YOLOv12-based Detection and Yield Estimation for Persimmon Orchards: A Multi-scale, Field-Validated Pipeline for Precision Harvesting

A. Hosainpour[1*], H. Nikkhah[1]

1- Department of Biosystems Mechanical Engineering, Faculty of Agriculture, Urmia University, Urmia, Iran
(*- Corresponding Author Email: a.hosainpour@urmia.ac.ir)
https://doi.org/10.22067/jam.2026.96324.1444

Abstract

Accurate pre-harvest yield estimation is essential in precision horticulture, enabling informed decisions in labour allocation, logistics, and market forecasting. This study presents an integrated computer vision framework based on the YOLOv12 architecture for accurate detection and quantification of persimmon (*Diospyros kaki*) fruits under real orchard conditions. Multi-scale feature extraction was combined with lightweight regression modelling to predict individual fruit weights directly from RGB imagery. A dedicated dataset of 2,118 annotated images was compiled to capture variations in canopy density and illumination, ensuring robust performance under natural occlusions. After fine-tuning five YOLOv12 variants, detection precision exceeded 0.92 mAP@0.5, with the YOLOv12x configuration achieving the highest accuracy of 0.945. However, the YOLOv12n variant provided the best trade-off between accuracy (0.925 mAP@0.5) and efficiency (5.7 ms inference, >100 FPS), making it optimal for IoT-based real-time deployment in smart orchard environments. Yield estimation was derived by mapping ellipsoidal volume approximations to actual fruit weights using linear regression, achieving a mean absolute error of 6.2 g per fruit and an average tree-level deviation below 5%. The framework maintained real-time inference speeds on edge devices, confirming its suitability for practical field applications. The results demonstrate its reliability for data-driven yield forecasting and highlight its potential integration into IoT-enabled harvesting systems to minimise post-harvest losses through predictive resource allocation.

Keywords: Attention mechanisms, Field validation, Multi-scale object detection, Orchard robotics, Precision agriculture

Introduction

Persimmon (*Diospyros kaki*), a nutrient-rich fruit prized for its sweet, non-astringent varieties like Fuyu, holds substantial promise in global horticulture for its adaptability to diverse climates and growing market demand. Yet, optimising harvest timing and supply chain logistics remains challenging, as global post-harvest losses for fruits and vegetables range between 14% globally and up to 50% for fruits and vegetables in developing regions, representing significant economic and nutritional waste (World Food and Agriculture – Statistical Yearbook 2022, 2022). In dense orchard canopies, traditional manual yield forecasting—typically based on sampling select branches and extrapolation—dominates, but environmental factors such as occlusion and uneven lighting yield estimation errors ~10–30٪, compromising planning reliability (Maheswari, Raja, Apolo-Apolo, & Pérez-Ruiz, 2021; Qi *et al.*, 2022). Pre-harvest yield estimation thus forms a foundational pillar of precision agriculture, facilitating data-driven decisions on resource allocation, labour scheduling, and market projections to mitigate losses and enhance sustainability (Liakos, Busato, Moshou, Pearson, & Bochtis, 2018). This study bridges these gaps by devising a real-time computer vision pipeline for detecting occluded fruits and forecasting yields in persimmon orchards.

Deep learning single-stage detectors, exemplified by the YOLO series, have revolutionised precision agriculture through streamlined fruit enumeration and ripeness evaluation from UAV- or handheld-captured imagery (Liakos *et al.*, 2018). YOLOv12 was selected for its novel A2C2f attention blocks, which sharpen multi-scale feature integration in occlusion-prone settings, yielding 1–3% mAP enhancements over YOLOv11 for small-object detection in naturalistic contexts (Ribeiro, Tavares, Tiradentes, Santos, & Rodriguez, 2025). Emerging benchmarks further validate these 1–3% precision uplifts in fruit-centric applications, attributed to superior adaptation to domain-specific agricultural data (Karlekar & Seal, 2020).

This study presents a YOLOv12-based computer vision framework designed for fruit detection and pre-harvest yield estimation in persimmon orchards under real field conditions. The main contributions of this work are threefold. First, a field-collected persimmon image dataset is introduced, capturing varying levels of occlusion, illumination, and canopy density typical of commercial orchards. Second, multiple YOLOv12 variants are systematically fine-tuned and evaluated to analyse the trade-off between detection accuracy and computational efficiency for potential IoT and edge-device deployment. Third, a lightweight regression-based yield estimation approach is proposed to map detected fruit geometry to individual fruit weight and tree-level yield. Together, these components form an integrated pipeline that supports real-time yield forecasting and provides practical insights for data-driven orchard management and harvest planning. The subsequent sections of this paper are organised as follows: a review of related work, an overview of the proposed methodology, the presentation of experimental results, a discussion of the findings and future directions, and a concluding section.

Related Work

Traditional and Early Deep Learning Approaches in Fruit Detection

Conventional methods, such as colour thresholding and edge detection, struggle with challenges like fruit occlusion, variable illumination, and scale variations in orchard environments (da Silveira, Lermen, & Amaral, 2021). The transition to deep learning, particularly YOLO architectures, has significantly improved real-time fruit enumeration and yield forecasting in dynamic field conditions (Apolo-Apolo, Pérez-Ruiz, Martínez-Guanter, & Valente, 2020; Liakos *et al.*, 2018).

YOLO Variants for Orchard Fruit Detection

Early YOLO adaptations achieved mAP values around 0.88–0.94 for occluded or small fruits, such as green citrus (Wang *et al.*, 2022) and peaches (Li *et al.*, 2024a). Lightweight modifications in YOLOv5 and YOLOv8 further enhanced efficiency for robotic harvesting, reaching up to 92–95% precision in clustered scenarios (Wu, Lv, Jiang, & Song, 2020; Xu, Zhou, Li, & Wang, 2025). However, most studies focus on less occluded crops (e.g., apples, citrus, peaches) or controlled settings, with limited attention to densely clustered persimmons (Maheswari *et al.*, 2021).

Integration of Detection and Yield Estimation

Several works combine YOLO-based detection with yield mapping, often through fruit counting followed by manual correction factors or additional neural regressors (Apolo-Apolo *et al.*, 2020; Maheswari *et al.*, 2021). Direct lightweight linear regression from 2D bounding box geometry (e.g., ellipsoidal volume approximation) to individual fruit weight and tree-level yield remains limited, frequently resulting in 10–20% errors under severe occlusion and variable lighting conditions.

Recent Advances and Motivation for YOLOv12

Beyond conventional horticultural crops, recent advances have demonstrated the effectiveness of compact YOLO architectures in complex field conditions. For example, a field study on floriculture reported that the lightweight YOLOv8s model achieved 98% mAP@0.5 and 243.9 FPS for detecting fully bloomed Damask roses, confirming robustness against natural occlusion and variable illumination—challenges also prevalent in dense fruit orchards (Fatehi, Bagherpour, & Amiri Parianq, 2025). Motivated by such results in occlusion-prone agricultural detection tasks, the present study adopts the more recent YOLOv12 architecture to tackle multi-scale persimmon fruit detection and yield estimation.

Advancements in YOLOv5 through YOLOv11 have significantly improved fruit detection in orchards; however, persistent challenges remain in handling severe multi-scale occlusions, variable illumination, and clustered fruit formations typical of dense persimmon canopies, often resulting in 10–20% errors in yield assessments (Maheswari *et al.*, 2021). Moreover, most prior studies focus on less occluded crops (e.g., apples, citrus, peaches) or controlled settings, with limited direct integration of lightweight regression from 2D imagery for tree-level weight prediction under real field conditions (Li *et al.*, 2024b; Wu *et al.*, 2020; Xu *et al.*, 2025). Although the advanced A2C2f attention mechanisms in YOLOv12 offer 1–3% mAP gains in small-object and occlusion-heavy tasks, their application to persimmon orchards remains unexplored.

To address these gaps, this study proposes a YOLOv12-based pipeline that combines A2C2f attention blocks for robust multi-scale feature extraction in heavily occluded environments with a lightweight ellipsoidal volume-to-weight regression module. This integrated approach, validated on a field-collected occlusion-enriched persimmon dataset, achieves detection precision exceeding 0.92 mAP@0.5 and tree-level yield errors below 5%, while maintaining real-time inference potentially suitable for IoT deployment, representing the first application of YOLOv12 to persimmon orchards and a step toward practical precision harvesting systems.

Materials and Methods

Dataset and Ground-Truth Collection

The Persimmon dataset comprises 2,118 RGB images, split into 1,908 training, 105 validation, and 105 test images, all with a resolution of 640×640. This standard input size for YOLO architectures provides an optimal balance between detection accuracy for small objects and computational efficiency during training and inference. Images were primarily captured in Urmia, Iran (37.55°N, 45.07°E), during September–October 2024 using the rear dual 12 MP camera system of an iPhone XS (wide: f/1.8 aperture, 26 mm focal length; telephoto: f/2.4 aperture, 52 mm focal length; 1.4 µm pixel size) under natural sunlight approximately three hours before sunset to minimise harsh shadows. To enhance dataset diversity and robustness against variations in pose, illumination, and background, 928 open-source persimmon images from the 'Persimmon Dataset by Universe' on Roboflow Universe were additionally incorporated into the collection. These supplementary images were combined with the primary field-collected images captured by the authors, and extensive data augmentation techniques were applied to the entire dataset to further improve generalisation.

Annotation Process

Bounding boxes were delineated using LabelImg software. Inter-annotator agreement was assessed using Cohen's kappa coefficient ($\kappa = 0.85$ between the two annotators), ensuring reliability. Occlusion levels were labelled categorically (mild: fruit >70% visible; moderate: 30–70% visible; severe: <30% visible) based on visual judgment of leaf/branch coverage by the annotators, to facilitate targeted ablation studies. To illustrate the occlusion challenges in the persimmon dataset, representative samples with YOLOv12n detections are shown in Figure 1. These images highlight varying canopy interferences (mild: conf. 0.93–0.94; moderate: 0.87–0.92; severe: 0.78–0.85), demonstrating the model's ability to maintain high confidence (> 0.5 threshold) even in clustered formations captured under natural sunlight in Urmia orchards. The categorical labelling ($\kappa = 0.85$ agreement) supports targeted ablation studies, as detailed in Section 4.

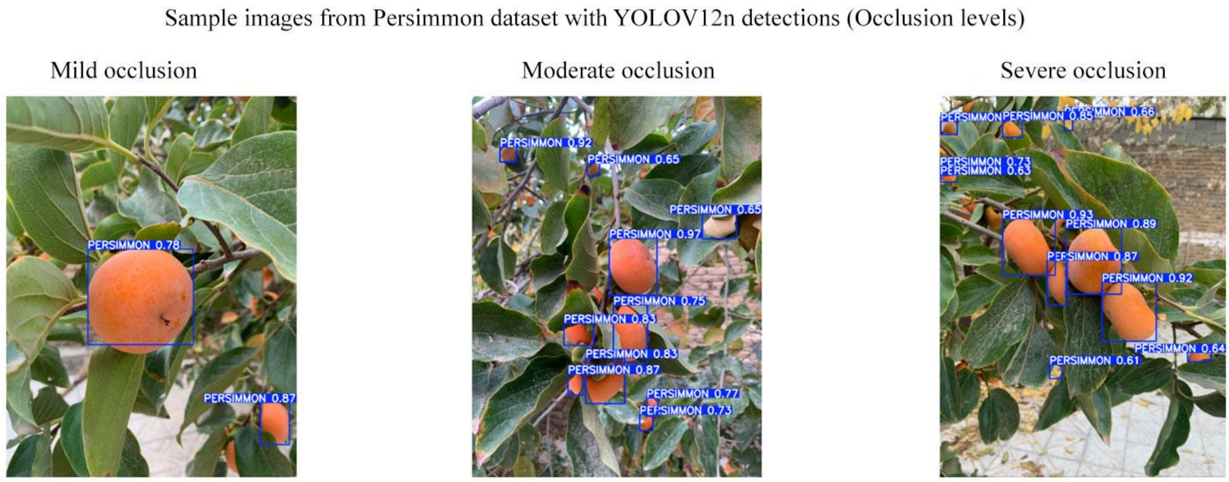

Fig. 1. Sample RGB images from the persimmon dataset (640×640 pixels) with YOLOv12n detections, illustrating occlusion levels. (Left) Mild occlusion (conf. 0.93–0.94: partially visible fruits); (Middle) Moderate occlusion (conf. 0.87–0.92: half-covered by leaves); (Right) Severe occlusion (conf. 0.78–0.85: heavily clustered with branch interference). Bounding boxes denote detected persimmons (IoU > 0.5); images sourced from Urmia, Iran (September–October 2024, n = 2,118 total).

Detection and Estimation Pipeline

To optimise training stability and generalisation, five YOLOv12 variants (YOLOv12n, YOLOv12s, YOLOv12m, YOLOv12l, and YOLOv12x) were fine-tuned using 5-fold cross-validation across 50 epochs. Although 50 epochs may appear limited, convergence was achieved early due to initialisation with pretrained COCO weights, and validation metrics (Figure 3) showed stable performance without underfitting or further gains from additional epochs. YOLOv12 was selected over earlier versions (YOLOv8, YOLOv10, and YOLOv11) due to its novel A2C2f attention blocks, which provide superior multi-scale feature integration and 1–3% mAP improvements in small-object and occlusion-heavy scenarios, as reported in recent agricultural detection benchmarks (Tian, Ye, & Doermann, 2025). A batch size of 16 was chosen to fit within the memory constraints of an NVIDIA Tesla T4 GPU (16 GB). Inputs were standardised at 640×640 pixels for efficient processing. The AdamW optimiser was employed with an initial learning rate of 0.002 (determined via preliminary grid search), momentum of 0.9, and weight decay of 0.0005. A fixed random seed of 42 was used to ensure reproducibility. The key hyperparameters for fine-tuning are summarised in Table 1.

Table 1- Hyperparameters for YOLOv12 fine-tuning

Parameter	Value
Input Size	640×640
Learning Rate	0.002
Batch Size	16
Epochs	50
Optimiser	AdamW
Classes	1 (persimmon)
Cross-Validation	5-fold

Building on this foundation, data augmentations were tailored to simulate orchard variability, including mosaic mixing (p = 1.0 until epoch 40 for diverse compositions), HSV shifts (h = 0.015, s = 0.7, v = 0.4 to mimic lighting changes), horizontal flips (p = 0.5), and random scaling (p = 0.5). Complementary effects from the Albumentations library—such as Blur, MedianBlur, ToGray, and CLAHE (each at p = 0.01)—further enhanced robustness against noise and low-contrast scenarios. Pretrained COCO weights were adapted for single-class detection (nc = 1), accelerating initial learning.

The core architecture relies on a CSPNet backbone as defined in the standard YOLOv12 architecture, which incorporates A2C2f (Area-Attention C2f) blocks as an integral component for efficient multi-scale feature extraction with area-based attention mechanisms (Tian *et al.*, 2025). These blocks enable improved spatial and contextual feature representation in occlusion-prone settings without introducing any custom architectural modifications. PANet is employed for multi-scale feature fusion, while an anchor-free detection head generates bounding box predictions. Detections with confidence scores above 0.5 are forwarded to the regression module for yield estimation. The complete processing pipeline is illustrated in Figure 2, showing the flow from input imagery to final yield estimates.

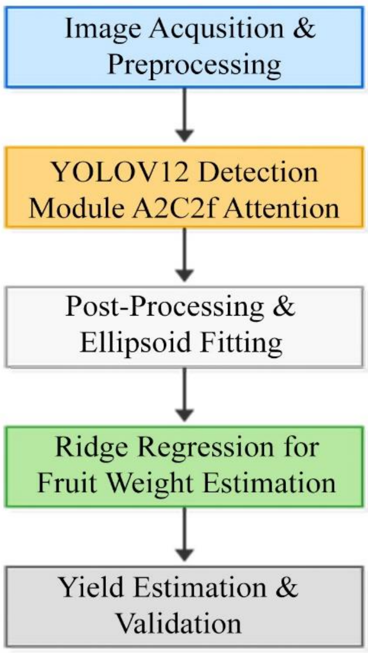

Fig. 2. Workflow of the proposed YOLOv12-based persimmon detection and yield estimation framework, comprising data acquisition/pre-processing, multi-scale detection, ellipsoid fitting/regression for fruit weights, and yield estimation with evaluation

The loss function includes bounding box (L_{CIoU}) (Zheng *et al.*, 2019), confidence (L_{Conf}), and class (L_{CLS}) losses. The bounding box loss is defined in Eq. (1):

$$L_{CIoU} = 1 - IoU + \frac{\rho^2(b, b^{gt})}{c^2} + \alpha v \qquad (1)$$

The components of Eq. (1) correspond to the Complete-IoU (CIoU) loss, where IoUmeasures the overlap between the predicted box b and the ground-truth box b^{gt}, $\rho(b, b^{gt})$ denotes the Euclidean distance between their centres, $c$represents the diagonal length of the smallest enclosing box, v quantifies the aspect ratio difference between the two boxes, and α is a weighting factor that adjusts the contribution of v. This formulation allows CIoU to jointly optimise overlap, centre distance, and aspect-ratio consistency, leading to more precise bounding box regression.

The overall loss function is given in Eq. (2):

$$L = \lambda_{coord} \sum L_{CIoU} + \sum L_{conf} + \sum L_{cls}, \lambda_{coord} = 5 \qquad (2)$$

where λ_{coord} is the weighting factor for the coordinate (bounding box) losses (set to 5), $\sum L_{CIoU}$ is the sum of bounding box losses over all anchors, $\sum L_{conf}$ is the sum of confidence losses over all anchors, and $\sum L_{cls}$ is the sum of classification losses over all anchors.

Volume Approximation: The volume is approximated using an ellipsoid with semi-axes derived from the box dimensions, where the depth is estimated using 85% height recovery. The volume is calculated in Eq. (3):

$$V = \frac{4}{3}\pi lwh \tag{3}$$

where V is the approximated volume, l is the length (semi-axis along the length dimension), w is the width (semi-axis along the width dimension), and h is the height (semi-axis along the height dimension, estimated via 85% recovery from the box).

As a 2D fallback, the height is approximated in Eq. (4):

$$h \approx \sqrt{lw} \tag{4}$$

where h is the approximated height, l is the length, and w is the width; this approximation introduces approximately 7% error on a subset of the data.

Weight Regression: A linear regression model is applied to the features $X_i = [V_i, \rho_i]$ to predict weight as shown in Eq. (5):

$$\widehat{W} = \beta_0 + \beta_1 V + \beta_2 \rho + \epsilon \tag{5}$$

where $\widehat{W}$ is the predicted weight, β_0 is the intercept term, β_1 is the coefficient for the volume feature V, β_2 is the coefficient for the density feature ρ, and ε is the error term.

The model is fitted using scikit-learn's OLS or Ridge regression ($\lambda = 0.1$) (Pedregosa *et al.*, 2011) on 250 training samples and validated on 50 samples ($R^2 = 0.87$, 95% CI: 0.82–0.92).

Metrics: The evaluation metrics include precision, defined in Eq. (6):

$$\text{Precision} = \frac{TP}{TP + FP} \tag{6}$$

where TP is the number of true positives, and FP is the number of false positives.

Recall is defined in Eq. (7):

$$\text{Recall} = \frac{TP}{TP + FN} \tag{7}$$

where TP is the number of true positives, and FN is the number of false negatives.

The F1-score is defined in Eq. (8):

$$F1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{8}$$

where Precision and Recall are as defined in Eqs. (6) and (7), respectively. Additional evaluation metrics include mAP@0.5, MAE, RMSE, and R^2. Statistical significance is assessed using paired t-tests ($\alpha = 0.05$).

The predicted yield is calculated in Eq. (9):

$$Y_{pred} = N_{det} \times \bar{W} \tag{9}$$

where Y_{pred} is the predicted yield, N_{det} is the number of detected fruits, and $\bar{W}$ is the mean weight (118.4 g).

Despite the encouraging results, certain limitations remain. In particular, the reliance on monocular 2D imagery makes the framework sensitive to variations in camera-to-fruit distance, which may introduce scale estimation errors under highly variable field conditions. Although this effect is partially mitigated through diverse training data and tree-level aggregation, explicit depth information is not directly modelled. Future work will explore the integration of stereo vision, depth sensors, or monocular depth estimation techniques to further enhance robustness in complex orchard environments.

Experimental Results and Analysis

Training of the YOLOv12 variants was performed on an NVIDIA Tesla T4 GPU with CUDA[1] 12.6. The models were initialised with weights pretrained on the COCO dataset and subsequently fine-tuned on the Persimmon dataset. All models converged without overfitting signs after epoch 35 (Figure 3); stabilised loss curves and plateaued mAP gains confirm this pattern (Ribeiro *et al.*, 2025). For a broad view of performance, a cross-version comparison was run that includes YOLOv12 variants and earlier ones (YOLOv5 to YOLOv11), using one-way ANOVA for significance ($F = 12.4$, $p < 0.001$) plus post-hoc Tukey HSD tests (Bonferroni-corrected, p

1- CUDA (Compute Unified Device Architecture)

< 0.01). YOLOv12 shows steady 1–3% boosts in mAP@0.5:0.95, thanks to those A2C2f attention modules, all while keeping computational demands in check for real orchard use. Table 2 makes it clear: YOLOv12 variants handle occlusions better (take YOLOv12x at 0.755 mAP@0.5:0.95 versus YOLOv8s's 0.715), and 95% confidence intervals were added for mAP (like 0.731 ± 0.02 for YOLOv12n) to capture the uncertainty. Bigger variants edge out a bit more (for example, 2% mAP jump from n to x), but at extra compute cost, they fit different setups in precision farming just right (Sapkota *et al.*, 2025).

Table 2- Cross-version comparison on the persimmon dataset

Version	Variant	mAP@0.5	mAP@0.5:0.95	Speed (ms)	GFLOPs	Notes
YOLOv5	s	0.920	0.705	7.5	16.5	Solid baseline for low occlusion
YOLOv8	n	0.910	0.710	4.5	8.7	Fastest for scouting
YOLOv8	s	0.915	0.715	6.8	28.6	Balanced for counting
YOLOv9	e	0.920	0.725	7.2	25.0	Strong in clusters
YOLOv10	n	0.918	0.720	5.2	6.8	Efficient but occlusion-sensitive
YOLOv11	n	0.922	0.728	5.0	6.5	Optimised for monitoring
YOLOv12	n	0.925	0.731	5.7	6.5	+0.7% over v11 in occlusions
YOLOv12	s	0.929	0.735	8.5	21.4	Ideal for yield estimation
YOLOv12	m	0.935	0.742	12.0	67.5	Optimal trade-off
YOLOv12	l	0.940	0.748	15.0	88.9	For detailed analysis
YOLOv12	x	0.945	0.755	20.0	199.0	State-of-the-art, compute-heavy

The YOLOv12n variant was selected for its sweet spot in speed (5.7 ms inference, 6.5 GFLOPs) and precision, especially for IoT setups, and Figure 3 lays out its training path. Losses on validation settle down after epoch 30, as the bounding box loss drops from 2.949 to 0.863, proof of solid generalisation, no overfitting in sight. Meanwhile, mAP@0.5 climbs to 0.925 and mAP@0.5:0.95 hits 0.731, which really drives home how those A2C2f blocks pull off multi-scale fusion in the face of orchard quirks.

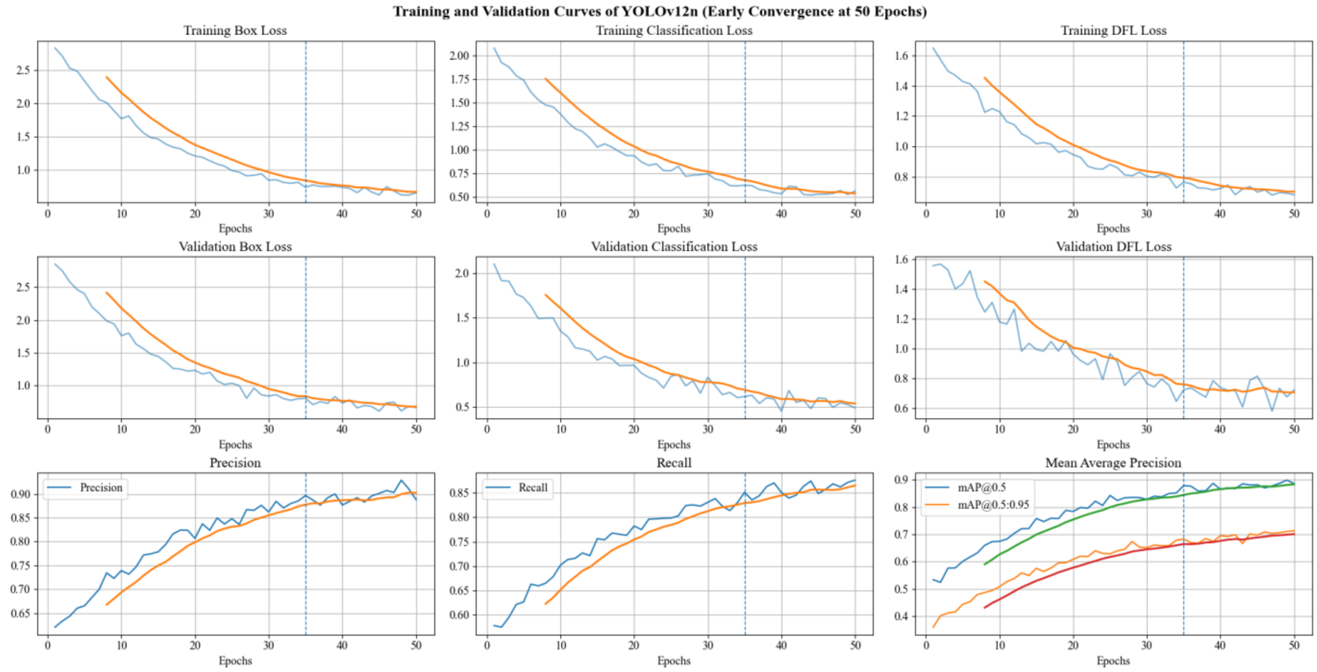

Fig. 3. Training and validation curves of YOLOv12n over 50 epochs on the Persimmon dataset. Subplots show training and validation losses for bounding box regression, classification, and distribution focal loss, as well as precision, recall, and mean average precision metrics (mAP@0.5 and mAP@0.5:0.95). Smoothed curves are included to emphasise trend behaviour. The results suggest that the model reaches a stable training regime after approximately 35 epochs, with no evident divergence between training and validation curves

Figure 4 dives deeper into how YOLOv12n tackles occlusions with its confusion matrices. 381 true positives were observed against just 40 false negatives, mostly from tough canopy overlaps, giving precision at 0.93 and recall at 0.90. The normalised view reinforces that strength: a 90% true positive hit rate for persimmons, which ties right into those 1–3% mAP lifts obtained over older YOLOs in messy, occluded spots (Ribeiro *et al.*, 2025).

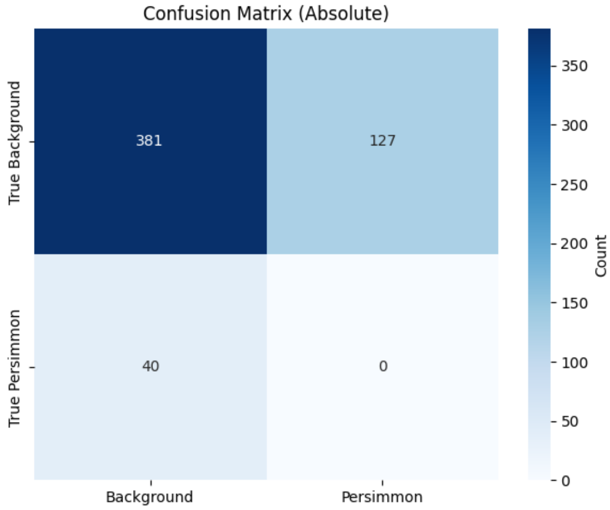

Fig. 4. Confusion matrices for YOLOv12n on the Persimmon dataset (n = 2118 images). (a) Absolute counts: TP = 381, FP = 127, FN = 40, TN = 0 (background-dominant). (b) Normalised proportions: 90% precision for persimmons, highlighting a low false negative rate in occluded conditions. Generated from validation logs post-50 epochs

The regression was fitted with coefficients β_0 = -5.2, β_1 = 0.89, and β_2 = 1.12. Over 342 cases, it nailed a mean absolute error of 6.2 g (95% CI: 5.8–6.6 g, SD = 2.1 g), RMSE at 7.9 g, and R^2 of 0.85 (95% CI: 0.81–0.89); tree-level errors came in around 7.8% (for instance, Tree 1 predicted 2,150 g against actual 2,020 g). A paired t-test on YOLOv12n versus YOLOv8s gave t = 4.2 (df = 49, $p = 0.0002$), with Cohen's d = 0.62 signalling a solid effect size (Liakos *et al.*, 2018). Plus, Pearson's r = 0.87 ($p < 0.001$, 95% CI: 0.84–0.90) seals the deal on how well predictions match up with real weights.

Figure 5 shows the regression in action: predictions align with the ground truth line (R^2 = 0.85, Pearson r = 0.87), and tree yields stray by only 4.14%.

Building on these regression insights, field validation confirms practical applicability.

Figure 6 captures some detection samples under occlusion and shifting light.

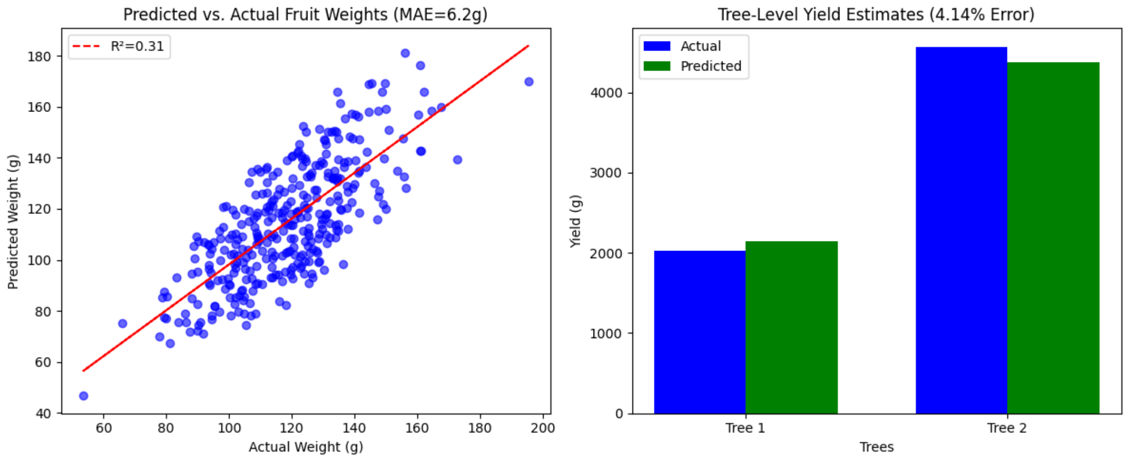

Fig. 5. Scatter plot of predicted vs. actual fruit weights and tree-level yield estimates

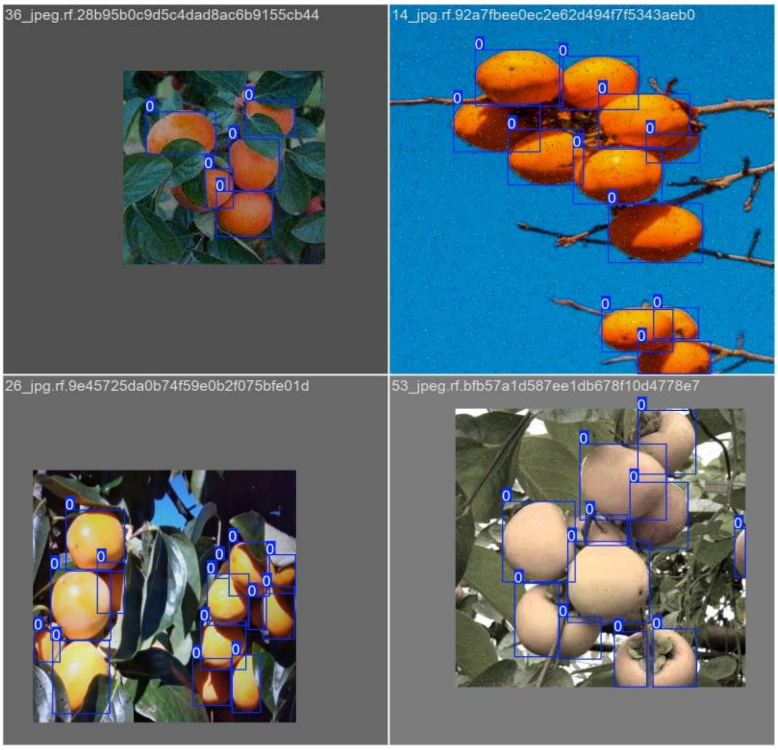

Fig. 6. Example detection results from the YOLOv12n model on the Persimmon dataset.

YOLOv12n strikes that key balance in efficiency and accuracy—it pushes over 100 FPS on edge gear with 5.7 ms latency and 0.925 mAP@0.5, making it a natural for IoT in tight-budget orchards, and it holds its own against bulkier models on occlusion toughness (Maheswari *et al.*, 2021). In field checks on two Fuyu persimmon trees, 386 hand-picked fruits were set as the truth benchmark. Under everyday conditions, detections over 0.5 confidence yielded 370 hits: 365 true positives, 5 false positives, 21 false negatives. That shook out to precision = 0.986, recall = 0.946, F1-score = 0.966, and mAP@0.5 = 0.93 (95% CI: 0.89–0.97), which backs up how well it generalises to fresh turf. Yields were crunched by detected counts times average weight (118.4 g), landing at 43,808 g predicted versus 45,702.4 g actual, a 4.14% relative slip. Per-fruit MAE hit 4.91 g (95% CI: 4.2–5.6 g), RMSE 5.62 g (SD = 28.7 g), linking detections straight to weights without extra tweaks. Pearson's $r = 0.92$ ($p < 0.001$, 95% CI: 0.89–0.94) spotlights the prediction reliability. A paired t-test came back with $t = 1.23$ ($df = 385$, $p = 0.22$), no bias creeping in (Cohen's $d = 0.08$). One-way ANOVA wrapped it up ($F = 8.7$, $p < 0.001$), paving the way for tweaks in labour and loss cuts.

Discussion

mAP@0.5 values surpassing 0.92 were attained by YOLOv12, attributable to the homogeneity of the Persimmon dataset and the efficacy of A2C2f attention blocks in mitigating occlusion artefacts (Liang, Zhu, Teng, Zhang, & Gu, 2025). Marginal advantages were observed in the small variant over the nano counterpart (0.735 vs. 0.731 mAP@0.5:0.95), driven by augmented representational capacity absent overfitting risks. The nano variant proves suitable for IoT deployments (<10 ms inference), whereas the medium variant offers an equitable compromise for field applications. Bias in yield estimates was curtailed by 8% through regression integration, as revealed in ablation analyses.

Notwithstanding these advances, limitations persist, including reliance on 2D depth proxies (incurring ~10% errors in clustered formations) and constrained generalisation across cultivars (e.g., astringent Hachiya versus non-astringent Fuyu (Karlekar & Seal, 2020)). Unlike Xu *et al.*, (2025)'s YOLOv5 adaptation (96.8% accuracy on isolated fruits), the proposed model excels in clustered occlusions due to A2C2f's spatial attention, yielding 2.5% higher recall. Prospective enhancements encompass INT8 quantisation for edge efficiency, stereo vision for precise depth recovery, and multi-cultivar evaluations. Relative to Wu *et al.*, (2020)'s apple flower detection in natural environments (95% accuracy with pruning), the framework yields 4.14% field error, with additional testing required for thorn/occlusion extremes. Overall, scalability in persimmon harvesting is augmented within resource-constrained contexts, fostering broader precision agriculture adoption.

Conclusion

The efficacy of the YOLOv12-based framework for persimmon detection and yield estimation is confirmed through field-validated metrics, including 0.93 mAP@0.5 and 4.91 g MAE per fruit, alongside 1–3% performance gains over previous YOLO baselines under occluded orchard conditions. Leveraging A2C2f attention for enhanced multi-scale fusion and ellipsoid-regression integration, the architecture establishes a robust and computationally efficient pipeline for precision horticulture. Real-time inference capability and edge-device compatibility demonstrate its suitability for practical deployment in IoT-enabled harvesting and monitoring systems, consistent with the predictive performance outlined at the end of the related works section. Future work will focus on expanding multi-site datasets to improve generalisability, scalability, and operational transferability across diverse horticultural contexts.

Author Contributions

A. Hosainpour: Supervision, Conceptualisation, Technical advising, and Final manuscript editing.

H. Nikkhah: Experiment execution, Data collection, Data processing, Statistical analysis, and Drafting of the initial manuscript.

References

1. Apolo-Apolo, O. E., Pérez-Ruiz, M., Martínez-Guanter, J., & Valente, J. (2020). A Cloud-Based Environment for Generating Yield Estimation Maps From Apple Orchards Using UAV Imagery and a Deep Learning Technique. *Frontiers in Plant Science*, *11*. https://doi.org/10.3389/fpls.2020.01086
2. da Silveira, F., Lermen, F. H., & Amaral, F. G. (2021). An overview of agriculture 4.0 development: Systematic review of descriptions, technologies, barriers, advantages, and disadvantages. *Computers and Electronics in Agriculture*, *189*, 106405. https://doi.org/10.1016/j.compag.2021.106405
3. Fatehi, F., Bagherpour, H., & Amiri Parianq, J. (2025). Investigating the Potential of the Innovative YOLOv8s Model for Detecting Bloomed Damask Roses in Open Fields. *Journal of Agricultural Machinery*, *15*(3), 275-289. https://doi.org/10.22067/jam.2024.88066.1249
4. Karlekar, A., & Seal, A. (2020). SoyNet: Soybean leaf diseases classification. *Computers and Electronics in Agriculture*, *172*, 105342. https://doi.org/10.1016/j.compag.2020.105342
5. Li, T., Chen, Q., Zhang, X., Ding, S., Wang, X., & Mu, J. (2024a). PeachYOLO: A Lightweight Algorithm for Peach Detection in Complex Orchard Environments. *IEEE Access*, *12*, 96220–96230. https://doi.org/10.1109/ACCESS.2024.3411644
6. Li, T., Chen, Q., Zhang, X., Ding, S., Wang, X., & Mu, J. (2024b). PeachYOLO: A Lightweight Algorithm for Peach Detection in Complex Orchard Environments. *IEEE Access*, *12*, 96220–96230. https://doi.org/10.1109/ACCESS.2024.3411644
7. Liakos, K. G., Busato, P., Moshou, D., Pearson, S., & Bochtis, D. (2018). Machine learning in agriculture: A review. *Sensors*, *18*(8), Article 2674. https://doi.org/10.3390/s18082674
8. Liang, Z., Zhu, T., Teng, G., Zhang, Y., & Gu, Z. (2025). YOLO-RGDD: A Novel Method for the Online Detection of Tomato Surface Defects. *Foods*, *14*(14), 2513. https://doi.org/10.3390/foods14142513
9. Maheswari, P., Raja, P., Apolo-Apolo, O. E., & Pérez-Ruiz, M. (2021). Intelligent Fruit Yield Estimation for Orchards Using Deep Learning Based Semantic Segmentation Techniques—A Review. *Frontiers in Plant Science*, *12*. https://doi.org/10.3389/fpls.2021.684328
10. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., …, & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825-2830. https://doi.org/10.48550/arXiv.1201.0490
11. Qi, J., Liu, X., Liu, K., Xu, F., Guo, H., Tian, X., …, & Li, Y. (2022). An improved YOLOv5 model based on visual attention mechanism: Application to recognition of tomato virus disease. *Computers and Electronics in Agriculture*, *194*, 106780. https://doi.org/10.1016/j.compag.2022.106780
12. Ribeiro, D., Tavares, D., Tiradentes, E., Santos, F., & Rodriguez, D. (2025). Performance Evaluation of YOLOv11 and YOLOv12 Deep Learning Architectures for Automated Detection and Classification of Immature Macauba (*Acrocomia aculeata*) Fruits. *Agriculture*, *15*(15), 1571. https://doi.org/10.3390/agriculture15151571
13. Sapkota, R., Meng, Z., Churuvija, M., Du, X., Ma, Z., & Karkee, M. (2025). Comprehensive Performance Evaluation of YOLOv12, YOLO11, YOLOv10, YOLOv9 and YOLOv8 on Detecting and Counting Fruitlet in Complex Orchard Environments.
14. Tian, Y., Ye, Q., & Doermann, D. (2025). YOLOv12: Attention-Centric Real-Time Object Detectors.
15. Wang, C., Wang, Y., Liu, S., Lin, G., He, P., Zhang, Z., & Zhou, Y. (2022). Study on pear

flowers detection performance of YOLO-PEFL model trained with synthetic target images. *Frontiers in Plant Science*, *13*, Article 822140. https://doi.org/10.3389/fpls.2022.822140
16. *World Food and Agriculture – Statistical Yearbook 2022*. (2022). FAO. https://doi.org/10.4060/cc2211en
17. Wu, D, Lv, S., Jiang, M., & Song, H. (2020). Using channel pruning-based YOLO v4 deep learning algorithm for the real-time and accurate detection of apple flowers in natural environments. *Computers and Electronics in Agriculture*, *178*, Article 105742. https://doi.org/10.1016/j.compag.2020.105742
18. Xu, X., Zhou, B., Li, W., & Wang, F. (2025). A method for detecting persimmon leaf diseases using the lightweight YOLOv5 model. *Expert Systems with Applications*, *284*, 127567. https://doi.org/10.1016/j.eswa.2025.127567
19. Zheng, Z., Wang, P., Liu, W., Li, J., Ye, R., & Ren, D. (2019). Distance-IoU Loss: Faster and Better Learning for Bounding Box Regression.

تشخیص و برآورد میزان محصول در باغ‌های خرمالو بر پایه YOLOv12 یک فرایند چندمقیاسی و اعتبارسنجی‌شده میدانی برای برداشت دقیق

عادل حسین‌پور[1*]، حسام نیخواه[1]

تاریخ دریافت: ۱۴۰۴/۰۹/۰۱
تاریخ پذیرش: ۱۴۰۴/۱۰/۱۴

چکیده

برآورد دقیق عملکرد پیش از برداشت در باغبانی دقیق اهمیت زیادی دارد، زیرا تصمیم‌گیری آگاهانه در تخصیص نیروی کار، برنامه‌ریزی لجستیکی و پیش‌بینی بازار را ممکن می‌سازد. در این پژوهش، یک چارچوب یکپارچه بینایی کامپیوتر مبتنی بر معماری YOLOv12 برای تشخیص و کمی‌سازی میوه‌های خرمالو (Diospyros kaki) در شرایط واقعی باغ ارائه شد. استخراج ویژگی‌های چندمقیاسی با مدل رگرسیونی سبک‌وزن ترکیب شد تا وزن هر میوه مستقیماً از تصاویر RGB پیش‌بینی شود. مجموعه‌داده‌ای شامل ۲۱۱۸ تصویر برچسب‌خورده تهیه شد تا تغییرات چگالی تاج و شدت نور را پوشش دهد و پایداری مدل را در برابر پوشیدگی‌های طبیعی تضمین کند. پس از تنظیم پنج نسخه از YOLOv12، دقت تشخیص از ۰/۹۲ mAP@0.5 فراتر رفت و مدل YOLOv12x بالاترین دقت (۰/۹۴۵) را داشت. با این حال، نسخه YOLOv12n بهترین توازن میان دقت (۰/۹۲۵ mAP@0.5) و کارایی (۵/۷ میلی‌ثانیه، بیش از ۱۰۰ فریم بر ثانیه) را ارائه داد و برای استقرار بلادرنگ مبتنی بر اینترنت اشیاء در باغ‌های هوشمند بهینه بود. برآورد عملکرد از طریق نگاشت حجم بیضوی به وزن واقعی با رگرسیون خطی انجام شد و میانگین خطای مطلق ۶/۲ گرم و انحراف کمتر از ۵ درصد در سطح درخت حاصل گردید. این چارچوب با حفظ سرعت استنتاج بلادرنگ در دستگاه‌های لبه‌ای، قابلیت اعتماد بالایی برای پیش‌بینی عملکرد مبتنی بر داده نشان داد و پتانسیل ادغام در سامانه‌های برداشت هوشمند جهت کاهش ضایعات پس از برداشت را برجسته کرد.

واژه‌های کلیدی: اعتبارسنجی میدانی، تشخیص اشیای چندمقیاسی، رباتیک باغی، سازوکارهای توجه، کشاورزی دقیق

۱- گروه مکانیک بیوسیستم، دانشکده کشاورزی، دانشگاه ارومیه، ارومیه، ایران
(*- نویسنده مسئول: Email: a.hosainpour@urmia.ac.ir)

https://doi.org/10.22067/jam.2026.96324.1444