agronomy

MDPI

Article

Missing-Depth-Tolerant Oriented Object Detection for Day-Night Apple-Harvesting Perception

Yinxiang Xu [1], Hang Wang [1], Zhanbo Gao [1], Shuhe Meng [2], Hongjian Zhang [1], Linlin Sun [1], Linlong Jing [1], Xinpeng Cao [1], Yongxian Wang [1], Shuo Kang [1,3,*] and Wei Pan [1,*]

[1] College of Mechanical and Electrical Engineering, Shandong Agricultural University, Taian 271018, China; xuyinxiang666@gmail.com (Y.X.)
[2] College of Mechanical and Electrical Engineering, Qingdao University, Qingdao 266071, China
[3] Shandong Key Laboratory of Intelligent Production Technology and Equipment for Facility Horticulture, Taian 271018, China
* Correspondence: kangshuo@sdau.edu.cn (S.K.); panwei@sdau.edu.cn (W.P.)

Abstract

Apple-harvesting perception must remain usable when depth measurements are unavailable. This study evaluates a single four-channel YOLOv8s-OBB detector that accepts RGB-D input when depth is available and RGB-Z input, with a zero-filled fourth channel, when depth is absent. Training-only input-level modality dropout exposes the detector to both sensing states without changing the inference graph. Apples are assigned to three operational visual categories—ripe_pickable, ripe_unpickable, and unripe—and oriented bounding boxes provide image-plane geometric cues. On an image-level split of 542 orchard images, p_{drop} = 0.5 gave the highest observed full-test mAP_{50} and mAP_{50-95} among the tested settings, at 0.7199 and 0.5380. Under full-test forced RGB-Z input, mAP_{50} was 0.7064, compared with 0.5906 without modality dropout and 0.6952 for the RGB-only baseline. On the paired night-time subset, mAP_{50} curve AUCs under three graded missing-depth masks were 0.7225–0.7355, compared with 0.6088–0.6444 without dropout. These results demonstrate empirical missing-depth tolerance within a single detector under the evaluated conditions. The dropout setting is empirical, and the results are single-run point estimates. Daytime images lack paired depth, so day/night comparisons do not isolate illumination robustness; the image-level split also does not establish scene-independent generalization. OBB outputs are image-plane cues, and the exploratory geometry-only simulation does not validate robotic grasping or field harvesting success.

Keywords: RGB-D sensing; modality dropout; multimodal fusion; sensor-failure robustness; orchard robotics

Check for updates

Academic Editors: Maofang Gao and Dimitrios Kateris

Received: 15 August 2026
Revised: 19 September 2026
Accepted: 26 September 2026
Published: 8 October 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

Robotic apple harvesting is being developed to reduce seasonal labor pressure and to improve the consistency of fruit picking in commercial orchards, but the perception module must first identify harvestable targets and provide action-relevant spatial cues in unstructured canopies [1–4]. Apples are frequently occluded by leaves, branches, stems, bags, and neighboring fruits, so the visible fruit contour may be incomplete, truncated, or visually merged with the background. Recent apple-detection studies in clustered and in situ orchard scenes show that occlusion, fruit overlap, and variable visibility remain central barriers to reliable target perception [5–8]. For an autonomous harvester, such errors are not

only image-level localization failures; they can affect target selection, approach-direction reasoning, collision avoidance, and subsequent manipulation feasibility.

Illumination variation forms a second practical limitation for orchard perception. Daytime images may contain strong shadows, backlighting, overexposed regions, and color shifts caused by canopy structure, whereas night-time operation relies on artificial illumination and often produces different contrast, reflection, and noise patterns. Recent low-light and night-time apple studies further indicate that degraded edge cues, non-uniform exposure, and enhancement quality can influence localization stability [9,10]. A perception module intended for smart agricultural deployment should therefore remain useful across day-night scenes, while also acknowledging that illumination robustness and sensor availability are difficult to separate in real orchard datasets.

RGB-D sensing can complement appearance with geometric information when fruit color, boundary contrast, or visible texture are unreliable. However, depth information may be degraded or unavailable because of sensing range, reflective surfaces, calibration drift, synchronization error, occlusion, or hardware disturbance. RGB-D apple and fruit-perception studies demonstrate the value of color-depth complementarity when depth is available [11–13], and recent multimodal fruit detectors show that fusion architectures can improve perception in complex scenes under complete multimodal input [14,15]. This complete-depth assumption is fragile in orchard deployment, where a harvesting perception module should exploit depth when it exists while remaining compatible with RGB-only operation when depth is missing. Missing-modality learning provides a practical route for reducing over-dependence on a single sensor channel [16].

Most fruit detection pipelines still represent objects using horizontal bounding boxes (HBBs). HBBs are sufficient for localization and counting, but they provide limited information about the in-plane direction of the visible fruit region. In cluttered canopies, an axis-aligned box may include branches, leaves, and neighboring fruit, and its geometric center does not necessarily indicate an accessible approach direction. Rotated-box detectors and ProbIoU-style overlap formulations show that orientation-aware boxes can encode target extent together with in-plane angle [17,18]. In this study, OBB orientation is treated only as an intermediate cue for downstream approach-direction reasoning, not as a complete 3D grasp pose.

These observations leave four linked gaps for day−night apple-harvesting perception. First, apple detectors are often evaluated mainly as image-level localizers rather than action-oriented perception modules. Second, RGB-D and multimodal fruit-perception studies commonly demonstrate benefits when complete multimodal input is available, but fewer studies test RGB-D/RGB-Z compatibility under missing depth. Third, HBB outputs provide limited image-plane orientation for approach-direction reasoning. Fourth, missing-depth robustness is seldom evaluated with a reproducible forced RGB-Z plus graded partial-depth protocol on paired orchard RGB-D data.

Existing studies partly cover each component of this chain. Robotic harvesting and in situ apple perception studies define the operational need, target-journal context, and occlusion challenge [1–8]; low-light apple studies clarify the day-night appearance problem [9,10]; RGB-D and multimodal fruit perception show the value of depth when complete multimodal input is available [11–15]; missing-modality training shows how sensor dependence can be reduced [16]; and rotated-box detection methods provide orientation-aware localization primitives [17,18]. However, these lines of work are usually evaluated separately and do not jointly test a unified RGB-D/RGB-Z detector, an OBB-based image-plane approach cue, and a forced RGB-Z plus graded partial missing-depth robustness protocol on paired orchard RGB-D data.

This study evaluates an integrated missing-depth protocol for orchard perception. The contribution combines (1) a four-channel YOLOv8s-OBB interface that accepts RGB-D and zero-padded RGB-Z without changing the inference graph; (2) training-only input-level modality dropout; (3) forced RGB-Z evaluation using the same test images; (4) graded partial-depth stress tests on paired night-time RGB-D images, with an auxiliary feature-level channel-dropout baseline; and (5) exploratory use of OBB orientation as an image-plane cue. Four-channel input, modality dropout, and OBB prediction are established components, rather than individually new algorithms. The methodological contribution is their joint evaluation under a common sensing interface: complete-depth accuracy and missing-depth degradation are examined together, instead of assuming that depth remains available. The same-backbone HBB/OBB representation comparison and geometry-only simulation illustrate the possible use of the additional angle cue, while the class-weighting ablation remains auxiliary. The study does not establish superiority over existing fusion architectures on a shared benchmark or validate a complete manipulation pipeline.

1.1. Apple Harvesting and Orchard Perception

Vision-based perception has become a central component of robotic apple harvesting, supporting fruit detection, localization, maturity-related assessment, and target selection in orchard environments [1–8]. Recent orchard studies further emphasize that clustered fruit, occluder−occludee relationships, fruit visibility, lightweight deployment, and multi-view sensing remain central to reliable apple detection under natural canopy structure [5–8,19–24]. These systems reduce manual observation, enable continuous target perception, and provide a direct interface between sensing and robotic control. However, apples are often partially occluded by leaves, branches, stems, bags, and adjacent fruits, while background clutter and fruit clustering can obscure object boundaries. These conditions limit the reliability of detection results when the harvester must make a physical approach to the target rather than only count or localize fruit in an image.

Another limitation is that many perception systems are developed or evaluated mainly under daytime or relatively controlled acquisition conditions. In practical harvesting, the visual appearance of apples changes with shadows, backlighting, canopy density, artificial night-time illumination, and low-light noise. Recent low-light and night-time apple studies show that non-uniform exposure, degraded edge cues, and enhancement quality can substantially affect localization and recognition stability [9,10]. Therefore, orchard perception for apple harvesting requires models that are not only accurate under favorable images, but also robust to occlusion and illumination changes that directly affect downstream grasp planning.

Recent studies in agriculture have extended this line of work through adverse-light apple localization, improved YOLO-based detection, lightweight embedded deployment, and integrated robotic-harvester evaluation [25–29].

1.2. Oriented Object Detection and Grasp-Oriented Detection

Conventional object detection in agricultural scenes commonly relies on horizontal bounding boxes because they are simple to annotate, train, and evaluate. This representation is effective for many recognition and counting tasks, but it provides limited geometric information for robotic manipulation. In dense apple canopies, an HBB may include a large amount of irrelevant background and cannot describe the dominant orientation of the visible target region. As a result, the output of an HBB detector is only weakly connected to the pose requirements of an end-effector, especially when the approach direction must avoid branches or align with the accessible fruit surface.

OBB detection has been widely investigated in computer vision tasks where object rotation or dense arrangement is important. Rotated detectors such as R3Det refine object localization for oriented targets [17], and ProbIoU-based formulations provide overlap measures that can be applied to rotated boxes [18]. For apple harvesting, an OBB can serve as a more action-oriented perception primitive by providing orientation-aware information that may support grasp posture estimation and collision-aware approach selection. Nevertheless, existing fruit detection studies more often focus on detection accuracy than on whether the detected representation is suitable for manipulation. This leaves a gap between visual detection outputs and the geometric information required by a harvesting manipulator.

Complementary studies on three-dimensional obstacle-avoidance planning and dual-view fruit localization further show that detector outputs must be interpreted together with branch geometry and manipulator constraints [30,31]. For near-circular apples, the OBB angle is not treated here as biological fruit orientation or as a complete grasp direction; it only describes the visible image-plane region, particularly when occlusion or truncation makes that region elongated.

1.3. Multimodal Fusion and Missing-Depth Robust Learning

RGB-D and multimodal perception can improve orchard sensing by combining complementary information from appearance and geometry, especially when one source is affected by illumination, weak texture, occlusion, or background similarity [11–15]. Recent multimodal fruit-perception work shows that feature-fusion design can improve detection in complex scenes when complete multimodal input is available [14,15]. Compared with single-modality perception, multimodal fusion can provide a richer target description and may reduce sensitivity to specific image degradations. However, the benefit of fusion often depends on the assumption that all required modalities are available and reliable during inference.

Noise-tolerant RGB-D feature fusion has also been evaluated for outdoor fruit detection [32]. The present method does not explicitly model occlusion; occlusion is treated as a scene difficulty that may coincide with missing or unreliable depth.

That complete-depth assumption is difficult to guarantee in orchard deployment. Depth information may be missing or degraded because of sensing range, reflective surfaces, night-time illumination, calibration drift, synchronization error, occlusion, or hardware disturbance. Missing-modality learning and dropout-based training provide a practical way to reduce over-dependence on a single sensor channel [16]. If a detector is trained only for complete RGB-D input, its performance may become unstable when depth is unavailable. In this context, a single RGB-D/RGB-Z OBB detector that can exploit depth when available while remaining tolerant to missing depth addresses a practical gap between multimodal perception research and grasp-oriented apple harvesting.

2. Materials and Methods

2.1. Dataset Construction

The dataset was constructed to support apple OBB detection under daytime and night-time orchard conditions. The curated dataset contained 542 annotated images, including 201 daytime RGB samples and 341 night-time RGB-D samples with verified depth pairing. Daytime samples were therefore used as RGB/RGB-Z inputs, whereas the night-time subset contained paired RGB-D data. The available RGB images and annotation files were stored at a native resolution of 1280 × 720 pixels before being resized to 640 × 640 pixels for network training and evaluation. Images contained clustered fruit, leaves, branches, stems, and background support structures. Night-time samples were acquired under artificial illumination. Apple cultivar information and detailed sensor-platform and illumination

hardware metadata were not retained in the available records. Intel RealSense D455 depth cameras were used for the paired night-time RGB-D acquisition, with an imaging distance range of 0.6–6 m.

The dataset was divided by predefined split files into 380 training images, 81 validation images, and 81 test images. Each split retained both illumination conditions, with 141 daytime and 239 night-time images for training, 30 daytime and 51 night-time images for validation, and 30 daytime and 51 night-time images for testing. No image file was shared across splits. Because complete tree-row identifiers were not available in the current annotation records, the split should be interpreted as an image-level split rather than a strict tree- or row-level generalization test. This point is reported as a limitation to avoid overclaiming scene-independent deployment performance. The archived metadata do not allow us to determine whether images from the same acquisition sequence, tree, or nearby orchard location were distributed across different subsets. We therefore do not claim strict scene-level independence or row-level generalization.

Three operational visual categories were defined from RGB appearance and image-level accessibility: ripe_pickable denotes a visually mature apple whose visible surface and surrounding clearance suggest that it could be selected as a candidate harvesting target; ripe_unpickable denotes a visually mature apple that is not immediately accessible in the image because of severe occlusion, truncation, branch interference, dense cluster overlap, or an ambiguous approach path; unripe denotes fruit whose visible appearance indicates insufficient maturity for harvesting. These labels are operational image categories rather than physiological maturity measurements; firmness, soluble-solids content, and other destructive maturity indices were not available. Classes and OBBs were manually annotated from the RGB images, and depth maps were not used to determine either the class or the box. Each OBB was fitted to the visible fruit region, so its angle describes the visible image-plane contour rather than a biological fruit orientation. The annotations contained 5482 instances: 2661 ripe_pickable, 832 ripe_unpickable, and 1989 unripe. The dataset composition is summarized in Table 1. The annotations were completed by three annotators, and each annotation was cross-reviewed. However, no formal inter-annotator agreement coefficient or complete adjudication log was retained. Cross-review therefore does not establish a measured level of annotation agreement, and accessibility-related labels remain potentially subjective. Future annotation rounds should retain independent labels and adjudication decisions so that agreement and ambiguous cases can be assessed quantitatively.

Table 1. Dataset composition. The table summarizes image split, illumination condition, modality availability, and OBB annotation statistics.

Item Value	Item Value
Total images	542
Daytime RGB/RGB-Z samples	201
Night-time RGB-D samples	341
Training validation/test split	380/81/81 images
Apple categories	ripe_pickable; ripe_unpickable; unripe
OBB instances	5482 total: 2661 ripe_pickable, 832 ripe_unpickable, 1989 unripe

2.2. Overall Detector Design

The proposed missing-depth-tolerant RGB-D/RGB-Z OBB detector is designed as a smart harvesting perception module rather than as a new network architecture. It uses one four-channel input interface in both training and inference: RGB images are concatenated with either a valid depth map to form RGB-D input or a zero-valued depth plane to form RGB-Z input when depth is unavailable. The resulting tensor is processed by a YOLOv8s-

OBB detector with the standard backbone, neck, and oriented detection head, so the same trained detector can run under complete-depth and missing-depth conditions without changing the inference graph. Input-level modality dropout is applied only during training and does not add inference-time branches, sensor checks, or ensemble models. The overall workflow is shown in Figure 1.

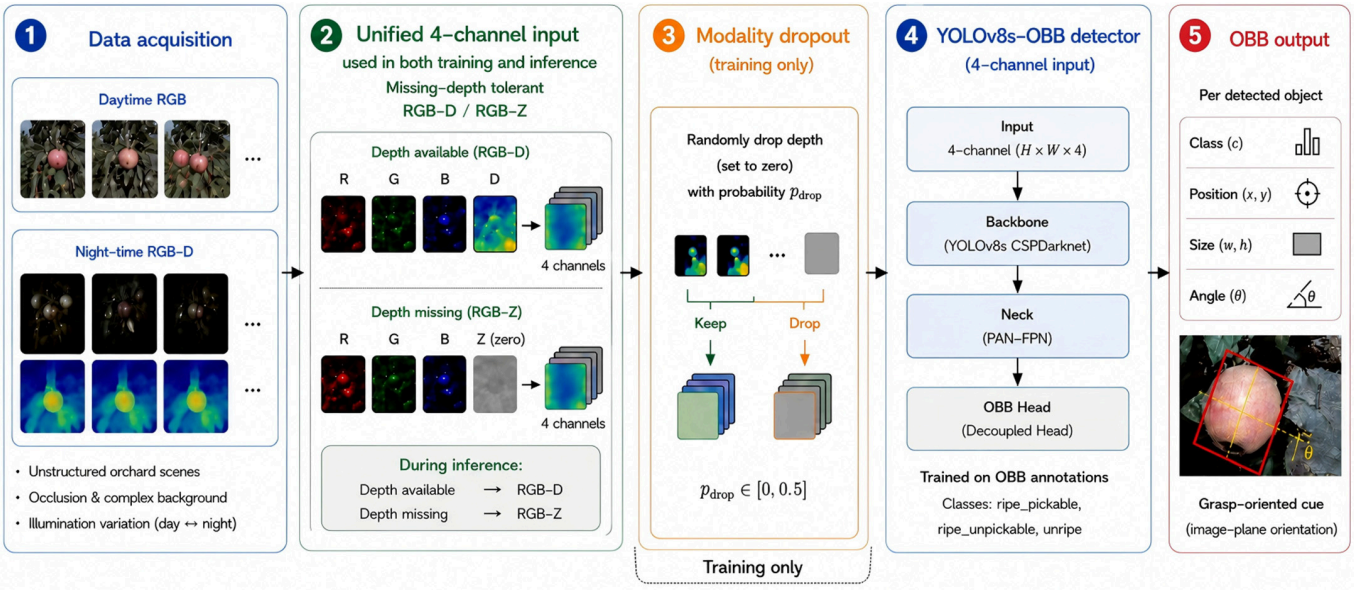

Figure 1. Missing-depth-tolerant RGB-D/RGB-Z OBB detection workflow. The four-channel input interface uses RGB-D when depth is available and zero-padded RGB-Z when depth is missing. The red box in the output example indicates the predicted oriented bounding box around the apple.

In this workflow, the detection output is treated as a perception cue for downstream reasoning rather than as a complete robotic grasp plan. The class prediction distinguishes ripe_pickable, ripe_unpickable, and unripe apples, whereas the OBB geometry provides target center, size, and image-plane orientation. That orientation can support later approach-direction reasoning, but it does not by itself determine a physically executable 3D grasp pose.

2.3. Input-Level Fusion and Missing-Depth Tolerance

Input-level fusion was adopted because it preserves a simple four-channel interface that remains compatible with the YOLOv8s-OBB implementation and avoids deployment complexity from dual-branch or late-fusion modules. For samples with valid depth, the input tensor was represented as X = [R, G, B, D], where D denotes the registered and normalized depth channel. RGB and depth images were paired before resizing so that the fourth channel remained spatially aligned with the RGB image after conversion to the 640 × 640 network input. Depth values marked as invalid by the sensor or preprocessing routine were treated as invalid pixels. Valid depth maps were normalized to the interval 0 to 1 before concatenation with RGB channels, whereas invalid pixels were assigned 0 after normalization.

To handle missing depth, D was replaced by a zero-valued plane with the same spatial size as the RGB image, producing an RGB-Z input. In the constructed dataset, daytime samples were represented as RGB-Z samples because paired depth was not available, whereas night-time samples used real depth when available. During evaluation, the same mechanism was also used to create a forced RGB-Z stress condition by setting the fourth channel to zero for all test images. Zero padding preserves input compatibility for the same

four-channel detector, but it should not be interpreted as reconstruction of the missing geometric information.

2.4. Training Strategy and Input-Level Modality Dropout

The detector was trained for OBB detection using the predefined training and validation splits. The RGB baseline used YOLOv8s-OBB pretrained weights. The RGB-D/RGB-Z variants used a four-channel YOLOv8s-OBB configuration in which the first convolutional layer was expanded to accept four input channels, while the remaining backbone, neck, and OBB head followed YOLOv8s-OBB. The pretrained RGB kernels were copied to the first three channels of the expanded layer, and the additional depth-channel kernel was initialized as the mean of the three RGB kernels before joint fine-tuning. Training was performed at an input size of 640 pixels with a batch size of 16, a maximum of 200 epochs, early-stopping patience of 50 epochs, mixed-precision computation, deterministic execution, and a fixed random seed of 42. AdamW optimization was used with a weight decay of 0.0005. Geometric and photometric augmentations included rotation-compatible augmentation up to 30 degrees for OBB labels, horizontal flipping, translation, scaling, mosaic augmentation, and color perturbation.

The resulting detector architecture is illustrated in Figure 2, highlighting the four-channel input stem and backbone, PAN-FPN multi-scale fusion, and three-scale OBB prediction head.

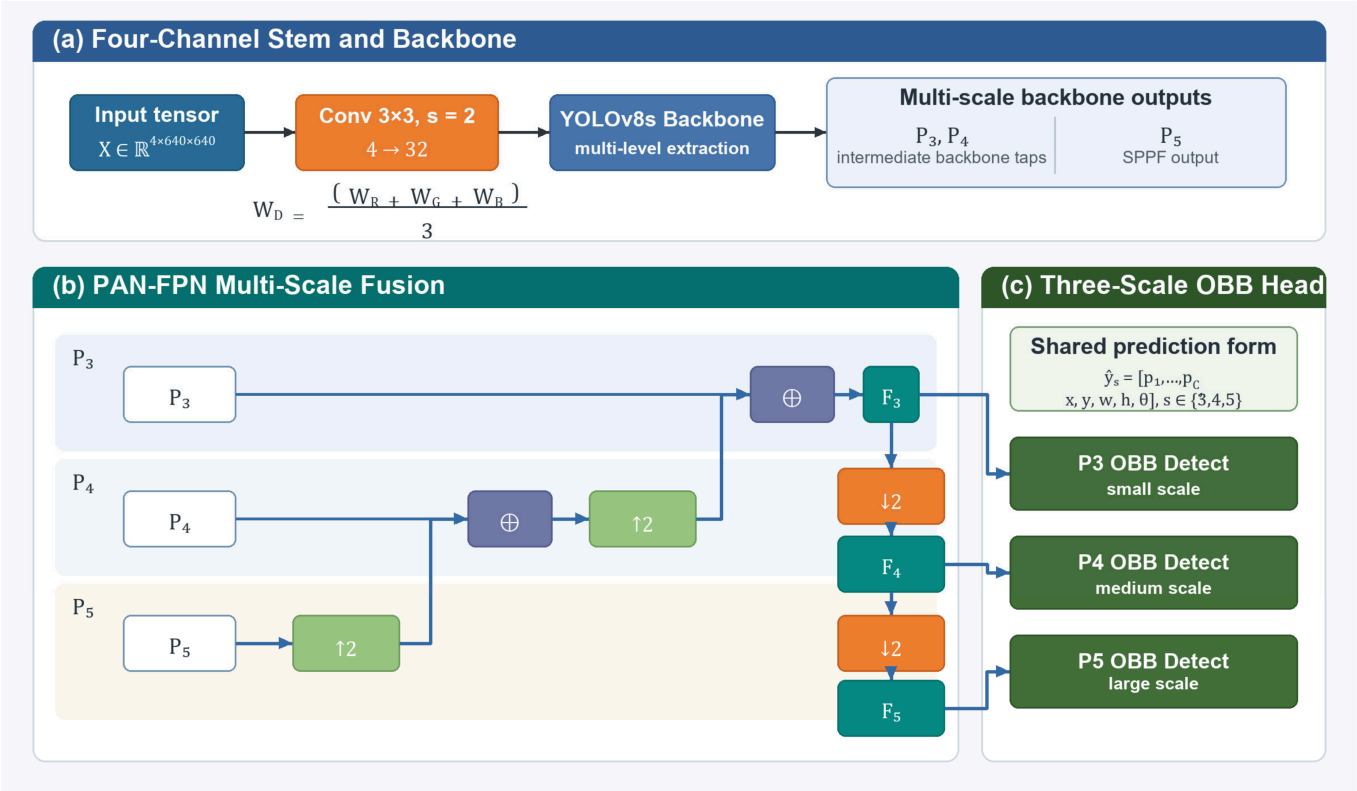

Figure 2. Four-channel YOLOv8s-OBB network architecture. The detector expands the input stem to accept RGB-D/RGB-Z tensors, initializes the depth-channel kernel from the mean RGB kernels, extracts multi-scale backbone features, fuses P3–P5 features through PAN-FPN, and predicts OBB outputs at three scales. (**a**) Four-channel input stem and backbone; (**b**) PAN-FPN multi-scale feature fusion; (**c**) three-scale OBB prediction head. Arrows indicate the direction of data and feature flow between modules.

The validation command used an IoU setting of 0.6 for prediction filtering during the YOLOv8s-OBB evaluation routine, with a maximum of 300 detections per image. This value should not be confused with the reported AP thresholds: mAP_{50} was computed at an IoU threshold of 0.50, whereas $mAP_{50\text{-}95}$ was averaged over IoU thresholds from 0.50 to 0.95, as defined in Equations (1)–(4) [33]. Thus, the reported mAP_{50} and $mAP_{50\text{-}95}$ values remain standard detection metrics rather than scores evaluated only at IoU = 0.6.

Unified evaluation used Ultralytics 8.3.253 with PyTorch 2.5.1 + cu121 on an NVIDIA GeForce RTX 4060 Laptop GPU with 8188 MiB memory. Reported timing values include preprocessing, model inference, and postprocessing per image.

To improve tolerance to missing depth, input-level modality dropout was introduced during training only. For each sample with valid depth, the depth channel D was replaced by zero with probability p_{drop} during the training augmentation stage, whereas samples without depth remained RGB-Z inputs. The tested endpoint probabilities were p_{drop} = 0.0, 0.2, 0.3, and 0.5. The schedule started with p_{start} = 0.1 during the first 10 warm-up epochs and then linearly increased the depth-drop probability toward the selected endpoint value according to Equation (5). This design follows the missing-modality intuition of ModDrop-like learning [16] while tailoring it to orchard RGB-D sensing: the detector is exposed to both complete RGB-D inputs and depth-missing RGB-Z inputs, encouraging use of depth when it is informative while preserving RGB-based discriminative ability when depth is absent. The dropout operation was disabled during validation and inference.

2.5. Auxiliary Feature-Level Baseline and Partial Missing-Depth Protocol

An auxiliary feature-level channel-dropout baseline was implemented to compare input-level modality dropout with feature-level missing-modality regularization under the same four-channel RGB-D/RGB-Z input, YOLOv8s-OBB backbone, training schedule, and p_{drop} = 0.3 endpoint. Training-only channel dropout was applied immediately after the first stem convolution. A fixed 50% subset of stem output channels was selected once using the L1 magnitude of the depth-input kernels. This heuristic provides a reproducible early-fused-feature baseline; it is not treated as a theoretical identification of uniquely depth-sensitive channels or as a primary method. The forward hook was disabled during validation and inference and removed from saved checkpoints.

Partial missing-depth robustness was evaluated on the night test subset containing paired real depth. Three deterministic masks were applied only to initially positive depth pixels: an upper-region mask removed the spatially highest valid pixels; a seeded random-block mask accumulated rectangular missing regions; and a far-threshold mask removed pixels with the highest values in the normalized depth channel. For each mask, missing ratios of 0%, 25%, 50%, 75%, and 100% were used as evenly spaced stress levels spanning complete to absent depth. These levels are evaluation settings and should not be interpreted as measured probabilities of real sensor failure. Reported outputs included precision, recall, mAP_{50}, $mAP_{50\text{-}95}$, absolute change from 0%, relative degradation (S0–Sp)/S0, and trapezoidal curve AUC. Precision, recall, AP, mAP, and IoU followed Equations (1)–(4), while the modality-dropout schedule and angular MAE used Equations (5) and (6), respectively. The 100% point is a forced RGB-Z cross-check on the night subset and is therefore distinct from the full-test forced RGB-Z values in Table 2. The three masks were selected to represent interpretable spatial failure modes that are easy to reproduce: loss in an upper image region, contiguous block-shaped holes, and loss of distant/high-normalized-depth pixels. They are deliberately stress-test patterns rather than a probability model of any particular sensor failure, and they do not reproduce all combinations of holes, noise, multipath artifacts, or RGB-depth misregistration. AUC was calculated over the missing-depth fraction r in [0, 1],

using the trapezoidal rule between the five sampled levels; it summarizes performance across the tested stress levels rather than a sensor-failure probability distribution.

Table 2. Unified -evaluator detection performance. Values compare the RGB baseline, RGB-D detectors with input-level modality dropout, and the feature-level channel-dropout baseline.

Detector	p_{drop}	Full mAP_{50}	Full mAP_{50-95}	RGB-Z mAP_{50}	Delta RGB-Z mAP_{50}	Night mAP_{50}	Day mAP_{50}
Baseline RGB	-	0.6952	0.5195	0.6952	+0.1046	0.7179	0.6603
RGB-D no dropout	0.0	0.7041	0.5277	0.5906	0.0000	0.7575	0.6178
Input-level modality dropout	0.2	0.6930	0.5135	0.6880	+0.0974	0.7233	0.6376
Input-level modality dropout	0.3	0.6804	0.5049	0.6857	+0.0951	0.6906	0.6653
Input-level modality dropout	0.5	0.7199	0.5380	0.7064	+0.1158	0.7433	0.6809
Feature-level channel-dropout baseline	0.3	0.6800	0.5084	0.6273	+0.0367	0.7005	0.6412

The principal detection and robustness indicators were calculated using standard definitions. For each IoU threshold tau, true positives, false positives, and false negatives were accumulated over all test images after confidence sorting:

$$\text{Precision} = \text{TP}/(\text{TP} + \text{FP}) \quad (1)$$

$$\text{Recall} = \text{TP}/(\text{TP} + \text{FN}) \quad (2)$$

$$\text{AP}(\tau) = \int_0^1 \text{P}_\tau(\text{R})\, \text{dR} \quad (3)$$

$$\text{mAP} = \text{mean}_c(\text{AP}_c);\ \text{mAP}_{50\text{-}95} = \text{mean}_{\tau\,=\,0.50:0.95}[\text{mAP}(\tau)] \quad (4)$$

The training-only depth-drop probability at epoch e was scheduled from the warm-up value p_{start} to the selected endpoint p_{drop}:

$$p\,(e) = p_{start} + (p_{drop} - p_{start})\ \min(\max((e - \text{ewarm})/(E - \text{ewarm}), 0), 1) \quad (5)$$

For the auxiliary orientation analysis, Equation (6) evaluates angular error with a wrapped 180-degree-periodic difference because an oriented rectangle is invariant under a 180-degree rotation. This metric was used only for evaluation and was not included in the training objective. N is the evaluated pair count, and θ_i and θ_i^* are paired angles in radians; matching details were not retained, limiting reproducibility.

$$\text{MAE}\theta = [180/(\pi\, N)]\ \Sigma_i\ |\,\text{mod}(\theta_i - \theta_i^* + \pi/2, \pi) - \pi/2\,| \quad (6)$$

2.6. OBB Detection Head

For the OBB detector, each apple instance is represented by a rotated bounding box $b = (x_c, y_c, w, h, \theta)$, where (x_c, y_c) denotes the box center, w and h denote the side lengths, and θ denotes the image-plane rotation angle relative to the horizontal axis. This is consistent with the *xywhr* representation used by the YOLOv8s-OBB implementation; the predicted box can also be converted to a four-corner polygon for visualization and evaluation. In the loss notation below, the k-th predicted box is denoted as b_k, and the matched ground-truth box is denoted as b_k^*.

For the k-th prediction unit on the multi-scale feature maps, the decoded prediction vector contains class scores $p_{k,1}, \ldots, p_{k,C}$ and the five OBB geometry terms ($x_{c,k}$, $y_{c,k}$, w_k, h_k, θ_k), where C is the number of apple categories. Compared with an HBB head, whose regression target is (x_c, y_c, w, h), the OBB head adds θ and therefore changes the regression target from an axis-aligned box to a rotated box.

Following the standard YOLOv8s-OBB training formulation and ProbIoU-based rotated-box regression [18,33], the optimization objective consists of the classification loss, the rotated-box ProbIoU localization loss, and the distribution focal loss: LOBB = Lcls + λbox LProbIoU + λdfl LDFL. Here, Lcls denotes the classification loss, LProbIoU denotes the rotated-box localization loss based on ProbIoU overlap between predicted and ground-truth OBBs, and LDFL denotes distribution focal loss for box distribution regression. The predicted orientation parameter is optimized implicitly through the rotated-box ProbIoU loss, whose geometric formulation depends on the predicted box orientation. No additional independent periodic angle-consistency loss was introduced in the training objective.

When the optional grasp-oriented class-weighting setting is enabled, only the rotated-box ProbIoU localization term is replaced by Lcw = (1/Npos) Σk∈P qk αc,k [1 − ProbIoU(bk, bk*)]. Here, P denotes the positive prediction set, qk is the TAL assignment quality weight, Npos is the normalization term, bk and bk* are the predicted and matched ground-truth OBBs, and αc,k is the class weight assigned to the category of the k-th positive sample. The classification loss and DFL term remain unchanged. These class weights apply to the complete rotated-box ProbIoU localization term and do not define a dedicated angle loss.

This grasp-oriented class weighting is treated as an optional ablation rather than the default setting in the main results, where uniform class weighting is used. The weights [1.5, 0.8, 0.5] scale the class-specific rotated-box ProbIoU localization contributions for ripe_pickable, ripe_unpickable, and unripe, respectively; they do not define an angle-specific loss. Angular MAE is used later only as an auxiliary evaluation metric for the ablation, as defined in Equation (6), and was not included in the training objective.

The OBB output is therefore a five-parameter image-plane detection primitive rather than a complete 3D grasp pose. Figure 3 illustrates the standard HBB head, the YOLOv8s-OBB head, and the optional grasp-oriented class-weighted rotated-box regression used in the ablation. The meaningful head-type comparison in this study is the HBB-vs-OBB output comparison under the same backbone and neck, asking whether the additional orientation cue is useful for geometry-only approach-direction reasoning.

2.7. Exploratory Orientation-Aware Interpretation and Simulation Setup

For orientation-aware interpretation, each detected OBB was converted into a set of manipulation-relevant cues, including the target center, box dimensions, orientation angle, confidence score, and semantic class. The ripe_pickable detections were regarded as candidate harvesting targets, whereas the ripe_unpickable and unripe detections provided non-target or avoidance information for downstream decision making. The center of the OBB indicates the approximate image-plane target location, while the orientation angle provides a candidate in-plane approach direction for a later end-effector planner.

The OBB detection output was then used in a geometry-only grasp simulation. The comparison contrasts an OBB-derived approach cue with an HBB interpretation from the same backbone, where only axis-aligned extent and center information are available. Depth, when present, can support later spatial interpretation; when it is absent, the detector still returns an RGB-derived orientation-aware result. The output is therefore treated as an image-plane cue for target selection and approach-direction reasoning, not as a substitute for 3D pose estimation or closed-loop manipulation. The geometric assumptions of the HBB- and OBB-based grasp strategies are shown in Figure 4.

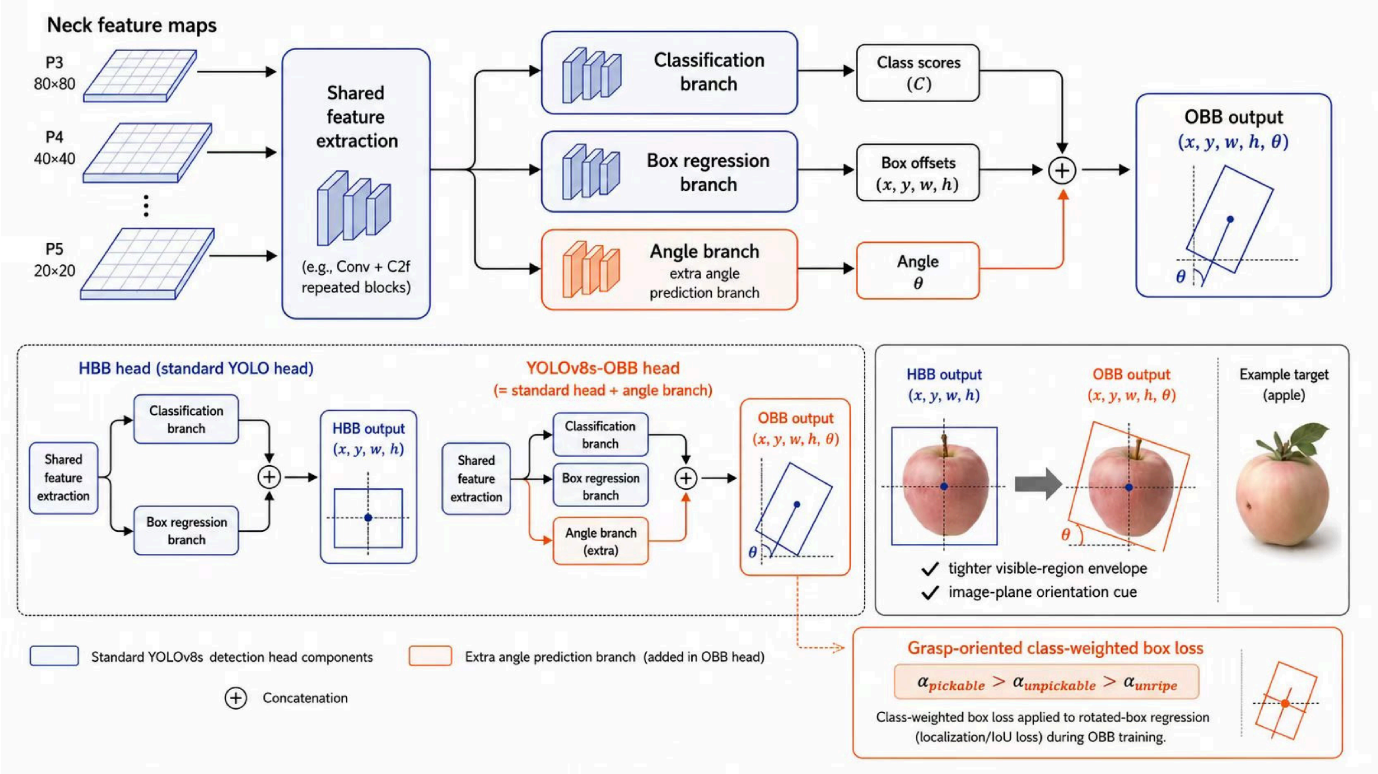

Figure 3. HBB and OBB head comparison. The OBB head adds the angle term theta to the HBB box representation, and the optional class-weighted ProbIoU term is used only in the ablation.

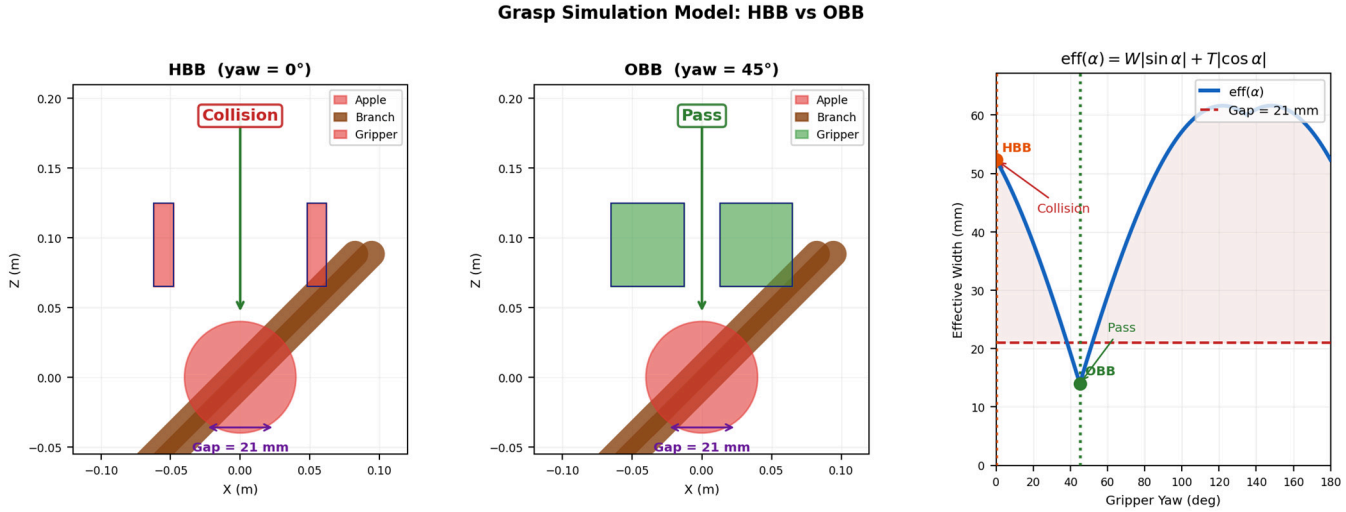

Figure 4. Geometry -only grasp simulation setup. HBB and OBB outputs are mapped to simplified approach-direction assumptions for image-plane comparison. The green dotted vertical line in the right panel marks the OBB-aligned gripper yaw of 45° used in the simulation.

3. Results

The results move from quantitative detection performance to missing-depth robustness and representative qualitative examples.

3.1. Overall Detection Performance

The overall detection and forced RGB-Z comparison is reported in Table 2, evaluated with the same configuration for every detector and scenario. The RGB-only baseline obtained 0.6952 full-test mAP_{50} and 0.5195 $mAP_{50\text{-}95}$. The four-channel detector without

input-level modality dropout achieved 0.7041 full-test mAP_{50} but decreased to 0.5906 under forced RGB-Z input. With p_{drop} = 0.5, input-level modality dropout gave the highest observed full-test result among the tested settings (0.7199 mAP_{50} and 0.5380 mAP_{50-95}) and the highest observed forced RGB-Z result (0.7064 mAP_{50}), an absolute gain of 0.1158 over no dropout. The feature-level channel-dropout baseline reached 0.6273 under forced RGB-Z input and is reported as an auxiliary comparison.

The pattern suggests that input-level modality dropout is not merely recovering performance under forced RGB-Z. With the unified evaluator, p_{drop} = 0.5 was the best observed setting in the current experiment and produced the highest full-test mAP_{50} (0.7199), forced RGB-Z mAP_{50} (0.7064), and day-only mAP_{50} (0.6809) among the four-channel variants. Its advantage over the RGB-only baseline under forced RGB-Z was modest (+0.0112), so the main result is a single RGB-D/RGB-Z detector that degrades gracefully when depth is absent. Figure 5 summarizes the input-level modality-dropout comparison. A second result is the partial missing-depth robustness pattern reported in Figure 6. This empirical setting should not be interpreted as a universally optimal dropout probability.

3.2. Ablation Analysis of Input-Level Modality Dropout

The effect of the input-level modality dropout probability is visualized in Figure 5 using the unified evaluator. Full-test mAP_{50} values were 0.7041, 0.6930, 0.6804, and 0.7199 at p_{drop} = 0.0, 0.2, 0.3, and 0.5, respectively. Forced RGB-Z mAP_{50} increased from 0.5906 without dropout to 0.6880, 0.6857, and 0.7064. The p_{drop} = 0.5 setting produced the highest observed full-test and forced RGB-Z mAP_{50} values among the tested settings and was selected as the input-level modality dropout representative for the partial missing-depth robustness analysis.

The ablation indicates that training-time input-level modality dropout can reduce dependence on the depth channel while retaining a single unchanged inference graph. The non-monotonic intermediate values also show that p_{drop} is an empirical robustness setting rather than a universally optimal probability.

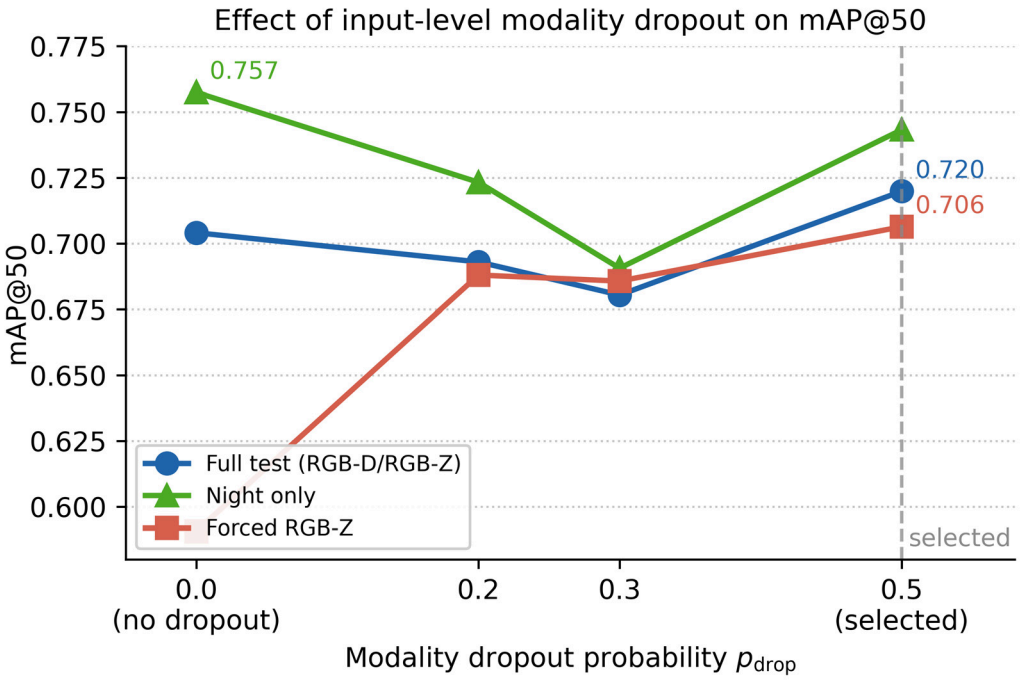

Figure 5. Input -level modality dropout ablation. mAP_{50} is shown under full-test, night-only, and forced RGB-Z conditions.

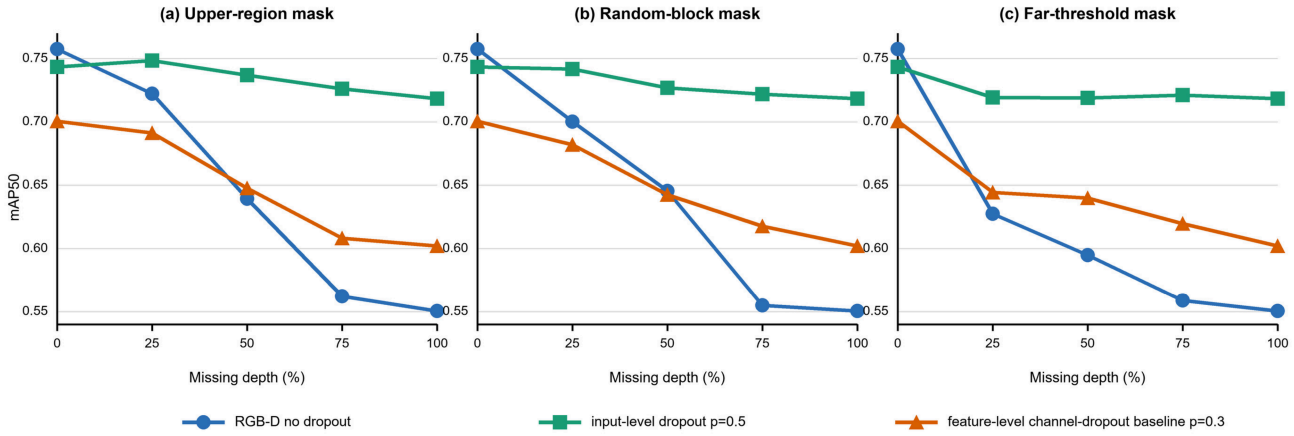

Figure 6. Partial missing-depth robustness. Curves show mAP_{50} for upper-region, random-block, and far-threshold masks on paired night RGB-D test images. (**a**) Upper-region mask; (**b**) random-block mask; (**c**) far-threshold mask.

3.3. Class-Wise Performance and Occlusion-Sensitive Categories

Under the unified full-test evaluation for the p_{drop} = 0.5 detector, ripe_pickable achieved AP50/AP50-95 of 0.907/0.764, ripe_unpickable achieved 0.509/0.314, and unripe achieved 0.744/0.536. On the NVIDIA GeForce RTX 4060 Laptop GPU, the same evaluation recorded 0.43 ms preprocessing, 5.85 ms inference, and 3.68 ms postprocessing per image (9.96 ms total for these components). These matched-model results identify ripe_unpickable as the most difficult category; its lower performance is consistent with the smaller number of instances and with a label that depends on both visual maturity and image-level accessibility. The 9.96 ms value is a laboratory reference for the stated laptop GPU and software pipeline; it should not be read as a validated real-time guarantee for an embedded agricultural-robot platform.

The ripe_unpickable category contains 832 instances, compared with 2661 ripe_pickable and 1989 unripe instances. Its lower AP is compatible with both class imbalance and ambiguity in an accessibility-related label, but these summaries do not distinguish their effects. When only part of a mature fruit is visible, its appearance can overlap with ripe_pickable fruit despite differences in annotated accessibility. The current evidence therefore identifies a difficult category without attributing a measured fraction of its errors to imbalance, occlusion, or annotation subjectivity.

Qualitative failure-mode taxonomy for ripe_unpickable. The p_{drop} = 0.5 detector achieved AP50/AP50-95 of 0.509/0.314 for ripe_unpickable, compared with 0.907/0.764 for ripe_pickable and 0.744/0.536 for unripe. Class-specific precision and recall, a confusion matrix, and per-instance prediction logs were not retained in the archived evaluation outputs. AP summaries alone cannot recover confusion counts or distinguish the contributions of class imbalance and label subjectivity. Figure 7 presents qualitative detector-output examples. Figure 8 contains annotated dataset examples, rather than detector-error measurements. It illustrates severe occlusion and truncation, which may reduce contour evidence; branch interference and weak surface contrast, which may make accessibility or maturity labels ambiguous; and dense cluster overlap, which may complicate instance separation. The difficult-case discussion is a qualitative failure-mode taxonomy, not a measured decomposition of prediction errors. Assessing these hypotheses requires matched predictions and ground truth, with error counts stratified by class and occlusion condition.

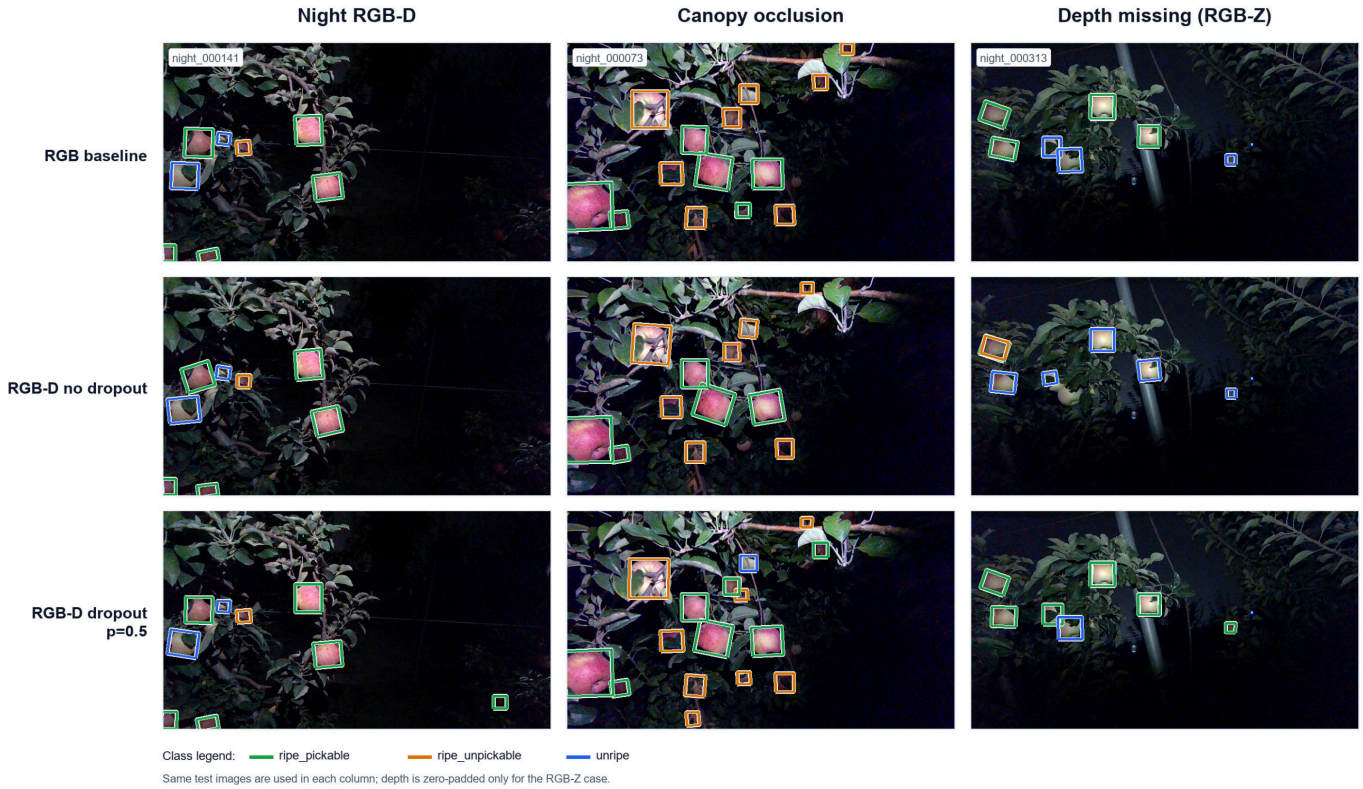

Figure 7. Qualitative detection examples. Columns share the same test image across detector rows, with OBB colors denoting ripe_pickable, ripe_unpickable, and unripe classes.

3.4. Scenario-Wise Robustness Under Day, Night, and Forced RGB-Z Input

Returning to Table 2, the no-dropout four-channel detector achieved 0.7575 mAP_{50} on night images with real depth but decreased to 0.5906 when the full test set was forced to RGB-Z. Input-level modality dropout at p_{drop} = 0.5 achieved 0.7433 on the night subset and 0.7064 under full-test forced RGB-Z, whereas the feature-level channel-dropout baseline reached 0.7005 and 0.6273, respectively. These scenario results motivate evaluation across intermediate missing-depth severities rather than relying only on a single all-zero endpoint.

The day-only and night-only results are scenario-specific observations under the available modality composition, not a controlled estimate of illumination robustness. At p_{drop} = 0.5, mAP_{50} was 0.7433 on night-time RGB-D images and 0.6809 on daytime RGB-Z images. Both illumination and input modality differ between these subsets. The within-night missing-depth experiment holds the RGB images fixed while altering the depth channel; its results support missing-depth tolerance under the evaluated night-time acquisition condition, while not providing an independent estimate of illumination robustness or resolving the day/night confounding.

3.5. Partial Missing-Depth Robustness

After the overall and scenario-wise comparisons, the robustness analysis isolates missing depth on paired night RGB-D images.

Figure 6 isolates partial missing-depth robustness on paired night RGB-D images. The curves compare the no-dropout RGB-D detector, the selected input-level modality dropout detector at p_{drop} = 0.5, and the auxiliary feature-level channel-dropout baseline at p_{drop} = 0.3. Each detector was tested at 0%, 25%, 50%, 75%, and 100% missing depth under upper-region, deterministic random-block, and far-threshold masks.



Figure 8. Dataset-level difficult cases. Annotated examples show severe canopy occlusion, ambiguous surface features, and dense cluster overlap. (**a**) Partly hidden apples in a dark canopy; (**b**) fruit obscured by leaves and branches; (**c**) partly occluded fruit in dense foliage; (**d**) low-contrast fruit in a dark scene; (**e**) fruit with weak contrast against foliage; (**f**) fruit boundaries under backlighting; (**g**) clustered fruit in a dark canopy; (**h**) adjacent overlapping apples; (**i**) densely clustered apples among branches. Colored boxes show manual ground-truth OBB annotations, not detector predictions.

The detector trained with input-level modality dropout produced the flattest curves for all three masks. Its mAP_{50} AUC ranged from 0.7225 to 0.7355, compared with 0.6088–0.6444 for the no-dropout detector and 0.6388–0.6496 for the feature-level channel-dropout baseline. The range-based summary highlights degradation behavior across masks rather than treating the 100% missing-depth point as the only robustness evidence.

At 100% missing depth on the night subset, mAP_{50} was 0.5505 without dropout, 0.7183 with input-level modality dropout, and 0.6021 with the feature-level channel-dropout baseline. Relative to each detector's 0% score, these values correspond to mAP_{50} decreases of 27.33%, 3.36%, and 14.05%, respectively. The identical 100% values across mask types confirm the night-subset forced RGB-Z cross-check. The feature-level channel-dropout baseline improved severe-missingness robustness relative to no dropout, but it remained below the simpler input-level modality dropout strategy across the reported curve AUCs and endpoints.

To provide a qualitative view of these robustness patterns, Figure 7 compares predictions from the RGB baseline, the RGB-D detector without dropout, and the RGB-D detector trained with input-level modality dropout at $p_{drop} = 0.5$ on representative night RGB-D, canopy-occlusion, and forced RGB-Z cases.

3.6. Dataset-Level Difficult Cases Associated with Failure Modes

The qualitative examples are followed by dataset-level difficult cases so that scene difficulty is separated from strict prediction-error analysis.

Figure 8 is included as a dataset-level view of difficult cases, not as a strict prediction-error panel. The selected examples focus on severe canopy occlusion, ambiguous surface features, and dense cluster overlap. The overlaid OBBs are manually labeled dataset annotations, so they indicate ground-truth category regions rather than confidence-scored detector predictions. These cases help explain why ripe_unpickable and heavily occluded apples remain difficult, without claiming a complete qualitative evaluation of detector errors.

In the severe canopy occlusion cases, large parts of the fruit surface are hidden by leaves, branches, or neighboring objects. Incomplete contours make class assignment and OBB fitting less stable, especially for ripe_unpickable apples, where the label depends on both maturity and image-level accessibility.

In the ambiguous surface feature cases, fruit boundaries and surface appearance are weakly separated from the surrounding canopy. Illumination changes, shading, reflection, and color similarity reduce the reliability of RGB appearance cues and missing or invalid depth further limits geometric disambiguation.

In dense cluster-overlap cases, adjacent apples have overlapping visible regions and similar color, texture, or depth structure. An OBB can describe an oriented visible region, but the present qualitative examples do not establish improved instance separation relative to HBBs.

4. Discussion

Taken together, the results show that input-level modality dropout improves missing-depth robustness of the four-channel RGB-D/RGB-Z detector under both the full-test forced RGB-Z condition and graded partial-depth stress. Without dropout, forced RGB-Z mAP_{50} was 0.5906; with $p_{drop} = 0.5$, it increased to 0.7064. The partial missing-depth AUC remained between 0.7225 and 0.7355, compared with 0.6088–0.6444 without dropout. This supports a degradation-pattern claim rather than only a single endpoint comparison.

The auxiliary feature-level channel-dropout baseline reached 0.6273 full-test forced RGB-Z mAP_{50} and partial-curve AUCs of 0.6388–0.6496. It improved severe-missingness performance relative to no dropout but remained below input-level modality dropout in these reported experiments. This baseline uses a fixed L1-based channel-selection heuristic and was evaluated at $p_{drop} = 0.3$, whereas the selected input-level model uses $p_{drop} = 0.5$. At the matched endpoint of 0.3, full-test forced RGB-Z mAP_{50} was 0.6857 for input-level dropout and 0.6273 for the feature-level baseline. These comparisons support the evaluated input-level strategy, but do not establish superiority over feature-level fusion or regularization in general.

The prediction examples in Figure 7 visually support the quantitative finding that input-level modality dropout mainly improves robustness when the fourth channel is absent or unreliable, while occluded ripe_unpickable fruit remains challenging.

Figure 9 provides an illustrative feature-response visualization. Heat maps were obtained from the channel-wise mean of the absolute intermediate backbone activations, followed by upsampling and min-max normalization. The displayed input-level dropout model uses $p_{drop} = 0.3$; it is different from the $p_{drop} = 0.5$ model used for the main quantitative results. These activations are qualitative examples and do not independently establish the mechanism underlying missing-depth tolerance.

Compared with previous RGB-D and multimodal fruit perception studies, the contribution of the present detector is RGB-D/RGB-Z missing-depth compatibility within one OBB detector rather than higher absolute accuracy on a shared benchmark. Earlier

RGB-D apple and fruit-localization studies demonstrated the value of combining color and geometric information for detection and localization [11–13]. More recent multimodal networks, such as MLP-based multimodal tomato detection [14] and MT-WavYOLO for occluded orchard fruits [15], show that feature-level fusion and 3D-aware fusion can improve fruit perception when complete multimodal input is available. Recent apple-detection studies in agricultural robotics and computer vision venues further show strong progress in occlusion-aware, real-time, lightweight, and localization-oriented perception [5–8,19–24]. In contrast, this method keeps the fusion interface simple: a zero-compatible four-channel input is used with YOLOv8s-OBB, and missing-depth robustness is improved through training-time input-level modality dropout. Table 3 summarizes this positioning.

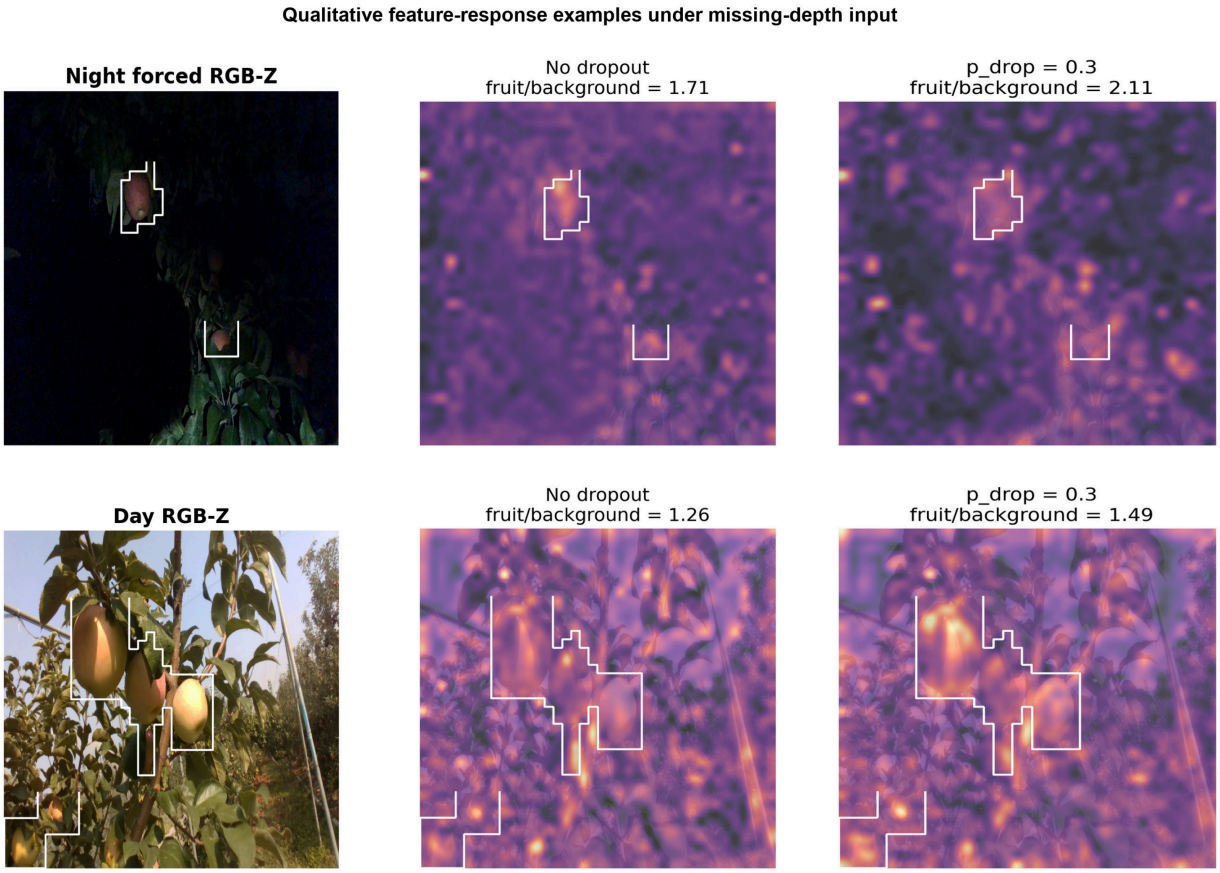

Figure 9. Illustrative feature-response visualization for RGB-Z inputs. The input-level modality-dropout model shown here uses $p_{drop} = 0.3$, whereas the main quantitative model uses $p_{drop} = 0.5$. Heat maps show normalized intermediate activations and are not a quantitative comparison of the selected main detector. White contours outline the annotated apple regions.

OBBs add an image-plane angle to the center and extent available from HBBs. Within the defined geometry-only model, the OBB-based strategy achieved an 80.0% geometry-only pass rate, compared with 40.0% for HBB, supporting the utility of the additional image-plane orientation cue for simplified approach-direction reasoning. This is a representational and geometry-level result; the present evidence does not establish that the angle is an accurate 3D approach direction. Appendix A.1 compares simplified approach strategies in a geometry-only Monte Carlo model and does not validate robot dynamics, fruit detachment, 3D localization, or closed-loop collision avoidance. The archived outputs do not support reconstruction of angular MAE or another orientation-specific metric for the main $p_{drop} = 0.5$ detector. The quantitative orientation results in Appendix A.2 therefore remain auxiliary evidence and are not transferred to the main detector. Direct validation

requires matched OBB predictions and ground truth, an explicit matching rule and sample count, and angle analysis that accounts for the instability of near-circular boxes. A separate robot experiment is required to connect orientation accuracy to harvesting performance. The class-weighted auxiliary variant is not the default: its lower angular MAE coincided with lower overall mAP_{50} and a lower angular-threshold pass rate.

Table 3. Positioning relative to representative orchard perception studies. The comparison summarizes fusion or representation choices and missing-depth handling.

Study	Target/Task	Fusion or Representation	Missing-Depth Handling	Position Relative to This Study
Gene-Mola et al. [11]	Fuji apple detection using RGB-D camera data	Multimodal deep learning with RGB-D and radiometric cues	Complete RGB-D input assumed	Shows value of color-depth complementarity; no forced RGB-Z protocol
Chen et al. [14]	Tomato detection in complex scenes	MLP-based multimodal fusion architecture analysis	Complete multimodal input used	Recent multimodal fruit detector; less focused on single-detector RGB-D/RGB-Z compatibility
Fu et al. [15]	Occluded orchard-fruit robotic harvesting	MT-WavYOLO with multi-task learning and 3D frustum fusion	Complete 3D-aware input setting	Strong recent multimodal/3D-aware orchard method; not a missing-depth stress-test interface
Shi et al. [9]	Low-light apple detection	Knowledge distillation with geometry-consistent feature alignment	Low-light robustness considered; depth missingness not addressed	Relevant day/night robustness context, but not an RGB-D/RGB-Z interface
This study	Day-night apple OBB detection with image-plane grasp-oriented cues	Four-channel RGB-D/RGB-Z input-level fusion plus training-only modality dropout	Zero-padded RGB-Z input explicitly evaluated under forced missing-depth stress	Interface-level contribution; improves full-test forced RGB-Z mAP_{50} from 0.5906 to 0.7064, adds a reproducible partial missing-depth robustness protocol, and treats the feature-level channel-dropout baseline as auxiliary

The class-wise AP results identify ripe_unpickable as the weakest category in the evaluated main model. Its smaller instance count and accessibility-based definition make class imbalance and label ambiguity plausible explanations, alongside occlusion. These explanations remain hypotheses because the available summaries do not contain class-specific confusion counts or controlled label-ablation results. Cross-review by three annotators improves the documented review process but cannot substitute for a retained quantitative agreement assessment.

The dataset-level difficult cases in Figure 8 further clarify the main limitations of the current detector. Because the overlays are ground-truth annotations rather than prediction outputs, the figure should not be read as a strict failure-analysis panel. Instead, it illustrates orchard conditions that are associated with likely errors: severe occlusion, weak surface contrast, and overlapping fruit instances. These conditions directly affect ripe_unpickable apples, for which visual evidence is often incomplete, and the accessibility boundary is less objective than simple fruit presence.

Several limitations bound these findings. The dataset contains 542 images and uses an image-level split; acquisition-sequence, tree, and row identifiers are incomplete, so

scene-independent generalization cannot be established. Daytime RGB-Z and night-time RGB-D observations confound illumination with depth availability. The synthetic depth masks are reproducible stress tests, but do not reproduce the full range of sensor noise, multipath effects, or registration errors. Zero padding preserves the interface without reconstructing missing geometry. Cultivar and detailed illumination hardware metadata are unavailable, limiting acquisition reproducibility and external validity. Each detector setting was trained once with seed 42; deterministic execution aids repeatability but does not quantify training variability. The reported values are point estimates without confidence intervals, and small differences cannot be interpreted as statistically established improvements. The choice $p_{drop} = 0.5$ is descriptive of the tested settings and requires confirmation on independent data and multiple seeds. Follow-up work should retain complete acquisition and annotation records, use sequence- or tree-disjoint evaluation, acquire paired RGB-D data across day/night conditions, repeat training, and test a documented robot platform under measured depth failures.

5. Conclusions

This study evaluated missing-depth-tolerant RGB-D/RGB-Z OBB detection for day-night apple-harvesting perception. The work provides a single four-channel detector, training-time input-level modality dropout, and a reproducible robustness protocol. The same inference graph accepts RGB-D input when depth is available and RGB-Z input when the fourth channel is zero padded.

In the reported single-run comparison, $p_{drop} = 0.5$ increased full-test forced RGB-Z mAP_{50} from 0.5906 without dropout to 0.7064, compared with 0.6952 for the dedicated RGB-only baseline. The paired night-time stress tests yielded mAP_{50} AUCs of 0.7225–0.7355 across the three mask types. The principal result is compatibility with available and missing depth through a single inference graph; the gain over the RGB-only baseline is modest. The selected dropout setting is the best observed among the tested values, not a universal optimum. The image-level split, day/night–modality confounding, and absence of repeated runs limit claims about generalization, illumination robustness, and statistical stability.

For smart apple-harvesting perception, the detector can use valid depth when available and still return detections when the fourth channel is partly or fully unavailable. OBB output remains an image-plane geometric cue. Neither the exploratory geometry-only simulation nor the detector output establishes a three-dimensional grasp pose, robot collision-avoidance performance, or field harvesting success. The next step is integration with a documented RGB-D vision-manipulator platform and evaluation of the complete perception-control loop under measured depth failures.

Author Contributions: Conceptualization, Y.X. and S.K.; methodology, Y.X., X.C., and S.K.; software, Y.X., H.Z., and Z.G.; validation, H.W., S.M., L.S., and X.C.; formal analysis, Y.X., Y.W., and Z.G.; investigation, H.W. and S.M.; resources, S.M., H.Z., L.J., and Y.W.; data curation, Y.X., L.J., and H.W.; writing—original draft preparation, Y.X. and L.S.; writing-review and editing, H.Z., L.S., L.J., X.C., Y.W., S.K., and W.P.; visualization, Z.G., X.C., and L.S.; supervision, H.Z., L.S., L.J., X.C., Y.W., S.K., and W.P.; project administration, L.J., Y.W., and W.P.; funding acquisition, W.P. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Youth Fund of the Natural Science Foundation of Shandong Province, grant number ZR2026QC1458; the China Agriculture Research System, grant number CARS-25; and the Shandong Province Key R&D Plan, grant number 2023TZXD061.

Institutional Review Board Statement: Not applicable. This study did not involve humans or animals.

Data Availability Statement: The raw data supporting the conclusions of this article will be made available by the authors on request.

Acknowledgments: The authors acknowledge the institutional support of the College of Mechanical and Electrical Engineering, Shandong Agricultural University, and the Shandong Key Laboratory of Intelligent Production Technology and Equipment for Facility Horticulture.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analysis, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

AP	Average precision
AUC	Area under the curve
DFL	Distribution focal loss
HBB	Horizontal bounding box
IoU	Intersection over union
MAE	Mean absolute error
mAP	Mean average precision
OBB	Oriented bounding box
PAN-FPN	Path aggregation network–feature pyramid network
RGB-D	Red-green-blue plus depth
RGB-Z	RGB plus a zero-filled fourth channel
TAL	Task-aligned assignment

Appendix A

This appendix retains the exploratory geometry-only simulation and class-weighting ablation for transparency. These analyses do not constitute physical manipulator validation or field harvesting results.

Appendix A.1. Exploratory Geometry-Only Analysis of HBB and OBB Outputs

A restricted geometry-only simulation was used to connect the detector output with the intended harvesting use case.

Appendix A.1 reports 100 Monte Carlo trials of simplified HBB- and OBB-based approach strategies using the geometry in Figure 4. The HBB-based strategy had a geometry-only pass rate of 40.0% and an interference rate of 60.0%; the corresponding OBB-based rates were 80.0% and 20.0%. These values describe outcomes within the specified geometric model. They are not detector localization or angular-accuracy metrics, physical collision-avoidance measurements, or observed harvesting success rates.

Figure 4 illustrates the model assumptions: a branch angle of 45 degrees, a gap width of 21 mm, a finger width of 60 mm, and a finger thickness of 14 mm. The HBB strategy uses a fixed horizontal gripper direction, whereas the OBB strategy uses an oriented image-plane cue. The effective finger width changes with rotation, illustrating how an assumed direction can alter passage through the simplified gap. The difference in Monte Carlo pass rates is conditional on these assumptions and does not isolate detector-angle accuracy or demonstrate OBB superiority in physical manipulation.

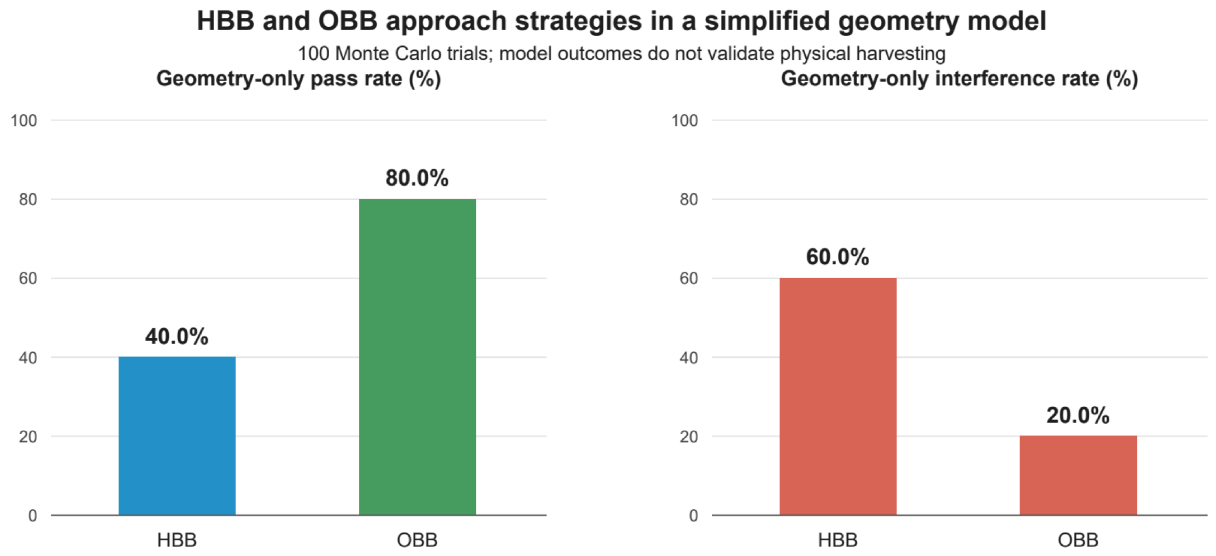

Figure A1. Geometry-only Monte Carlo model outcomes over 100 trials. Pass and interference rates compare simplified HBB- and OBB-based approach strategies; these percentages do not represent physical grasp or harvesting success.

Appendix A.2. Auxiliary Class-Weighting Ablation

The separate class-weighting ablation compares uniform weights [1.0, 1.0, 1.0] with weights [1.5, 0.8, 0.5] for ripe_pickable, ripe_unpickable, and unripe. Table A1 reports overall mAP_{50}, per-class AP50, angular MAE, and an angular-threshold pass rate for the evaluated elongated ripe_pickable boxes. This pass rate is a different metric from the Monte Carlo geometry-only pass rate in Appendix A.1. The ablation values are not orientation measurements for the main $p_{drop} = 0.5$ detector.

Table A1. Auxiliary class-weighting ablation. Angular MAE and angular-threshold pass rate concern only the evaluated elongated ripe_pickable instances; they do not characterize the main $p_{drop} = 0.5$ detector.

Variant	Class Weights	Overall mAP_{50}	Ripe_ Pickable AP_{50}	Ripe_ Unpickable AP_{50}	Unripe AP_{50}	Angular MAE (deg)	Angular-Threshold Pass Rate
Uniform	[1.0, 1.0, 1.0]	0.6952	0.8953	0.4763	0.7140	11.16	75.61%
Grasp-oriented	[1.5, 0.8, 0.5]	0.6869	0.8882	0.4529	0.7196	9.84	74.36%

Note: The geometry indicators were computed only for elongated ripe_pickable ground-truth boxes with aspect ratio ≥ 1.3. Angular-threshold pass rate denotes the reported fraction passing the 15-degree angular-error criterion, previously termed geometry pass rate. The archived summary does not retain the evaluated pair count or detailed matching protocol. These indicators should be interpreted as auxiliary, with this reproducibility limitation.

Class weighting reduced the reported angular MAE from 11.16 degrees to 9.84 degrees, a decrease of 1.32 degrees. However, overall mAP_{50} fell from 0.6952 to 0.6869, ripe_pickable AP50 from 0.8953 to 0.8882, ripe_unpickable AP50 from 0.4763 to 0.4529, and the angular-threshold pass rate from 75.61% to 74.36%. Only unripe AP50 increased, from 0.7140 to 0.7196. A lower mean angular error can coexist with a lower threshold pass rate because the metrics summarize different features of the error distribution; the archived aggregate values cannot identify the underlying change. Without repeated runs or instance-level errors, the small MAE difference does not establish more stable orientation prediction. The weighted variant is retained as an auxiliary ablation, and uniform class weighting remains the default for the main detector.

References

1. Bac, C.W.; van Henten, E.J.; Hemming, J.; Edan, Y. Harvesting robots for high-value crops: State-of-the-art review and challenges ahead. *J. Field Robot.* **2014**, *31*, 888–911. [CrossRef]
2. Silwal, A.; Davidson, J.R.; Karkee, M.; Mo, C.; Zhang, Q.; Lewis, K. Design, integration, and field evaluation of a robotic apple harvester. *J. Field Robot.* **2017**, *34*, 1140–1159. [CrossRef]
3. Zhang, K.; Lammers, K.; Chu, P.; Li, Z.; Lu, R. An automated apple harvesting robot—From system design to field evaluation. *J. Field Robot.* **2024**, *41*, 2384–2400. [CrossRef]
4. Tianjing, Y.; Mhamed, M. Developments in automated harvesting equipment for the apple in the orchard: Review. *Smart Agric. Technol.* **2024**, *9*, 100491. [CrossRef]
5. Chu, P.; Li, Z.; Zhang, K.; Chen, D.; Lammers, K.; Lu, R. O2RNet: Occluder-occludee relational network for robust apple detection in clustered orchard environments. *Smart Agric. Technol.* **2023**, *5*, 100284. [CrossRef]
6. Chiang, C.; Monney, A.; Monney, P.; Christen, D. ASPEN study case: Real time in situ apples detection and characterization. *Smart Agric. Technol.* **2024**, *8*, 100506. [CrossRef]
7. Wen, S.; Zhou, J.; Hu, G.; Zhang, H.; Tao, S.; Wang, Z.; Chen, J. PcMNet: An efficient lightweight apple detection algorithm in natural orchards. *Smart Agric. Technol.* **2024**, *9*, 100623. [CrossRef]
8. Villacres, J.; Vougioukas, S. Assessing a multi-camera system to enhance fruit visibility for robotic harvesting in a V-trellised apple orchard. *Comput. Electron. Agric.* **2024**, *224*, 109164. [CrossRef]
9. Shi, Y.; Ma, Y.; Geng, L.; Chu, L.; Li, B.; Li, W. Knowledge distillation with geometry-consistent feature alignment for robust low-light apple detection. *Sensors* **2025**, *25*, 4871. [CrossRef]
10. Xu, L.; Ni, T.; Zhou, F.; Lin, Q.; Chen, L.; Luo, Q.; Xia, K.; Xue, X.; Zheng, X.; Wu, S.; et al. Nighttime apple harvesting image enhancement based on a global-local feature fusion diffusion model. *Comput. Electron. Agric.* **2026**, *246*, 111672. [CrossRef]
11. Gené-Mola, J.; Vilaplana, V.; Rosell-Polo, J.R.; Morros, J.R.; Ruiz-Hidalgo, J.; Gregorio, E. Multi-modal deep learning for Fuji apple detection using RGB-D cameras and their radiometric capabilities. *Comput. Electron. Agric.* **2019**, *162*, 689–698. [CrossRef]
12. Fu, L.; Gao, F.; Wu, J.; Li, R.; Karkee, M.; Zhang, Q. Application of consumer RGB-D cameras for fruit detection and localization in field: A critical review. *Comput. Electron. Agric.* **2020**, *177*, 105687. [CrossRef]
13. Gené-Mola, J.; Ferrer-Ferrer, M.; Hemming, J.; van Dalfsen, P.; de Hoog, D.; Sanz-Cortiella, R.; Rosell-Polo, J.R.; Morros, J.R.; Vilaplana, V.; Ruiz-Hidalgo, J.; et al. AmodalAppleSize_RGB-D dataset: RGB-D images of apple trees annotated with modal and amodal segmentation masks for fruit detection, visibility and size estimation. *Data Brief* **2024**, *52*, 110000. [CrossRef]
14. Chen, W.; Rao, Y.; Wang, F.; Zhang, Y.; Wang, T.; Jin, X.; Hou, W.; Jiang, Z.; Zhang, W. MLP-based multimodal tomato detection in complex scenarios: Insights from task-specific analysis of feature fusion architectures. *Comput. Electron. Agric.* **2024**, *221*, 108951. [CrossRef]
15. Fu, H.; Li, T.; Feng, Q.; Chen, L. MT-WavYOLO: Bridging multi-task learning and 3D frustum fusion for non-destructive robotic harvesting of occluded orchard fruits. *Comput. Electron. Agric.* **2026**, *242*, 111335. [CrossRef]
16. Neverova, N.; Wolf, C.; Taylor, G.W.; Nebout, F. ModDrop: Adaptive multi-modal gesture recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *38*, 1692–1706. [CrossRef]
17. Yang, X.; Yan, J.; Feng, Z.; He, T. R3Det: Refined single-stage detector with feature refinement for rotating object. *Proc. AAAI Conf. Artif. Intell.* **2021**, *35*, 3163–3171. [CrossRef]
18. Murrugarra-Llerena, J.; Kirsten, L.N.; Zeni, L.F.; Jung, C.R. Probabilistic Intersection-Over-Union for Training and Evaluation of Oriented Object Detectors. *IEEE Trans. Image Process.* **2024**, *33*, 671–681. [CrossRef]
19. Hani, N.; Roy, P.; Isler, V. MinneApple: A benchmark dataset for apple detection and segmentation. *IEEE Robot. Autom. Lett.* **2020**, *5*, 852–858. [CrossRef]
20. Wang, D.; He, D. Apple detection and instance segmentation in natural environments using an improved Mask Scoring R-CNN model. *Front. Plant Sci.* **2022**, *13*, 1016470. [CrossRef]
21. Chu, P.; Li, Z.; Lammers, K.; Lu, R.; Liu, X. Deep learning-based apple detection using a suppression mask R-CNN. *Pattern Recognit. Lett.* **2021**, *147*, 206–211. [CrossRef]
22. Jia, W.; Tian, Y.; Luo, R.; Zhang, Z.; Lian, J.; Zheng, Y. Detection and segmentation of overlapped fruits based on optimized Mask R-CNN application in apple harvesting robot. *Comput. Electron. Agric.* **2020**, *172*, 105380. [CrossRef]
23. 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]
24. Li, T.; Fang, W.; Zhao, G.; Gao, F.; Wu, Z.; Li, R.; Fu, L.; Dhupia, J. An improved binocular localization method for apple based on fruit detection using deep learning. *Inf. Process. Agric.* **2023**, *10*, 276–287. [CrossRef]
25. 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]
26. Yang, L.; Zhang, T.; Zhou, S.; Guo, J. AAB-YOLO: An improved YOLOv11 network for apple detection in natural environments. *Agriculture* **2025**, *15*, 836. [CrossRef]

27. Olguín-Rojas, J.C.; Vasquez, J.I.; López-Canteñs, G.J.; Herrera-Lozada, J.C.; Mota-Delfin, C. A lightweight YOLO-based architecture for apple detection on embedded systems. *Agriculture* **2025**, *15*, 838. [CrossRef]
28. Xu, Y.; Qiao, X.; Ding, L.; Li, X.; Chen, Z.; Yue, X. Enhanced YOLOv5 with ECA module for vision-based apple harvesting using a 6-DOF robotic arm in occluded environments. *Agriculture* **2025**, *15*, 1850. [CrossRef]
29. Jin, T.; Han, X.; Wang, P.; Lyu, Y.; Chang, E.; Jeong, H.; Xiang, L. Performance evaluation of robotic harvester with integrated real-time perception and path planning for dwarf hedge-planted apple orchard. *Agriculture* **2025**, *15*, 1593. [CrossRef]
30. Yan, B.; Quan, J.; Yan, W. Three-dimensional obstacle avoidance harvesting path planning method for apple-harvesting robot based on improved ant colony algorithm. *Agriculture* **2024**, *14*, 1336. [CrossRef]
31. Liu, Z.; Rasika, D.; Abeyrathna, R.M.; Sampurno, R.M.; Nakaguchi, V.M.; Ahamed, T. A new occlusion-avoidance and dual-view fruit localization method with a 6-DoF manipulator for orchard harvesting. *Comput. Electron. Agric.* **2025**, *237*, 110634. [CrossRef]
32. Sun, Q.; Chai, X.; Zeng, Z.; Zhou, G.; Sun, T. Noise-tolerant RGB-D feature fusion network for outdoor fruit detection. *Comput. Electron. Agric.* **2022**, *198*, 107034. [CrossRef]
33. Ultralytics. Metrics Reference: Batch_Probiou. Available online: https://docs.ultralytics.com/reference/utils/metrics/ (accessed on 13 July 2026).

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.