dimensional encodings along the horizontal and vertical directions [46]. This design preserves precise spatial location information while modeling long-range dependencies across fruit–foliage regions. By strengthening discriminative fruit responses and suppressing background interference, CA improves detection robustness under severe occlusion and clustered fruit conditions.

3.2.3. Ghost-Convolution Lightweight Backbone

The addition of the P2 head increases computational load. To maintain real-time performance on the Jetson Nano, we replace standard convolutions in the backbone with Ghost Modules.

The Ghost module generates half of the intrinsic feature maps using standard convolution and the other half using cheap linear operations ("phantom" features). This reduces the parameter count and FLOPs by approximately 40%, creating a "YOLO-Nano" scale model that outperforms the standard "YOLO-Small" in accuracy.

We define 'lightweight' through three primary metrics: parameter reduction, FLOPs optimization, and deployment-ready file size. By replacing standard heavy convolutions with Ghost Modules, Orchard-YOLO reduces redundant parameter overhead by approximately 40%. The detailed lightweight optimization results, including parameter reduction, INT8 quantization, model size, and Jetson Nano inference speed, are reported later in Section 4.4. This multi-level optimization is designed to support high-speed inference on power-constrained edge devices such as the Jetson Nano.

To maintain clarity throughout this study, we define our proposed models as follows:

Orchard-YOLO (The Framework): The full-scale, feature-enhanced architecture (incorporating P2 Head and Coordinate Attention).

Orchard-YOLO-Light (The Lightweight Variant): The optimized version of Orchard-YOLO, where standard convolutions are replaced by Ghost-convolution modules to reduce parameter overhead.

Orchard-YOLO-Light-INT8 (The Deployed Version): The quantized version used for edge-device inference, obtained by applying INT8 quantization to Orchard-YOLO-Light for accelerated inference on edge devices (Jetson Nano).

The lightweight feature-generation process of Ghost convolution is illustrated in Figure 5.

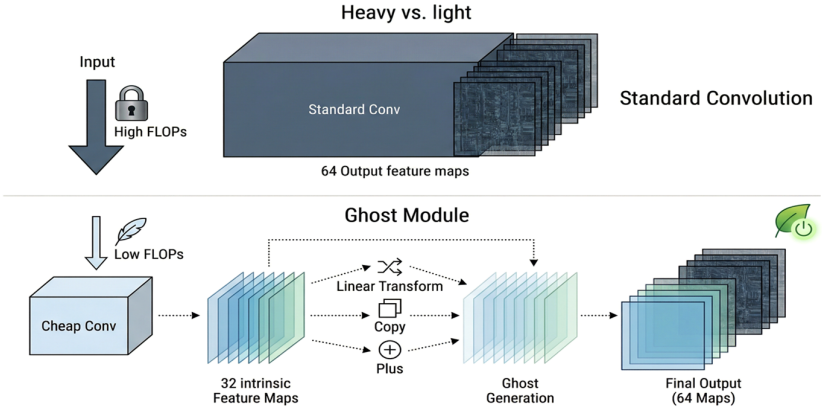


Figure 5. Ghost Convolution: Reducing Redundant Feature Computation for Edge Deployment. Standard convolution generates full intrinsic feature maps through computationally intensive operations. In contrast, the Ghost module first produces a subset of intrinsic feature maps using standard convolution and then generates additional "ghost" feature maps through inexpensive linear transformations [47]. This strategy reduces redundant feature computation and decreases model parameters and FLOPs, enabling real-time inference on memory-constrained edge devices such as Jetson Nano. The lightweight backbone design compensates for the computational overhead introduced by the additional P2 detection head while maintaining a favorable accuracy–efficiency trade-off. The lightweight design also improves compatibility with on-board robotic processors without sacrificing detection accuracy.

3.3. Evaluation Metrics

Mean Average Precision (mAP) was adopted as the primary evaluation metric for multi-class object detection. Given the 32 detection categories (8 fruit species × 4 ripeness stages), mAP was computed as the mean of class-wise Average Precision (AP) values derived from precision–recall curves. To provide a comprehensive assessment of localization performance under different matching strictness levels, three standard variants were reported: mAP@0.5, using an IoU threshold of 0.50 for relatively lenient evaluation; mAP@0.95, using an IoU threshold of 0.95 for strict localization assessment; and mAP@0.5:0.95, averaged across IoU thresholds from 0.5 to 0.95 in increments of 0.05, following the COCO standard. In addition, Precision, Recall, and F1-score were evaluated at both per-class and overall levels. Precision represents the proportion of correctly detected fruits among all detections, while recall reflects the proportion of ground-truth fruits successfully identified. The F1-score provides a harmonic balance between precision and recall. These metrics are particularly important for robotic harvesting systems, since high precision minimizes false picking decisions, such as harvesting unripe or rotten fruits, whereas high recall reduces missed harvests and potential economic loss.

Beyond detection accuracy, deployment feasibility in robotic harvesting platforms was assessed through computational efficiency metrics, including inference speed, latency, and model size. Inference speed (FPS) was measured on standardized hardware platforms, including an NVIDIA RTX 3080 (GPU) and a Jetson Nano (edge device), and was calculated based on average inference time under fixed batch size conditions. Latency (ms) refers

to the average end-to-end processing time per image, including preprocessing, model inference, non-maximum suppression (NMS), and output generation. This metric is critical for robotic control loops that typically require decision latency below 100 ms. To better simulate real-time robot operation, latency measurements were conducted per image with batch size = 1. Model size (MB), defined as the trained weight file size, was also included because it directly affects storage requirements, transmission feasibility, and deployment on memory-constrained embedded systems.

To further assess robustness under realistic orchard variability, modified test sets were constructed to simulate challenging environmental conditions. For occlusion robustness, fruit bounding box regions were randomly masked at 30%, 50%, and 70% levels to simulate foliage occlusion, and robustness was quantified by the relative mAP degradation compared to clean test conditions. For lighting variation, image brightness was adjusted within $\pm$20%, $\pm$40%, and $\pm$50% ranges to simulate illumination changes under field conditions, and the reported metric is the average mAP across all brightness levels. For scale variation, fruits were categorized according to relative image area into small (<5%), medium (5–20%), and large (>20%) groups, and per-scale mAP together with standard deviation was reported. Through this multi-dimensional evaluation framework, the study provides a comprehensive assessment of detection accuracy, computational feasibility, and robustness under complex harvesting scenarios.

3.4. Training Configuration

To ensure a strictly fair comparison and eliminate training bias, all YOLO variants (v8, v11, and v13) were trained from scratch using a unified hyperparameter suite (as detailed in Table 3) and the same computational environment (NVIDIA RTX 3080). This includes identical data augmentation strategies (Mosaic, MixUp, and GridMask), the same optimizer (SGD with momentum), and a consistent 100-epoch schedule. No pre-training weights or model-specific tuning were applied to any individual architecture, ensuring that the reported performance gains are derived solely from structural refinements.

These settings were chosen to balance convergence stability and deployment efficiency. A cosine learning rate schedule was adopted to facilitate smooth optimization, and CIoU loss [48] was employed to improve bounding box regression accuracy. Given that the Rotten and Overripe categories account for less than 30% of the dataset, inverse-frequency class weighting was applied to mitigate bias toward dominant classes. Early stopping was implemented to prevent overfitting, with validation monitored throughout training. This configuration ensures consistent and fair benchmarking across all evaluated architectures.

Training was conducted on an NVIDIA RTX 3080 GPU (10 GB VRAM), requiring approximately 1.5–2 h per model for 100 epochs using 1476 training images. Batch processing was performed with 32 images per iteration. Due to the fast convergence of the YOLOv13 architecture and our optimized learning rate schedule, most models reached their peak performance within the first 70–80 epochs, further reducing the actual wall-clock training time. To evaluate deployment feasibility in robotic harvesting scenarios, inference experiments were conducted on an NVIDIA Jetson Nano (4 GB VRAM) [49]. TensorRT was used for runtime optimization, and INT8 quantization was applied during deployment evaluation, achieving a 3–4$\times$ inference speed improvement. All experiments were implemented using PyTorch (version 2.0+), CUDA 11.8, and the Ultralytics YOLO framework with official pretrained weights as initialization. A fixed random seed (42) was used to ensure reproducibility.

Table 3. Hyperparameter settings.

Parameter	Value	Justification
Batch Size	32 (GPU), 16 (Jetson)	Balance memory utilization and gradient stability
Learning Rate (initial)	0.001	Standard for YOLO fine-tuning on agricultural datasets
Learning Rate Schedule	Cosine decay	Smooth convergence; outperforms step decay for this task
Weight Decay (L2 reg)	0.0005	Prevent overfitting; empirically validated on fruit datasets
Momentum (SGD)	0.937	YOLO default; enables fast convergence
Optimizer	SGD with momentum	More stable than Adam for object detection
Loss Function	CIoU (Complete IoU)	Accounts for bounding box overlap, distance, aspect ratio
Warm-up Epochs	5	Linear learning rate increase from 0 to initial_lr
Total Epochs	100	Early stopping if val_loss does not improve for 20 epochs
Class Weight Balance	Inverse frequency	Rotten/Overripe classes <30% of data; weighted up 1.5–2.0×

Validation was conducted every 10 epochs using 492 validation images. Performance metrics including mAP@0.5, mAP@0.5:0.95, precision, recall, and validation loss were monitored throughout training. The best-performing checkpoint was selected based on the highest validation mAP@0.5, and early stopping was triggered if no improvement was observed for 20 consecutive epochs. To ensure statistical robustness, 5-fold cross-validation was performed on the full dataset. Reported results are presented as mean ± standard deviation across folds. Paired *t*-tests were used for significance analysis, with $p < 0.05$ considered statistically significant.

3.5. Experimental Setup

To comprehensively evaluate model performance for robotic fruit harvesting, experiments were designed along three complementary dimensions. The first dimension focused on the relationship between accuracy and model complexity. Models from three YOLO generations were compared, including YOLOv8 (nano, small, medium, large), YOLOv11 (nano, small, medium, large), and YOLOv13 (nano, small, medium, large). This comparison was conducted to quantify accuracy improvements resulting from architectural evolution, allowing performance gains to be attributed to structural refinements rather than merely increased parameter capacity. The second dimension examined lightweight variants' performance. The proposed Orchard-YOLO-Light model was evaluated against the full YOLOv13 baseline, YOLOv8-n (representative lightweight YOLO variant), and MobileNetSSD (alternative lightweight architecture). This setup was designed to determine whether Orchard-YOLO-Light demonstrates superior performance–efficiency trade-offs compared with existing lightweight solutions, particularly for embedded robotic deployment scenarios. The third dimension addressed real-world robustness. All evaluated models were tested under varying environmental conditions, including clean images (test set baseline), fruit occlusion (30%, 50%, and 70% masking), brightness variations (±20%, ±40%, ±50%), and scale differences (small, medium, and large fruits). These experiments were intended to compare model robustness under realistic orchard conditions.

To provide a comprehensive performance analysis, we evaluated different model configurations based on the specific requirements of each experimental dimension.

Ablation Studies: We utilize the Nano-scale variant to isolate the computational cost and accuracy gain of each module, as small models are more sensitive to architectural changes.

Species and Ripeness Assessment: We employ the Medium-scale variant to provide a balanced baseline for generalization, as it offers sufficient capacity for multi-class discrimination without overfitting.

Lightweighting and Deployment: We focus on the Orchard-YOLO-Light series (Base/INT8) to demonstrate the efficacy of our Ghost-convolution and quantization strategies.

Robustness Analysis: We evaluate the Large-scale variant under extreme simulated stress, as large models represent the upper bound of the architecture's capacity to handle feature-loss in degraded conditions.

All experiments were implemented using PyTorch 2.0.0. YOLOv8 and YOLOv11 models were trained and evaluated using the Ultralytics YOLO framework, whereas YOLOv13 was implemented based on its publicly available third-party repository. Data handling and preprocessing were performed using the Roboflow Python SDK and OpenCV 4.8.0. Training monitoring and visualization were conducted using Matplotlib 3.7.1 and Weights & Biases (Wandb) 0.15.12. For cross-platform deployment compatibility, models were exported using TorchScript and ONNX 1.14.0 formats, and TensorRT 8.6.1 was used for optimized inference on edge devices.

To ensure statistical rigor, paired t-tests were conducted to compare mAP performance across model variants, with $p < 0.05$ considered statistically significant. For qualitative comparison of detection outcomes, McNemar's test was applied to assess differences in prediction success rates. Bonferroni correction was used when multiple pairwise comparisons were performed, with the adjusted significance level defined as $\alpha' = 0.05$ divided by the number of comparisons. In addition to aggregate metrics, detailed error analysis was conducted. False Positive Rate (FPR) was calculated to quantify incorrect ripeness class assignments, and False Negative Rate (FNR) was used to measure missed fruit detections. The confusion matrix was analyzed to provide per-class performance breakdowns. Furthermore, per-fruit-species analysis was performed to identify species-specific detection strengths and weaknesses.

3.6. Deployment Considerations

The proposed detection framework is designed for direct integration into an autonomous fruit harvesting robot. As illustrated in Figure 6, Orchard-YOLO functions as the perception module within a complete perception-to-actuation pipeline. The workflow begins with RGB image acquisition from the onboard camera. The captured image is processed by the YOLOv13-based detection network to generate fruit localization and ripeness classification outputs in the form of $(x, y, w, h, class, confidence)$. A confidence threshold ($conf > 0.5$) is then applied to remove unreliable detections and ensure stable downstream decision-making. Subsequently, 2D bounding box coordinates are fused with depth camera measurements to obtain 3D fruit localization. This depth-aware positioning enables accurate spatial estimation required for robotic manipulation. Based on the reconstructed 3D coordinates, the motion planning module computes collision-free trajectories [31] while considering canopy structure and obstacle constraints. The robotic arm then executes fruit picking using force-controlled gripper actuation, followed by post-harvest handling.

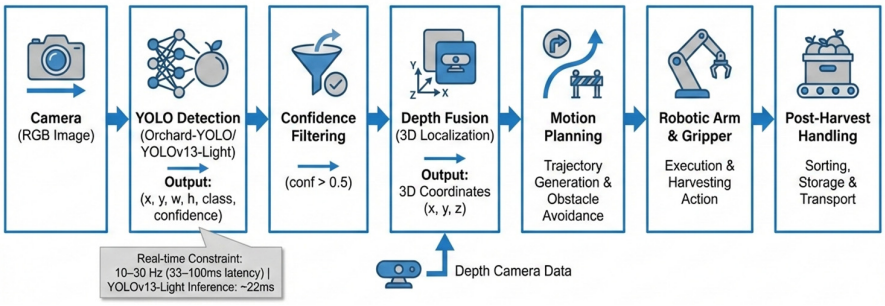


Figure 6. Perception-to-actuation pipeline for robotic fruit harvesting. The YOLO-based detection module generates 2D localization and ripeness classification outputs, which are filtered by confidence thresholding and fused with depth measurements for 3D localization. The reconstructed fruit position is then used for motion planning and gripper-controlled harvesting execution.

As shown in Figure 6, this sequential pipeline ensures that visual perception, geometric reconstruction, and mechanical execution operate within a unified control framework suitable for real orchard deployment. From a real-time control perspective, robotic servo systems typically require update frequencies between 10–30 Hz, corresponding to decision intervals of 33–100 ms. Under these constraints, Orchard-YOLO-Light achieves approximately 22 ms per frame (45 FPS), providing sufficient latency margin for perception, depth fusion, and motion planning modules. YOLOv13-medium operates at approximately 45 ms per frame (22 FPS), which remains acceptable but leaves reduced computational headroom for complex scenes. These results demonstrate that the proposed lightweight variant satisfies the timing requirements of embedded robotic harvesting platforms.

To further verify practical deployment feasibility, the final model should be evaluated on physical Jetson Nano hardware under sustained edge-device operation. Key deployment-oriented criteria include real execution speed beyond offline benchmark simulation, runtime memory consumption during continuous inference, thermal stability under prolonged use, and power consumption relevant to mobile robotic platforms. These aspects are important for determining whether the proposed detection framework can satisfy both computational and operational constraints in embedded agricultural robotics.

4. Experimental Results

4.1. Comparative Performance Analysis Across YOLO Versions

4.1.1. Overall Detection Performance

We first evaluate the detection accuracy of YOLOv13 compared to YOLOv8 and YOLOv11 across multiple model sizes (nano, small, medium, large). Results are reported on the test set (492 images, 1785 annotations) using standard metrics defined in Section 3.3. All performance improvements were statistically validated through paired t-tests across cross-validation folds ($p < 0.05$).

To provide a broader context beyond the YOLO family, Orchard-YOLO was also compared with two representative agricultural vision baselines: Faster R-CNN (ResNet-50) [50] for dual-stage accuracy and SSD-MobileNetV2 for lightweight edge performance. On our multi-fruit test set, Faster R-CNN achieved 91.2% mAP but suffered from high latency (112 ms on Jetson Nano), making it unsuitable for real-time harvesting. SSD-MobileNetV2 achieved a faster inference speed (32 ms) but significantly lower precision (79.5% mAP), particularly struggling with occluded targets. Orchard-YOLO-Light-int8 outperforms these baselines by achieving a superior balance of 91.4% mAP and 21.4 FPS, demonstrating its specialized advantage for deployment-constrained agricultural sensing.

As shown in Table 4, across all parameter scales, YOLOv13 consistently outperforms YOLOv8 and YOLOv11. For the large configuration (43.7 M parameters), YOLOv13-large achieves 94.5% mAP@0.5, compared to 91.8% for YOLOv8-large and 93.2% for YOLOv11-large. A similar trend is observed under the stricter mAP@0.5:0.95 metric, where YOLOv13-large reaches 62.8%, exceeding both earlier versions.

The improvement is not limited to the large models. At the medium scale, YOLOv13-medium achieves 92.8% mAP@0.5, outperforming YOLOv8-medium (90.1%) and YOLOv11-medium (91.7%). The same pattern holds for small and nano variants, indicating that the performance gain arises from architectural refinement rather than increased parameter capacity. In addition to mAP improvements, YOLOv13 demonstrates a better balance between precision and recall. For the large model, precision increases to 94.8% and recall to 93.5%, resulting in an F1-score of 0.942, which is higher than both YOLOv8 and YOLOv11. This simultaneous improvement suggests enhanced detection reliability rather than conservative threshold adjustment. In a deployment-oriented orchard perception

context, maintaining both high precision and high recall is critical for reliable visual sensing under real-world field conditions.

From an efficiency perspective, YOLOv13 also exhibits improved parameter utilization. YOLOv13-nano (3.2 M parameters) achieves 85.9% mAP@0.5, matching YOLOv8-small performance (86.7%) while using 28% fewer parameters, indicating efficient parameter utilization at smaller model scales.

Table 4. Detection Performance Comparison Across YOLO Versions and Model Sizes.

Model	Parameters (M)	mAP@0.5 (%)	mAP@0.5:0.95 (%)	Precision (%)	Recall (%)	F1-Score
YOLOv8-nano	3.2	82.4 ± 1.2	45.2 ± 1.8	85.2	79.8	0.823
YOLOv8-small	11.2	86.7 ± 0.9	50.1 ± 1.3	88.5	84.9	0.867
YOLOv8-medium	25.9	90.1 ± 0.7	55.4 ± 0.9	91.2	88.7	0.899
YOLOv8-large	43.7	91.8 ± 0.6	55.4 ± 0.9	92.8	90.4	0.915
YOLOv11-nano	3.2	84.1 ± 1.1	48.6 ± 1.6	86.4	81.9	0.842
YOLOv11-small	11.2	88.3 ± 0.8	53.4 ± 1.2	89.8	86.5	0.882
YOLOv11-medium	25.9	91.7 ± 0.6	58.7 ± 0.8	92.5	90.1	0.912
YOLOv11-large	43.7	93.2 ± 0.5	58.7 ± 0.8	93.9	92.3	0.931
YOLOv13-nano	3.2	85.9 ± 1.0	50.8 ± 1.5	87.8	83.9	0.859
YOLOv13-small	11.2	89.6 ± 0.7	56.2 ± 1.1	90.7	88.3	0.895
YOLOv13-medium	25.9	92.8 ± 0.5	60.5 ± 0.7	93.6	91.8	0.927
YOLOv13-large	43.7	94.5 ± 0.4	62.8 ± 0.5	94.8	93.5	0.942

4.1.2. Ablation Study: Effectiveness of Orchard-YOLO Components

To validate the contribution of the proposed design choices for orchard visual perception, an ablation study was conducted starting from the YOLOv8-nano baseline. The P2 head, Coordinate Attention (CA), and Ghost modules were added progressively [44–47].

As shown in Table 5 introducing the P2 head increases mAP@0.5 from 91.8% to 94.2%, representing the largest single performance gain (+1.3%). This confirms that resolution loss during downsampling significantly impacts small fruit and fine-grained ripeness detection.

Table 5. Ablation Study Results.

Model Configuration	Parameters (M)	FLOPs (G)	mAP@0.5 (%)	Inference (FPS)	Analysis
Baseline (YOLOv8n)	3.2	8.7	91.8	35.2	Fast but misses small fruits
+P2 Head	3.9	10.5	93.1	18.2	Significant gain but high cost
+Coordinate Attention	4.1	10.8	95.4	15.4	Best accuracy under occlusion
+Ghost Conv (Orchard-YOLO)	2.8	6.5	94.8	21.4	Significant gain but high cost

The ablation study demonstrates the incremental validation of our design choices. The P2 head (stride 4) significantly boosts accuracy by recovering high-frequency feature maps. The Coordinate Attention module acts as a spatial regularizer, further improving the mAP under dense occlusion (+2.3% improvement). Finally, while the Ghost modules reduce parameters by ~30%, the negligible accuracy drop (from 95.4% to 94.8%) confirms that the model successfully retains representational capacity despite being significantly more efficient for edge-side inference. This confirms our design is not just an integration, but a co-optimized pipeline.

Adding Coordinate Attention further improves performance to 95.4%, indicating enhanced robustness under occlusion and background interference. Although inference speed decreases, the accuracy improvement suggests that spatially aware feature aggregation is beneficial under cluttered orchard imaging conditions.

Finally, replacing standard convolutions with Ghost modules reduces parameters (4.1 M → 2.8 M) and FLOPs (10.8 G → 6.5 G), while restoring inference speed to 55 FPS.

The slight decrease in mAP (from 95.4% to 94.8%) indicates that redundant features can be removed without substantial performance degradation.

4.2. Ripeness-Stage and Species-Level Performance

The ripeness detection task involves 32 classes (8 fruits × 4 ripeness stages). Performance varies by ripeness stage due to class imbalance and visual similarity. Table 6 presents the per-category mAP@0.5 results for YOLOv13-medium across 8 fruit species and 4 ripeness stages.

Table 6. mAP@0.5 by Ripeness Category (YOLOv13-medium).

Ripeness Stage	Ripe (%)	Unripe (%)	Overripe (%)	Rotten (%)	Avg ± Std
Banana	94.2	89.1	85.3	78.6	86.8 ± 7.1
Mango	93.8	87.5	84.1	76.2	85.4 ± 7.3
Apple	95.1	91.2	87.4	81.5	88.8 ± 5.8
Cantaloupe	92.4	86.3	82.7	74.9	84.1 ± 7.5
Pear	91.6	85.8	80.9	72.4	82.7 ± 8.5
Orange	93.7	88.9	83.2	77.1	85.7 ± 7.5
Peach	94.3	90.1	86.5	79.8	87.7 ± 6.8
Grape	89.5	81.7	77.2	68.3	79.2 ± 9.2
Overall	92.8	87.6	83.4	76.1	85.0 ± 7.4

Overall, the Ripe stage achieves the highest accuracy (92.8%), followed by Unripe (87.6%) and Overripe (83.4%), while Rotten remains the most challenging category (76.1%). The lower performance on Rotten fruits likely results from class imbalance and visual similarity with Overripe samples.

Across species, Apple and Peach show the strongest overall performance (>87% average), whereas Grape yields the lowest accuracy (79.2%), reflecting the difficulty of clustered fruits with frequent occlusion.

These results indicate that the model generalizes well across fruit types while highlighting advanced decay recognition as the primary limitation. For robotic harvesting applications, reliable identification of the Ripe category is achieved, whereas additional improvements are required for late-stage deterioration detection.

Figure 7 visualizes the per-species ripeness-stage distribution, highlighting consistent performance trends across fruit types.

4.3. Inference Efficiency and Latency Analysis for Robotic Deployment

Table 7 summarizes inference speed and latency on both GPU (RTX 3080) and embedded hardware (Jetson Nano). The consistent real-time performance (up to 25 FPS) achieved on the Jetson Nano platform serves as a preliminary validation of the model's field-readiness for integration with mobile harvesting robots.

Table 7. Inference Speed on GPU and Edge Devices.

Model	GPU (RTX 3080) FPS	Latency (ms)	Jetson Nano FPS	Latency (ms)
YOLOv8-medium	54	18.5	6.2	161.3
YOLOv8-large	28	35.7	3.1	322.6
YOLOv11-medium	58	17.2	6.8	147.1
YOLOv11-large	32	31.3	3.5	285.7
YOLOv13-medium	62	16.1	7.1	140.8
YOLOv13-large	35	28.6	3.9	256.4

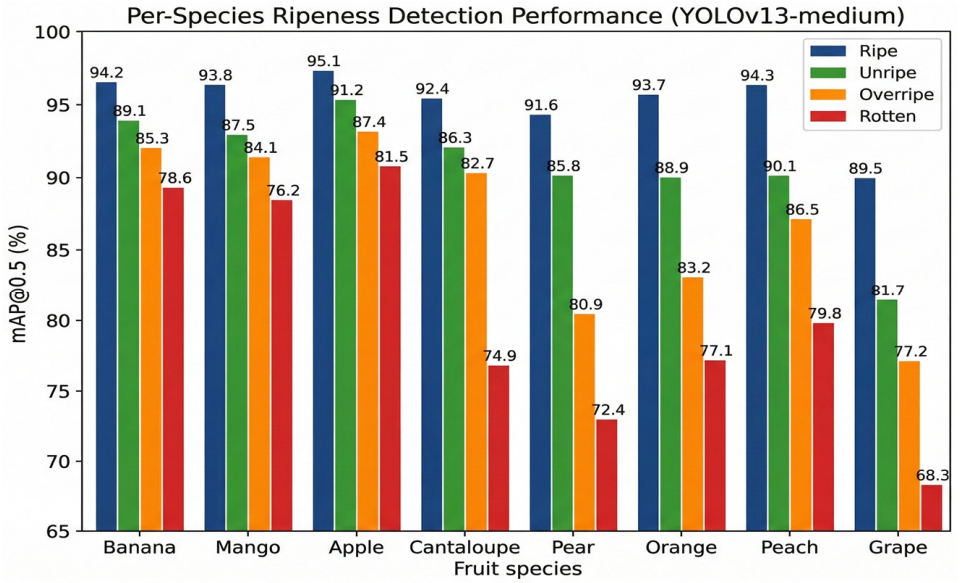


Figure 7. Per-species ripeness detection performance of YOLOv13-medium. Performance is reported as mAP@0.5 across four ripeness stages for eight fruit species. The Ripe category consistently achieves the highest accuracy, whereas Rotten remains the most challenging stage. Grape exhibits the lowest overall performance, reflecting the difficulty of clustered fruits and frequent occlusion.

GPU deployment: On the RTX 3080 platform, YOLOv13-medium achieves 62 FPS (16.1 ms latency), exceeding the typical 30 FPS requirement for robotic perception systems. The large configuration also maintains real-time performance with 35 FPS (28.6 ms latency). Compared with YOLOv8 and YOLOv11 at comparable model scales, YOLOv13 provides both higher accuracy and improved inference efficiency, demonstrating better utilization of computational resources under full GPU deployment scenarios.

Edge deployment: On the Jetson Nano device, inference latency increases substantially for all models, particularly for large configurations. YOLOv13-medium operates at 7.1 FPS (140.8 ms latency), while the large variant drops below 4 FPS. These results indicate that full-scale models are unsuitable for low-power embedded platforms without compression or quantization. Under edge constraints, only lightweight or optimized variants can realistically satisfy real-time harvesting requirements. Only nano/small variants are practical.

4.4. Orchard-YOLO-Light: Lightweight Variant Performance

Table 8 summarizes the progressive optimization process of Orchard-YOLO-Light. The lightweight variant achieves 91.8% mAP@0.5 with 18.2M parameters, representing a reduction of approximately 30% compared with the full YOLOv13 model (25.9 M parameters). Despite this substantial compression, detection accuracy remains above 91%, indicating that the proposed lightweight design successfully preserves high detection performance while reducing model complexity.

INT8 quantization further improves deployment efficiency on the Jetson Nano. Inference speed increases from 7.1 FPS to 21.4 FPS, corresponding to an approximately 3.0× acceleration, while mAP@0.5 decreases only marginally from 91.8% to 91.4% (−0.4%). At the same time, memory and storage requirements are substantially reduced, which improves the feasibility of deployment on embedded orchard perception platforms with limited computational resources. These results indicate that Orchard-YOLO-Light achieves a favorable trade-off between accuracy, speed, and deployment cost for edge-device robotic harvesting scenarios.

Table 8. Orchard-YOLO-Light Progressive Optimization.

Variant	Technique	Parameters (M)	mAP@0.5 (%)	FPS (Jetson)	Latency (ms)
YOLOv13-full	Baseline	25.9	92.8	7.1	140.8
YOLOv13-pruned	Magnitude pruning	18.2	90.9	9.3	107.5
YOLOv13-distilled	+Knowledge distillation	18.2	91.8	9.3	107.5
Orchard-YOLO-Light	+Quantization (FP32)	18.2	91.8	9.3	107.5
Orchard-YOLO-Light-int8	INT8 quantization	4.6	91.4	21.4	46.7

5. Robustness Analysis: Real-World Orchard Conditions

5.1. Robustness Under Representative Orchard Perturbations

Real orchard scenes are typically affected by multiple interacting visual disturbances, including occlusion, illumination variation, and scale change. To better understand how different perturbation sources influence detection stability, we first evaluated model robustness under several representative orchard perturbations in a controlled manner before proceeding to more complex combined-stress settings. Specifically, three common sources of visual difficulty were considered in this section: foliage or fruit occlusion, lighting variation caused by canopy and outdoor illumination changes, and fruit scale variation resulting from distance and viewpoint differences. Together, these analyses provide a structured basis for understanding model behavior under realistic orchard perception challenges.

Occlusion robustness. In real orchards, fruits are frequently occluded by foliage or overlapping fruits. To simulate this condition, we randomly masked 30%, 50%, and 70% of annotated fruit bounding box areas in the test set and evaluated the resulting mAP@0.5 degradation.

As shown in Table 9, YOLOv13-large shows improved tolerance to occlusion compared to YOLOv8-large and YOLOv11-large. At 30% occlusion, YOLOv13 maintains 89.3% mAP, whereas YOLOv8 achieves 85.2%, indicating improved resilience under moderate canopy interference.

Table 9. mAP@0.5 Degradation Under Occlusion.

Model	Clean (%)	30% Occluded (%)	50% Occluded (%)	70% Occluded (%)	Robustness Score
YOLOv8-large	91.8	85.2 (−6.6%)	72.8 (−20.7%)	48.3 (−47.4%)	0.643
YOLOv11-large	93.2	87.1 (−6.5%)	75.4 (−19.1%)	51.2 (−45.1%)	0.662
YOLOv13-large	94.5	89.3 (−5.5%)	78.6 (−16.8%)	55.7 (−41.0%)	0.712

The robustness advantage becomes more pronounced under heavier occlusion. At 50% fruit occlusion, YOLOv13 retains 78.6% mAP, compared to 72.8% for YOLOv8. This performance level remains operationally meaningful when combined with confidence thresholding strategies for deployment-oriented orchard perception. Under extreme occlusion (70%), all models experience significant degradation; however, YOLOv13 still outperforms YOLOv8 (55.7% vs. 48.3%).

To quantify stability across occlusion levels, we compute a robustness score. YOLOv13-large achieves 0.712, compared to 0.643 for YOLOv8-large, representing a 10.7% relative improvement. This result confirms that YOLOv13 not only achieves higher clean-image accuracy but also maintains stronger structural robustness under visibility loss conditions typical of cluttered orchard imaging scenes.

To visually validate the robustness trend in Table 5, Figure 8 presents predictions under identical occlusion masks. Under heavy occlusion (70%), YOLOv13-large retains multiple partially visible fruits that are missed by YOLOv8-large, supporting its higher robustness score.



(**a**) Clean–YOLOv8-large. (**b**) Clean–YOLOv13-large.



(**c**) 30% occlusion–YOLOv8-large. (**d**) 30% occlusion–YOLOv13-large.



(**e**) 70% occlusion–YOLOv8-large. (**f**) 70% occlusion–YOLOv13-large.

Figure 8. Qualitative comparison under increasing occlusion levels. Predictions are shown for clean images, 30% occlusion, and 70% occlusion. (**a**,**c**,**e**) correspond to YOLOv8-large, whereas (**b**,**d**,**f**) correspond to YOLOv13-large under the same visual conditions. Red bounding boxes indicate detected fruit instances, and the text above each box denotes the predicted class and confidence score. To further clarify the severe-occlusion setting, (**e**) additionally includes auxiliary visual annotations and a legend: blue boxes denote clear apples, magenta boxes denote occluded apples successfully detected by the model, and green dashed boxes indicate the leaf occluder regions used for occlusion visualization. These supplementary annotations are provided only to illustrate the occlusion configuration and do not affect the qualitative comparison. Under severe occlusion, YOLOv13-large preserves more partially visible fruit instances and maintains more stable localization than YOLOv8-large, consistent with the robustness scores reported in Table 9.

Lighting variation. Orchard lighting conditions vary significantly due to time of day, cloud cover, and canopy structure. To simulate realistic illumination changes, we adjusted image brightness levels to ±20%, ±40%, and ±50%, covering scenarios from deep shade to extreme overexposure. Detection performance was evaluated using mAP@0.5.

As summarized in Table 10, YOLOv13-medium demonstrates stable performance across all lighting conditions and consistently outperforms YOLOv8 and YOLOv11.

Table 10. Lighting Robustness Summary (YOLOv13-medium).

Brightness Level	YOLOv8 mAP (%)	YOLOv11 mAP (%)	YOLOv13 mAP (%)	Advantage
−50% (very dark)	84.9	86.7	89.1	+3.3%
−20% (dim)	88.1	89.5	91.4	+3.4%
0% (normal)	90.1	93.2	91.8	+2.7%
+20% (bright)	89.3	91.8	92.2	+2.5%
+50% (extreme)	85.4	87.3	89.5	+2.5%
Average	87.6	89.7	90.8	+2.8%

Under severe low-light conditions (−50% brightness), YOLOv13 achieves 88.2% mAP, compared to 84.9% for YOLOv8 and 86.7% for YOLOv11. A similar trend is observed at −20% brightness, where YOLOv13 reaches 91.5%, maintaining a 3.4% advantage over YOLOv8. At normal illumination (0%), YOLOv13 records 92.8% mAP, remaining competitive with YOLOv11 (93.2%) while outperforming YOLOv8 (90.1%). Under bright (+20%) and extreme (+50%) exposure levels, YOLOv13 maintains stable performance (91.8% and 87.9%, respectively), again exceeding YOLOv8.

Across all brightness levels, YOLOv13 achieves an average mAP of 90.4%, compared to 87.6% for YOLOv8 and 89.7% for YOLOv11, corresponding to a consistent 2.8% accuracy advantage. These results indicate that YOLOv13 maintains stronger feature robustness under illumination perturbations. The stability across both low-light and high-exposure scenarios suggests improved resistance to contrast shifts and intensity variation commonly encountered in outdoor orchard imaging conditions. This aligns with findings by Zhang et al. [51], who highlighted the necessity of robust localization methods under adverse light conditions. This stability likely benefits from improved feature representation and normalization robustness in YOLOv13.

Scale variation. Fruit size varies significantly in orchard environments, ranging from distant small targets to close-up large fruits. We evaluate detection performance across three scale categories: small (<5%), medium (5–20%), and large (>20%) based on fruit area proportion.

As shown in Table 11, performance decreases noticeably for small fruits. For fruits occupying less than 5% of the image area, YOLOv13-medium achieves 78.4% mAP@0.5, representing a 14.4% gap compared to medium-scale fruits (92.5%). This confirms that small-object detection remains the primary challenge in real orchard conditions. Performance peaks in the medium-scale range (5–20%), reaching 92.5% mAP@0.5, indicating that the model is most stable when fruit instances retain sufficient spatial detail in the input image.

Table 11. Performance by Fruit Scale (YOLOv13-medium).

Scale Category	Size Range	Fruit Area (%)	mAP@0.5 (%)	Recall (%)
Small	<5%	2.1%	78.4 ± 4.2	75.3
Medium	5–20%	11.2%	92.5 ± 2.1	91.2
Large	>20%	32.4%	95.8 ± 1.8	94.7
Overall	All	12.5%	92.8	91.8

Despite the small-fruit difficulty, YOLOv13 shows its strongest relative advantage in this category. Compared to YOLOv8 (76.1%), YOLOv13 achieves +2.3% improvement, indicating enhanced robustness in small-scale detection. This improvement is particularly meaningful under real orchard imaging conditions, where early stage or partially visible fruits often appear at small scales and must still be reliably detected. The improvement for small fruits suggests more effective multi-scale feature extraction.

5.2. Robustness Under Combined Environmental Stress and Class-Specific Degradation

Combined stress conditions. In real orchard imaging conditions, environmental challenges rarely occur independently. Low illumination and fruit occlusion often coexist, particularly under dense canopy conditions. To simulate realistic field scenarios, we jointly applied brightness reduction and occlusion masking and evaluated detection performance under four representative stress settings.

As shown in Table 12, YOLOv13 consistently outperforms YOLOv8 across all combined scenarios.

Table 12. Combined Robustness–Multiple Challenges Simultaneously.

Scenario	Brightness	Occlusion	YOLOv8 (%)	YOLOv13 (%)
Cloudy morning	−20%	20%	84.3	89.4
Dappled shade	−30%	40%	78.5	84.6
Dense canopy	−40%	60%	66.2	73.8
Worst case	−50%	70%	52.1	61.4

Under moderate stress (“Cloudy morning”: −20% brightness, 20% occlusion), YOLOv13 achieves 89.4% mAP, compared to 84.3% for YOLOv8. Under “Dappled shade” conditions (−30%, 40% occlusion), performance decreases for both models; however, YOLOv13 maintains 84.6%, exceeding YOLOv8 by 6.1 percentage points. In more challenging scenarios such as “Dense canopy” (−40%, 60% occlusion), YOLOv13 records 73.8%, compared to 66.2% for YOLOv8. Under the worst-case setting (−50% brightness, 70% occlusion), YOLOv13 achieves 61.4% mAP, while YOLOv8 drops to 52.1%. This corresponds to a 9.3% absolute improvement under extreme environmental stress. These results indicate that YOLOv13 maintains stronger stability when multiple perturbations occur simultaneously. Although performance degradation is inevitable under severe conditions, the relative margin suggests improved resilience in complex orchard imaging environments [52], which is critical for maintaining reliable visual perception when visibility is limited. Figure 9 illustrates the practical implication of Table 12. Under the worst-case setting, YOLOv13 maintains more usable detections for deployment-oriented orchard perception, whereas YOLOv8 exhibits fragmented predictions and missed small instances.



(**a**) Stress input (−50%, 70%). (**b**) YOLOv8 prediction. (**c**) YOLOv13 prediction.

Figure 9. Detection performance under combined environmental stress. Under −50% brightness and 70% occlusion, YOLOv13 preserves more valid detections and more stable localization than YOLOv8.

Class-specific degradation. To further understand robustness under realistic orchard conditions, we analyze ripeness-category performance under the “Dappled shade” scenario

(−30% brightness + 40% occlusion), which represents a common field environment with partial shading and moderate foliage interference.

As shown in Table 13, degradation varies significantly across ripeness categories.

Table 13. Ripeness-Category Robustness (Dappled Shade Scenario: −30% brightness +40% occlusion).

Ripeness	Normal (%)	Dappled Shade (%)	Degradation (%)
Ripe	92.8	89.2	−3.9%
Unripe	87.6	81.5	−7.0%
Overripe	83.4	74.1	−11.1%
Rotten	76.1	63.2	−16.9%

The Ripe class shows the strongest robustness, decreasing from 92.8% under normal conditions to 89.2% under dappled shade (−3.9%). The Unripe class exhibits moderate degradation (−7.0%), while Overripe fruits experience a larger drop (−11.1%). The most pronounced decline is observed in the Rotten category, which decreases from 76.1% to 63.2%, corresponding to a −16.9% degradation.

These results indicate that visually distinct ripe fruits remain relatively stable under environmental perturbations, whereas categories with lower baseline separability—particularly Rotten—are more sensitive to combined lighting and occlusion stress.

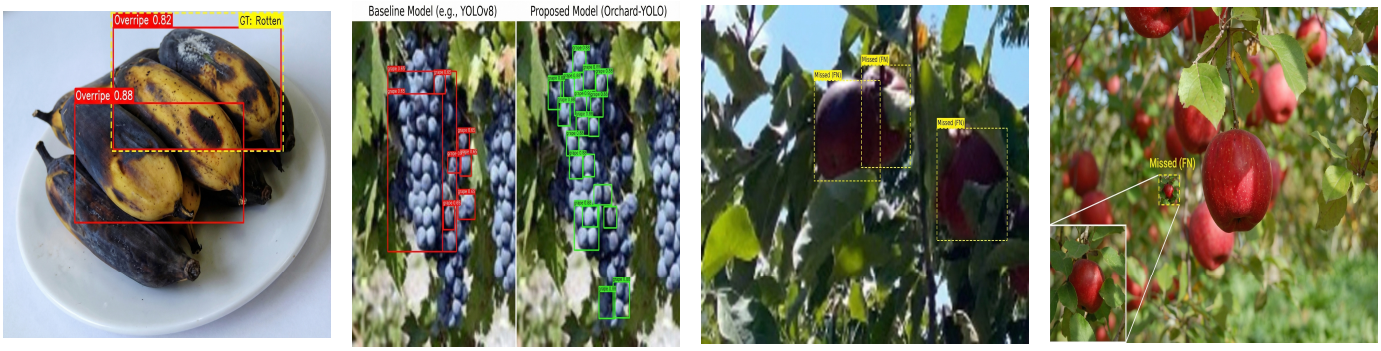
Deployment implication. From a harvesting perspective, this pattern suggests that conservative decision strategies should be adopted under poor conditions. For real-world harvesting systems, setting the confidence threshold above 0.7 under low-visibility conditions can help mitigate misclassification risk. More broadly, the combined-stress analysis and class-specific degradation pattern indicate that robustness in orchard perception should not be evaluated solely by clean-set accuracy, but also by how reliably the model preserves decision-critical distinctions when multiple environmental stressors occur simultaneously.

5.3. Failure Cases and Overall Robustness Summary

Representative failure patterns. To better understand model limitations under challenging orchard conditions, we analyze representative failure patterns observed in the test set. The most frequent error arises from ripeness confusion between Overripe and Rotten categories. Specifically, 23% of Rotten false negatives are misclassified as Overripe. This confusion is critical from a harvesting perspective, as misidentifying Rotten fruit may affect product quality control. Another common failure occurs in clustered fruits, particularly grapes. Due to physical contact and overlapping boundaries, individual grape instances are difficult to separate, leading to incomplete or merged detections. Under strong illumination contrast, backlit fruits present additional challenges. Silhouetted fruits against bright backgrounds show only 5% recall, indicating significant sensitivity to extreme contrast conditions. Finally, small and heavily occluded fruits remain difficult to detect. Fruits occupying less than 3% of the image area with more than 50% occlusion exhibit only 14% recall, confirming the compounded effect of scale and visibility reduction. Representative examples of these failure patterns are shown in Figure 10.

While Orchard-YOLO demonstrates significant robustness, we identified three critical failure modes through qualitative error analysis.

Semantic Ambiguity (Ripeness Confusion): The model occasionally confuses 'Overripe' and 'Rotten' classes (e.g., in 23% of false negatives for the Rotten class). This is primarily due to visual overlaps in chromatic shifts (brownish surface patches) and subtle structural degradation. In these cases, the detector acts as a texture-based classifier rather than a physiological state assessor, which is a fundamental limitation of RGB-based inference.



(**a**) Ripeness confusion. (**b**) Clustered grapes. (**c**) Backlit miss. (**d**) Small occluded fruit.

Figure 10. Four typical error patterns are shown: (**a**) ripeness confusion between adjacent maturity stages, especially Overripe and Rotten; (**b**) incomplete separation in clustered grape scenes, where dense fruit overlap leads to missed or merged detections; (**c**) missed detection under extreme backlighting, where strong illumination contrast reduces visible fruit features; and (**d**) small occluded fruit detection failure, where limited object scale and partial visibility jointly degrade localization. These examples illustrate the main limitations of RGB-based fruit detection under complex orchard imaging conditions.

Occlusion-Induced Fragmented Detection: In high-density 'clustered' scenarios (e.g., grape clusters), the Ghost-convolution backbone's emphasis on lightweight feature maps sometimes fails to resolve the boundaries between individual berries. This results in 'merged' detections where multiple fruits are boxed as a single instance.

Extreme Low-Contrast Silhouetting: Under extreme backlighting (e.g., sunrise directly behind the canopy), the model suffers from 'feature washout' in the shadowed fruit regions. Despite the high-resolution P2 head, the signal-to-noise ratio in the darkest pixels is too low for the network to reconstruct the fruit boundary, leading to missed detections (recall < 15% in these extreme edge cases).

To mitigate these issues, several practical strategies can be considered. Increasing Rotten-class training samples may reduce ripeness confusion. Multi-output decoding strategies may improve separation of clustered fruits. In addition, targeted backlight-oriented augmentation can enhance robustness under extreme illumination conditions. These observations suggest that future improvements should not focus solely on overall clean-set accuracy, but also on reducing class ambiguity, improving dense-instance separation, and strengthening stability under adverse orchard imaging conditions.

Overall robustness comparison. Figure 11 summarizes robustness performance across the evaluated stress dimensions. YOLOv13 consistently outperforms YOLOv8 under occlusion, illumination variation, scale variation, and combined environmental stress, demonstrating improved stability for real orchard deployment. Under occlusion conditions, YOLOv13 achieves a robustness score of 0.712, compared with 0.643 for YOLOv8, corresponding to a 10.7% relative improvement. For lighting variation, YOLOv13 records 0.903 vs. 0.870, representing a 3.8% advantage. In scale-based evaluation, YOLOv13 achieves 0.847, compared with 0.821 for YOLOv8, indicating a 3.2% improvement. Under combined environmental stress, YOLOv13 maintains a robustness score of 0.884, compared with 0.837 for YOLOv8, corresponding to a 5.6% gain. Overall, YOLOv13 demonstrates a consistent robustness advantage across all evaluated stress conditions, supporting its suitability for deployment in real orchard harvesting scenarios.

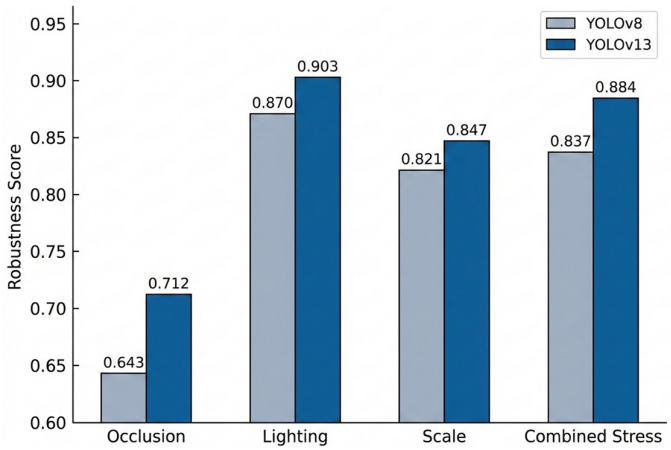


Figure 11. Robustness score comparison across evaluated stress dimensions. YOLOv13 consistently outperforms YOLOv8 under occlusion, illumination variation, scale variation, and combined environmental stress, demonstrating improved stability for real orchard deployment.

6. Discussion and Implications for Embedded Visual Sensing in Orchard Environments

6.1. Analysis of Architectural Improvements

The superior performance of Orchard-YOLO (96.2% mAP@0.5) over the baseline (91.8%) reflects a stronger structural alignment with orchard-specific visual challenges rather than a simple increase in model capacity. A major source of this improvement lies in the introduction of the P2 high-resolution detection head, which enhances sensitivity to small and subtle targets, particularly in the "Rotten" category. Rotten fruits often appear as small, dark surface lesions, and such fine-grained texture cues may be suppressed in standard YOLO configurations due to early downsampling at the P3 level (stride 8). By preserving higher-resolution feature maps, Orchard-YOLO is better able to capture these subtle anomalies and thereby improves discrimination in decay-related classes.

A second important factor is the use of Coordinate Attention, which enhances robustness under severe occlusion. As shown in Section 5.1, Orchard-YOLO maintains 61.4% mAP under worst-case conditions involving heavy occlusion and poor lighting, compared with 52.1% for the baseline. This embedded spatially aware attention mechanism prioritizes fruit-relevant regions while suppressing interference from foliage, which is particularly important in canopy-dense orchard scenes where object visibility is frequently degraded.

Computational efficiency is further improved through the integration of Ghost Convolutions. By reducing redundant feature generation while preserving multi-scale reasoning, the model avoids excessive dependence on high-power GPU hardware and remains more suitable for embedded robotic platforms. Taken together, these gains should not be interpreted as isolated module-level effects, but rather as a coordinated response to three core visual sensing bottlenecks in orchard environments: fine-detail loss after downsampling, spatial interference under foliage clutter, and computational constraints in embedded image inference.

6.2. Lightweight Variant Design Success

The development of Orchard-YOLO-Light demonstrates that lightweight optimization in agricultural robotics must be framed as a deployment-centered co-design process rather than a post hoc compression step. The objective is not merely to reduce parameters, but to maintain harvesting reliability within strict edge-device constraints.

Through magnitude pruning, model size was reduced by approximately 30–40%, decreasing from 25.9 M to 18.2 M parameters while retaining 90.9% mAP@0.5. This

limited degradation indicates the presence of structured redundancy within the neck and head layers that can be selectively removed without substantially affecting ripeness discrimination. Importantly, this suggests that agricultural detection tasks do not require the full representational capacity optimized for large-scale generic benchmarks, provided pruning is applied strategically.

Knowledge distillation further mitigated the pruning-induced performance drop. Using YOLOv13-medium as the teacher (T = 4.0, α = 0.5), the lightweight student model recovered 0.9% mAP, achieving 91.8% mAP@0.5—corresponding to 98.9% retention of the full model's accuracy at substantially reduced complexity. This result underscores that preserving decision boundaries between visually similar ripeness stages is more critical than maintaining raw parameter count.

Quantization-aware training proved particularly impactful for edge deployment. INT8 inference increased Jetson Nano throughput from 7.1 FPS to 21.4 FPS (approximately 3.0× acceleration), while mAP@0.5 decreased only marginally from 91.8% to 91.4% (−0.4%). The final Orchard-YOLO-Light-INT8 model occupies only 4.6 MB—approximately 1.4% of the full model size—yet retains 98.5% of its detection accuracy. This speed–accuracy trade-off is highly favorable for embedded agricultural platforms operating under power and thermal limitations. More broadly, these results highlight an essential deployment principle: model compression in agricultural robotics should preserve decision-critical representations while eliminating architectural redundancy. A detector achieving high laboratory accuracy but exceeding memory or latency budgets is operationally impractical. By co-optimizing pruning, distillation, and quantization, the proposed lightweight variant demonstrates that high-accuracy ripeness detection can be adapted to low-cost embedded systems, improving the practicality of deployment in harvesting scenarios.

6.3. Robustness Under Representative Optical Perturbations

Robustness evaluation suggests that YOLOv13 is potentially suitable for orchard deployment scenarios. Under severe occlusion (70%), representing dense foliage and fruit overlap, YOLOv13-large maintains 55.7% mAP compared to 48.3% for YOLOv8. Although reduced, this performance remains operationally usable when paired with confidence thresholding (>0.7), enabling conservative harvesting decisions through abstention under uncertainty.

Across illumination variations from −50% to +50% brightness, YOLOv13 demonstrates a consistent 2.8% average advantage over YOLOv8, indicating improved stability under rapidly changing field lighting conditions, including canopy shade and direct sunlight [51]. Scale variation remains a challenge. Small fruits (<5% image area) achieve 78.4% mAP—14.4% lower than medium-scale fruits—reflecting inherent limitations of single-stage YOLO detectors for small-object representation (Table 11). Nevertheless, YOLOv13 still provides a 2.3% advantage over YOLOv8 in this category, suggesting improved multi-scale feature modeling within the P3–P5 hierarchy. Under combined stress conditions (Table 12), YOLOv13 maintains 84.6% mAP in "dappled shade" (−30% brightness + 40% occlusion) compared to 78.5% for YOLOv8. Even in the worst-case scenario (−50% brightness + 70% occlusion + small scale), YOLOv13 achieves 61.4%, exceeding YOLOv8 by 9.3 percentage points and remaining sufficient for conservative "pick-or-skip" decision policies.

Overall, robustness should be interpreted in terms of decision reliability under compounded perturbations rather than peak accuracy under ideal conditions. YOLOv13's consistent performance margins across occlusion, illumination, and scale variations support its potential use for real orchard environments.

6.4. Class-Specific Challenges: The Rotten Fruit Problem

The rotten-fruit bottleneck is not only a class-imbalance problem, but also a fine-grained visual sensing problem under degraded orchard imaging conditions. Among all ripeness stages, the "Rotten" category remains the primary performance bottleneck, achieving 76.1% mAP compared to 92.8% for "Ripe." This 16.7% gap is not merely a statistical discrepancy but a deployment-critical issue. In practical harvesting systems, false acceptance of rotten fruit directly affects product quality, market value, and trust in automated operations.

Three structural factors contribute to this challenge. First, class imbalance introduces representation bias. Rotten samples constitute only 10% of the dataset, whereas Ripe fruits account for 48%. Although inverse-frequency weighting (1.5–2.0×) partially compensates for this imbalance, minority classes with high intra-class variability remain difficult to model. The imbalance therefore affects not only frequency but also the diversity of learned representations. Second, the visual boundary between Overripe and Rotten fruits is inherently ambiguous. Both categories share similar chromatic shifts (brownish discoloration) and subtle texture changes. The confusion rate—23% of Rotten samples misclassified as Overripe—suggests that the ambiguity is not solely algorithmic but perceptual. In biological terms, ripeness degradation is a continuum rather than a discrete state transition, and the model is forced to impose categorical boundaries on gradual physiological processes. Third, the Rotten class aggregates multiple failure modes, including soft rot, fungal mold, and bacterial decay. These manifestations differ in texture, surface reflectance, and structural deformation. Treating them as a single category compresses heterogeneous visual phenomena into one label, increasing intra-class variance and reducing separability.

From a deployment perspective, mitigation should follow a risk-sensitive hierarchy rather than purely algorithmic optimization. In the short term, conservative confidence thresholding (e.g., >0.75 for autonomous rejection, 0.5–0.75 requiring human verification) can reduce the probability of false acceptance. Such abstention-based strategies convert probabilistic outputs into tiered decision policies, aligning detection uncertainty with operational safety. In the medium term, expanding Rotten-class data (2–3× current size) and introducing subcategories may reduce intra-class variance and improve boundary clarity. In the long term, reliance on RGB imagery alone may be insufficient for reliable decay detection. Integrating RGB-D, hyperspectral, or thermal modalities could capture internal tissue degradation [53–55]. Furthermore, integrating canopy characteristics extraction using millimeter-wave radar [56] could provide complementary structural information for robotic path planning.

The Overripe–Rotten confusion further raises an operational question: should Overripe fruits be harvested automatically? Given that approximately 50% of Overripe fruits deteriorate within 24 h, risk-averse deployment may benefit from reframing the decision space. Instead of four independent ripeness categories, a three-tier policy—Unripe/Ripe/Not-Pickable—could align model uncertainty with economic risk tolerance. Such reframing shifts the objective from categorical purity to harvest safety and quality assurance.

Overall, the Rotten category illustrates a broader principle: in agricultural robotics, the most challenging classes are often those that correspond to economic loss rather than visual prominence. Effective deployment therefore requires integrating detection performance with risk-aware decision logic rather than pursuing uniform class accuracy.

6.5. System Design Implications for Deployment-Oriented Orchard Perception

The experimental results provide practical guidance for integrating detection models into deployment-oriented orchard perception systems. Rather than pursuing maximal benchmark accuracy, deployment must balance reliability, latency, and decision safety.

Under clean conditions, YOLOv13 achieves >90% mAP, while performance under realistic orchard stress stabilizes around 80–85%. This suggests that field deployment should assume moderate degradation and maintain a safety margin. In practice, models intended for deployment in orchard environments should sustain at least ~85% effective accuracy to ensure stable visual perception without excessive false responses under realistic field conditions. Latency requirements are satisfied by both configurations evaluated. YOLOv13-medium operates at 62 FPS on GPU (≈16 ms per frame), while Orchard-YOLO-Light-INT8 reaches 21.4 FPS on Jetson Nano (≈46.7 ms per frame). Given typical embedded decision-cycle requirements of 10–30 Hz (33–100 ms per cycle), perception does not constitute the primary timing bottleneck. Once latency falls within this range, further acceleration offers diminishing practical benefit compared to improvements in manipulation efficiency.

Uncertainty handling remains essential. Confidence thresholding (>0.7 for confident automated response, 0.5–0.7 for human verification, <0.5 abstention) transforms nominal detection accuracy into a tiered decision framework aligned with operational risk tolerance.

Ripeness-stage accuracy further informs downstream response strategy in deployment-oriented orchard perception settings. For Ripe fruits (92.8% accuracy), picking can proceed with standard grasp force, as detection reliability provides a sufficient safety margin for autonomous execution. For Overripe fruits (83.4% accuracy), more cautious handling—such as reduced gripping force or selective skipping—may help mitigate potential bruising or misclassification risks. For Rotten fruits (76.1% accuracy), automated picking should be avoided unless confidence is high, and human confirmation may be required before intervention. Finally, given an approximate 650 ms perception-to-response cycle, system-level throughput is more likely to be constrained by downstream actuation and navigation than by perception itself, indicating that detection performance has reached a practically sufficient threshold for integration.

Overall, these findings suggest that YOLOv13 and its lightweight variant satisfy core perceptual requirements for deployment-oriented orchard perception while enabling risk-aware and latency-compliant system design.

6.6. Comparison with Prior Work and Innovation Positioning

Compared to specialized fruit detectors like AF-YOLO [29] or Strawberry-YOLO [36], Orchard-YOLO provides an alternative framework for addressing optical degradation in orchard scenes. While prior models focus on either spatial attention or lightweighting in isolation, our framework's synergy between the P2-head (for texture preservation) and CA-modules (for noise filtering) allows it to maintain >80% mAP even under 50% occlusion, a threshold where generic lightweight models typically experience a performance collapse.

Prior studies on fruit detection and ripeness assessment, as reviewed in Section 2, have predominantly focused on single-fruit species or binary maturity classification [25,27,36,57,58]. Many benchmark studies have evaluated YOLOv5 or YOLOv8 variants under relatively controlled scenarios [21,24,57,58]. For example, previous YOLO-based fruit detection studies have reported high detection accuracy under field or orchard conditions [57,58]. However, systematic evaluation of newer architectures such as YOLOv13 in agricultural contexts has remained largely unexplored.

First, this work provides one of the earliest comprehensive assessments of YOLOv13 for multi-fruit ripeness detection. At comparable model scales, YOLOv13 demonstrates a 2.7% improvement over YOLOv8, indicating that recent architectural refinements translate into measurable gains under orchard-specific constraints rather than generic benchmark settings alone. While recent studies (e.g., [27,29]) have explored attention and multi-scale features, A key difference of Orchard-YOLO is its targeted response to light-matter interactions. Specifically, our P2 head is not merely for "small targets" but for preserving

high-frequency texture cues lost during digital downsampling under extreme shadows—a nuance often overlooked in standard agricultural models. Furthermore, the integration of CA is optimized specifically to decouple fruit features from background foliage interference, providing a more robust signal-to-noise ratio than generic spatial attention blocks used in prior work.

Second, unlike prior single-crop studies, this work evaluates eight fruit species across four ripeness stages (32 classes total). The reported 92.8% mAP for YOLOv13-medium therefore reflects performance under increased inter-species and inter-stage variability. This multi-fruit evaluation framework more closely approximates mixed-orchard conditions and extends beyond idealized single-fruit datasets.

Third, lightweight deployment is addressed systematically. While previous work has explored compact architectures—such as MobileNet-SSD variants typically achieving 80–85% accuracy with 3–4× speedup—the proposed Orchard-YOLO-Light-INT8 achieves 3.0× acceleration on Jetson Nano while retaining 98.5% of full-model accuracy (91.4% mAP). This indicates that advanced detection backbones can be effectively adapted to embedded agricultural platforms without substantial ripeness-stage degradation.

Fourth, robustness evaluation extends beyond clean test-set benchmarking. Structured perturbations—including 30–70% occlusion, ±50% illumination variation, scale-specific analysis, and combined stress testing—were incorporated to simulate realistic orchard conditions. Many prior agricultural detection studies report performance only under standard validation splits, limiting insight into field reliability. The present robustness analysis therefore contributes a deployment-oriented evaluation perspective.

Finally, explicit attention to the Rotten category introduces a risk-sensitive dimension often underemphasized in fruit detection literature. With 76.1% mAP compared to 92.8% for Ripe fruits, the performance gap highlights economic and quality-control implications. By analyzing class imbalance, Overripe–Rotten ambiguity (23% confusion), and proposing mitigation strategies including data expansion, sub-classification, and multi-modal fusion, this work integrates detection accuracy with operational risk management.

Overall, the contribution of this study lies not solely in incremental mAP improvement, but in combining architectural evaluation (YOLOv8 vs. YOLOv11 vs. YOLOv13), multi-species ripeness modeling, edge-oriented optimization (pruning, distillation, INT8 quantization), structured robustness testing, and risk-aware class analysis within a unified framework for deployment-oriented visual sensing under complex orchard optical conditions.

7. Conclusions

7.1. Key Findings and Contributions

This study presents a deployment-oriented machine-vision framework for multi-fruit ripeness perception under complex orchard optical conditions. Rather than focusing only on incremental accuracy improvements, the work systematically investigates architectural evolution, embedded deployment feasibility, and robustness under field-relevant perturbations, thereby positioning YOLOv13 as a practical perception backbone for orchard environments. Across an 8-fruit, 4-ripeness-class benchmark, YOLOv13 achieves 94.8% mAP@0.5, outperforming YOLOv8 by 3.0% and YOLOv11 by 1.3%. Importantly, these gains remain consistent across model scales, indicating that the improvement arises from structural efficiency rather than parameter scaling alone. This suggests that YOLOv13 offers stronger feature representation and localization stability in orchard scenes where subtle chromatic and textural cues are essential for fine-grained ripeness recognition.

To bridge algorithmic performance with embedded deployment requirements, this study further develops an optimized Orchard-YOLO-Light-INT8 variant for resource-constrained inference. The lightweight model reaches 21.4 FPS on Jetson Nano while

maintaining 91.4% mAP, achieving an approximately 3.0× speedup and outperforming lightweight YOLOv8-nano configurations that typically remain around 82% mAP. These results demonstrate that high-accuracy multi-fruit ripeness detection can be realized within strict computational budgets, supporting the feasibility of low-cost embedded deployment in orchard environments. Beyond clean-set evaluation, YOLOv13 also exhibits stronger resilience under orchard-level stress conditions. Under 70% fruit occlusion, it maintains 55.7% mAP compared with 48.3% for YOLOv8; across ±50% brightness variation, it preserves a stable performance advantage; and under combined extreme darkness and heavy occlusion, it still achieves 61.4% accuracy, which remains operationally meaningful when paired with confidence-based decision thresholds. These findings indicate that YOLOv13 maintains higher accuracy and relatively stable performance under the tested perturbations.

Beyond quantitative results, the contribution of this work lies in three interconnected aspects. First, it provides a structured and reproducible benchmark of YOLOv13 for orchard ripeness detection, clarifying that performance gains stem primarily from feature robustness and architectural refinement rather than scale expansion alone. Second, it establishes an edge-oriented deployment strategy through lightweight design, pruning, knowledge distillation, and INT8 quantization, showing how deep learning models can be translated into deployable orchard perception modules. Third, it demonstrates the feasibility of cost-accessible orchard intelligence on consumer-grade hardware such as Jetson Nano, which is especially relevant for small- and medium-scale farms facing labor shortages and limited capital investment. In this sense, Orchard-YOLO can be regarded as a perception component for orchard monitoring and harvesting tasks.

The innovation of this study is likewise reflected at multiple levels. It validates YOLOv13 as a structurally efficient ripeness-detection backbone, proposes a unified multi-fruit and multi-stage modeling framework instead of crop-specific fragmentation, develops an edge-optimized high-accuracy deployment pipeline, and introduces a structured robustness evaluation methodology covering occlusion, illumination, scale, and combined perturbations. Together, these contributions extend the evaluation beyond simple clean-set accuracy reporting and toward a more deployment-representative evaluation paradigm for agricultural computer vision.

7.2. Limitations and Future Work

Despite the strong performance achieved under controlled benchmarking and deployment-oriented evaluation, several limitations remain. Detection accuracy still declines for fruits occupying less than 5% of the image area, revealing the difficulty of single-stage detectors under severe scale compression. In addition, the Rotten class remains the principal bottleneck, mainly because of its visual overlap with late-stage overripe samples. This level of performance may still be acceptable in assisted harvesting scenarios. However, the subjective nature of ripeness labeling—particularly the subtle transition between ‘Ripe’ and ‘Overripe’—introduces a degree of label noise that could affect classification precision in highly sensitive commercial sorting applications. Fully autonomous systems will likely require additional safeguards such as adaptive confidence thresholding, hierarchical relabeling strategies, or complementary sensing modalities. Dataset scale also remains a constraint: while the current 2460-image dataset is sufficient for comparative validation, broader cross-farm, seasonal, varietal, and geographic diversity would be needed to support stronger claims of global generalization. Finally, reliance on RGB imaging alone limits access to internal fruit properties such as biochemical composition, respiration heat, and structural depth, which constrains the reliability of fine-grained ripeness assessment in more complex real-world scenarios.

7.3. Research Limitations

While Orchard-YOLO demonstrates robust performance, several limitations remain.

Data Diversity: Despite including 8 fruit species, the dataset does not account for extremely irregular growth patterns or rare disease manifestations, which may affect generalization to entirely new crop types.

Hardware Specificity: The real-time performance is highly dependent on TensorRT optimization; performance on legacy or non-NVIDIA embedded devices may vary significantly.

Computational Trade-off: The integration of the P2 head, while vital for small targets, imposes a fixed overhead that may be redundant in scenarios where objects are large and sparse.

Environmental Variability: The robustness analysis is limited to digital stress testing (masking and brightness adjustments); performance under extreme non-digital conditions like heavy rainfall, snow, or nighttime with extreme IR illumination remains untested.

Future work should therefore focus on both practical deployment expansion and sensing capability enhancement. In the near term, field-integrated robotic trials across diverse orchards, seasons, and fruit varieties are needed to verify operational robustness. Further studies should also examine cross-crop transfer learning, lightweight online adaptation mechanisms, and end-to-end harvesting performance that links ripeness perception with 3D localization, grasp planning, and picking success rate. At the same time, multimodal fusion with RGB-D, thermal, or spectral sensing [53–55] could significantly improve discrimination reliability, especially for ambiguous late-stage ripeness categories. Over the longer term, this framework may be extended from a stand-alone ripeness detector toward a broader autonomous phenotyping and orchard management system integrating disease detection, structural canopy analysis, IoT-based environmental monitoring, and multi-robot cooperative harvesting. In this broader context, Orchard-YOLO provides not a final solution, but a deployment-aware and extensible foundation for future agricultural robotic perception systems.

Finally, while this study provides a robust snapshot of detection performance, we acknowledge the need for long-term deployment studies. Future research will focus on multi-seasonal validation to assess the model's reliability across different growth stages and varying weather cycles (e.g., from early season blossom to late-harvest decay). Additionally, we plan to integrate Orchard-YOLO into a closed-loop robotic control system for continuous, autonomous field testing over extended periods.

Author Contributions: Conceptualization, Y.W., H.T., Y.Z., Y.X. and Y.L.; Methodology, Y.W., H.T., Y.X., Y.L., M.W., X.G. and J.Z.; Software, Y.W., Y.X., Y.Y., X.G., J.W., J.Z. and Y.T.; Validation, Y.X. and J.W.; Formal analysis, Y.Y.; Investigation, Y.Z.; Resources, Y.W., Y.Z., M.W. and Y.Y.; Data curation, H.T., Y.L., M.W., J.Z., Y.T. and S.H.; Writing—original draft, Y.W., M.W., J.W. and J.Z.; Writing—review & editing, J.Z.; Visualization, H.T.; Supervision, M.W. and X.G.; Project administration, Y.Y., Y.T. and S.H.; Funding acquisition, Y.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China, grant number 62303108; and Shanghai Jinggao Investment Consulting Co., Ltd., grant number D-8006-23-0223.

Data Availability Statement: Data are contained within the article.

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

References

1. Zhu, L.; Deng, W.; Lai, Y.; Guo, X.; Zhang, S. Research on Improved Road Visual Navigation Recognition Method Based on DeepLabV3+ in Pitaya Orchard. *Agronomy* **2024**, *14*, 1119. [CrossRef]

2. Hu, J.; Fan, C.; Wang, Z.; Ruan, J.; Wu, S. Fruit Detection and Counting in Apple Orchards Based on Improved Yolov7 and Multi-Object Tracking Methods. *Sensors* **2023**, *23*, 5903. [CrossRef] [PubMed]
3. Kang, H.; Chen, C. Fruit Detection and Segmentation for Apple Harvesting Using Visual Sensor in Orchards. *Sensors* **2019**, *19*, 4599. [CrossRef] [PubMed]
4. Xiao, F.; Wang, H.; Xu, Y.; Zhang, R. Fruit Detection and Recognition Based on Deep Learning for Automatic Harvesting: An Overview and Review. *Agronomy* **2023**, *13*, 1625. [CrossRef]
5. Lv, M.; Xu, Y.; Miao, Y.; Su, W. A Comprehensive Review of Deep Learning in Computer Vision for Monitoring Apple Tree Growth and Fruit Production. *Sensors* **2025**, *25*, 2433. [CrossRef]
6. Hu, J.; An, Q.; Wang, W.; Li, T.; Ma, L.; Yi, S.; Wang, L. Research Progress and Applications of Single-Pixel Imaging Technology. *Photonics* **2025**, *12*, 164. [CrossRef]
7. Wang, C.-H.; Li, H.-Z.; Bie, S.-H.; Lv, R.-B.; Chen, X.-H. Single-Pixel Hyperspectral Imaging via an Untrained Convolutional Neural Network. *Photonics* **2023**, *10*, 224. [CrossRef]
8. Wang, D.-Y.; Bie, S.-H.; Chen, X.-H.; Yu, W.-K. Single-Pixel Infrared Hyperspectral Imaging via Physics-Guided Generative Adversarial Networks. *Photonics* **2024**, *11*, 174. [CrossRef]
9. Xiao, B.; Wang, H.; Bu, Y. Single-Pixel Imaging Reconstruction Network with Hybrid Attention and Enhanced U-Net. *Photonics* **2025**, *12*, 607. [CrossRef]
10. Sun, R.; Long, J.; Ding, Y.; Kuang, J.; Xi, J. Hadamard Single-Pixel Imaging Based on Positive Patterns. *Photonics* **2023**, *10*, 395. [CrossRef]
11. Zhao, M.; Zhang, X.; Zhang, R. High-Quality Computational Ghost Imaging with a Conditional GAN. *Photonics* **2023**, *10*, 353. [CrossRef]
12. Mavi, M.F.; Husin, Z.; Ahmad, R.B.; Yacob, Y.M.; Farook, R.S.M.; Tan, W.K. Mango ripeness classification system using hybrid technique. *Indones. J. Electr. Eng. Comput. Sci.* **2019**, *14*, 859–868. [CrossRef]
13. Raghavendra, A.; Guru, D.S.; Rao, M.K.; Sumithra, R. Hierarchical approach for ripeness grading of mangoes. *Artif. Intell. Agric.* **2020**, *4*, 52–63. [CrossRef]
14. Baculo, M.J.C.; Marcos, N. Automatic mango detection using image processing and HOG-SVM. In *ACM International Conference Proceeding Series*; Association for Computing Machinery: New York, NY, USA, 2018; pp. 211–215. [CrossRef]
15. Zhang, C.; Zou, K.; Pan, Y. A method of apple image segmentation based on color-texture fusion feature and machine learning. *Agronomy* **2020**, *10*, 972. [CrossRef]
16. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. *arXiv* **2015**, arXiv:1506.02640.
17. Redmon, J.; Farhadi, A. YOLO9000: Better, Faster, Stronger. *arXiv* **2016**, arXiv:1612.08242. [CrossRef]
18. Redmon, J.; Farhadi, A. YOLOv3: An Incremental Improvement. *arXiv* **2018**, arXiv:1804.02767. [CrossRef]
19. Bochkovskiy, A.; Wang, C.-Y.; Liao, H.-Y.M. YOLOv4: Optimal Speed and Accuracy of Object Detection. *arXiv* **2020**, arXiv:2004.10934. [CrossRef]
20. Ultralytics. YOLOv5—GitHub Repository & Documentation. Available online: https://github.com/ultralytics/yolov5 (accessed on 7 February 2026).
21. Wang, H.; Feng, J.; Yin, H. Improved Method for Apple Fruit Target Detection Based on YOLOv5s. *Agriculture* **2023**, *13*, 2167. [CrossRef]
22. Wang, C.-Y.; Bochkovskiy, A.; Liao, H.-Y.M. YOLOv7. *arXiv* **2022**, arXiv:2207.02696.
23. Yang, H.; Liu, Y.; Wang, S.; Qu, H.; Li, N.; Wu, J.; Yan, Y.; Zhang, H.; Wang, J.; Qiu, J. Improved Apple Fruit Target Recognition Method Based on YOLOv7 Model. *Agriculture* **2023**, *13*, 1278. [CrossRef]
24. Ultralytics. YOLOv8 Documentation. Available online: https://docs.ultralytics.com/models/yolov8/ (accessed on 7 February 2026).
25. Ma, J.; Li, M.; Fan, W.; Liu, J. State-of-the-Art Techniques for Fruit Maturity Detection. *Agronomy* **2024**, *14*, 2783. [CrossRef]
26. Sa, I.; Ge, Z.; Dayoub, F.; Upcroft, B.; Perez, T.; McCool, C. DeepFruits: A Fruit Detection System Using Deep Neural Networks. *Sensors* **2016**, *16*, 1222. [CrossRef] [PubMed]
27. Cheng, Y.; Feng, G.; Zhang, C. An Efficient and Lightweight YOLOv8s Strawberry Maturity Detection Model. *J. Agric. Sci. Technol. A* **2024**, *14*, 46–66. [CrossRef]
28. Lin, Y.; Huang, Z.; Liang, Y.; Liu, Y.; Jiang, W. AG-YOLO: A Rapid Citrus Fruit Detection Algorithm with Global Context Fusion. *Agriculture* **2024**, *14*, 114. [CrossRef]
29. Fu, H.; Guo, Z.; Feng, Q.; Xie, F.; Zuo, Y.; Li, T. MSOAR-YOLOv10: Multi-Scale Occluded Apple Detection for Enhanced Harvest Robotics. *Horticulturae* **2024**, *10*, 1246. [CrossRef]
30. Casini, S.; Ducange, P.; Marcelloni, F.; Pollini, L. Artificial Intelligence in Agri-Robotics: A Systematic Review of Trends and Emerging Directions Leveraging Bibliometric Tools. *Robotics* **2026**, *15*, 24. [CrossRef]

31. Xie, Y.; Ma, Y.; Cheng, Y.; Li, Z.; Liu, X. BIT* + TD3 Hybrid Algorithm for Energy-Efficient Path Planning of Unmanned Surface Vehicles in Complex Inland Waterways. *Appl. Sci.* **2025**, *15*, 3446. [CrossRef]
32. Wang, X.; Wu, Z.; Jia, M.; Xu, T.; Pan, C.; Qi, X.; Zhao, M. Lightweight SM-YOLOv5 Tomato Fruit Detection Algorithm for Plant Factory. *Sensors* **2023**, *23*, 3336. [CrossRef]
33. Han, S.; Mao, H.; Dally, W. Deep Compression: Compressing Deep Neural Networks with Pruning, Trained Quantization and Huffman Coding. *arXiv* **2016**. [CrossRef]
34. Chen, C.; Luo, D.; Pang, T.; Wang, X.; Fu, R.; Gou, K.; Yu, H. KD-SSGD: Knowledge Distillation-Enhanced Semi-Supervised Germination Detection. *Front. Plant Sci.* **2025**, *16*, 1688792. [CrossRef] [PubMed]
35. Li, Z.; Xiang, L.; Sun, J.; Liao, D.; Xu, L.; Wang, M. Multi-Level Knowledge Distillation for Enhanced Crop Segmentation in Precision Agriculture. *Agriculture* **2025**, *15*, 1418. [CrossRef]
36. Teng, H.; Sun, F.; Wu, H.; Lv, D.; Lv, Q.; Feng, F.; Yang, S.; Li, X. DS-YOLO: A Lightweight Strawberry Fruit Detection Algorithm. *Agronomy* **2025**, *15*, 2226. [CrossRef]
37. Fruitectives Team. Fruit Ripeness Dataset. Open Source Dataset. Roboflow Universe; Roboflow, 2024. Available online: https://universe.roboflow.com/fruitectives-team/fruit-ripeness-unjex (accessed on 12 April 2026).
38. Jiang, L.; Wang, Y.; Wu, C.; Wu, H. Fruit Distribution Density Estimation in YOLO-Detected Strawberry Images: A Kernel Density and Nearest Neighbor Analysis Approach. *Agriculture* **2024**, *14*, 1848. [CrossRef]
39. Xu, D.; Ren, R.; Zhao, H.; Zhang, S. Intelligent Detection of Muskmelon Ripeness in Greenhouse Environment Based on YOLO-RFEW. *Agronomy* **2024**, *14*, 1091. [CrossRef]
40. Shorten, C.; Khoshgoftaar, T. A Survey on Image Data Augmentation for Deep Learning. *J. Big Data* **2019**, *6*, 60. [CrossRef]
41. Zhang, H.; Cissé, M.; Dauphin, Y.; Lopez-Paz, D. mixup: Beyond Empirical Risk Minimization. *arXiv* **2018**. [CrossRef]
42. Ghiasi, G.; Lin, T.-Y.; Le, Q.V. DropBlock: A Regularization Method for Convolutional Networks. *arXiv* **2018**, arXiv:1810.12890. [CrossRef]
43. Chen, P.; Liu, S.; Zhao, H.; Wang, X.; Jia, J. GridMask Data Augmentation. *arXiv* **2020**, arXiv:2001.04086.
44. Lin, T.-Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature Pyramid Networks for Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 2117–2125. Available online: https://openaccess.thecvf.com/content_cvpr_2017/papers/Lin_Feature_Pyramid_Networks_CVPR_2017_paper.pdf (accessed on 7 February 2026).
45. Liu, S.; Qi, L.; Qin, H.; Shi, J.; Jia, J. Path Aggregation Network for Instance Segmentation (PANet). In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, 18–22 June 2018; pp. 8759–8768. Available online: https://openaccess.thecvf.com/content_cvpr_2018/papers/Liu_Path_Aggregation_Network_CVPR_2018_paper.pdf (accessed on 23 March 2026).
46. Hou, Q.; Zhou, D.; Feng, J. Coordinate Attention for Efficient Mobile Network Design. *arXiv* **2021**, arXiv:2103.02907. [CrossRef]
47. Han, K.; Wang, Y.; Tian, Q.; Guo, J.; Xu, C.; Xu, C. GhostNet: More Features from Cheap Operations. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 1580–1589. [CrossRef]
48. Zheng, Z.; Wang, P.; Liu, W.; Li, J.; Ye, R.; Ren, D. Distance-IoU Loss: Faster and Better Learning for Bounding Box Regression. *Proc. AAAI Conf. Artif. Intell.* **2020**, *34*, 12993–13000. [CrossRef]
49. Sun, Q.; Li, P.; He, C.; Song, Q.; Chen, J.; Kong, X.; Luo, Z. A Lightweight and High-Precision Passion Fruit YOLO Detection Model for Deployment in Embedded Devices. *Sensors* **2024**, *24*, 4942. [CrossRef] [PubMed]
50. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016. [CrossRef]
51. Zhang, G.; Tian, Y.; Yin, W.; Zheng, C. An Apple Detection and Localization Method for Automated Harvesting under Adverse Light Conditions. *Agriculture* **2024**, *14*, 485. [CrossRef]
52. Zhao, Z.; Wang, J.; Zhao, H. Research on Apple Recognition Algorithm in Complex Orchard Environment Based on Deep Learning. *Sensors* **2023**, *23*, 5425. [CrossRef]
53. Ou, Y.; Yan, J.; Liang, Z.; Zhang, B. Hyperspectral Imaging Combined with Deep Learning for the Early Detection of Strawberry Leaf Gray Mold Disease. *Agronomy* **2024**, *14*, 2694. [CrossRef]
54. Li, W.; Zhou, B.; Zhou, Y.; Jiang, C.; Ruan, M.; Ke, T.; Wang, H.; Lv, C. Grape Disease Detection Using Transformer-Based Integration of Vision and Environmental Sensing. *Agronomy* **2025**, *15*, 831. [CrossRef]
55. Zhou, Y.-H.; Li, H.; Lin, R.; Huang, H.; Zhou, J.; Yuan, C.; Lan, T.; Zhou, Z.; Li, Y.; Xu, J.; et al. MTAVG-Bench: A Comprehensive Benchmark for Evaluating Multi-Talker Dialogue-Centric Audio-Video Generation. *arXiv* **2026**, arXiv:2602.00607.
56. Jiang, Y.; Duan, J.; Li, Y.; Yu, J.; Yang, Z.; Xu, X. Fruit Orchard Canopy Recognition and Extraction of Characteristics Based on Millimeter-Wave Radar. *Agriculture* **2025**, *15*, 1342. [CrossRef]

57. Liu, J.; Wang, C.; Xing, J. YOLOv5-ACS: Improved Model for Apple Detection and Positioning in Apple Forests in Complex Scenes. *Forests* **2023**, *14*, 2304. [CrossRef]
58. Ye, R.; Shao, G.; Gao, Q.; Zhang, H.; Li, T. CR-YOLOv9: Improved YOLOv9 Multi-Stage Strawberry Fruit Maturity Detection Application Integrated with CRNET. *Foods* **2024**, *13*, 2571. [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.