Comparative Evaluation of RGB, Depth, and Fused Input Modalities for Oil Palm Fresh Fruit Bunch Detection Using YOLOv11

Muhammad Zainal Muttaqin

January 29, 2026

Abstract

Automated detection of oil palm Fresh Fruit Bunches (FFBs) is essential for efficient harvesting operations. While RGB-based object detection has been widely studied, the potential contribution of depth information—either from hardware sensors or monocular depth estimation—remains underexplored for this domain. This paper presents a systematic comparison of six input configurations for FFB detection using YOLOv11n: RGB-only (A.1), real depth-only (A.2), early-fused RGB+real depth (A.3), synthetic depth-only (A.4a), early-fused RGB+synthetic depth (A.4b), and late fusion of dual backbones (A.5). All experiments are conducted under a unified augmentation protocol with five random seeds to ensure statistical robustness. Results show that early fusion of RGB with real depth (A.3) achieves the highest mAP50 of 0.8403±0.0161, marginally outperforming the RGB-only baseline (0.8385 ± 0.0249). Late fusion (A.5) ranks fourth (0.8084 ± 0.0304) with higher seed sensitivity, while depth-only configurations perform substantially worse, with synthetic depth (A.4a, 0.6533) falling below real depth (A.2, 0.7325). These findings indicate that RGB remains the dominant modality, early fusion outperforms late fusion, and depth fusion provides only marginal gains in the current experimental setting.

1 Introduction

Oil palm is one of the most economically important crops in Southeast Asia. Timely and accurate detection of Fresh Fruit Bunches (FFBs) on palm trees is a prerequisite for yield estimation, automated harvesting, and quality assessment. Recent advances in deep learning-based object detection, particularly the YOLO family of models, have enabled real-time detection from RGB imagery with competitive accuracy.

However, RGB images alone may be insufficient in challenging conditions such as occlusion, variable lighting, and cluttered canopy backgrounds. Depth information, whether captured by dedicated sensors (e.g., stereo cameras, LiDAR) or estimated from monocular images using foundation models such as Depth-Anything-V2, offers a complementary geometric signal that could improve detection robustness.

This study addresses the following research questions:

1. Does fusing depth information with RGB improve FFB detection accuracy compared to RGB alone?

2. How does real sensor-captured depth compare to synthetically estimated depth?
3. What is the effect of different fusion strategies (early 4-channel fusion vs. late feature-level fusion)?

We conduct six controlled experiments across five random seeds, enforcing uniform augmentation and evaluation protocols, to provide statistically grounded answers.

2 Methodology

2.1 Dataset

The dataset consists of RGB images and corresponding depth maps of oil palm trees captured with an Intel RealSense D435i stereo camera. Each image is annotated with bounding boxes for the single class `fresh_fruit_bunch`. The dataset is split into training, validation, and test sets with a 70:20:10 ratio. All evaluations are performed on the held-out test set.

Raw depth maps are recorded in millimeters (16-bit PNG). During preprocessing, depth values are clipped to the range 600–6000 mm (0.6–6.0 m), normalized to 0–255, and replicated across three channels to form a pseudo-RGB depth image compatible with standard YOLO input layers.

2.2 Model Architecture

All experiments use **YOLOv11n** (Nano variant) as the base detector. For 4-channel input experiments (A.3, A.4b), the first convolutional layer is modified to accept 4 input channels by zero-initializing the additional depth channel weights. For the late fusion experiment (A.5), two frozen YOLOv11n backbones (one for RGB, one for depth) extract features independently, which are concatenated at the neck level before the detection head.

2.3 Input Modalities

Six experimental configurations are evaluated:

- **A.1 – RGB Only:** Standard 3-channel RGB input (baseline).
- **A.2 – Real Depth Only:** 3-channel pseudo-RGB from real depth sensor.
- **A.3 – RGB + Real Depth (Early Fusion):** 4-channel input concatenating RGB and real depth.
- **A.4a – Synthetic Depth Only:** 3-channel pseudo-RGB from Depth-Anything-V2 Large estimates.
- **A.4b – RGB + Synthetic Depth (Early Fusion):** 4-channel input concatenating RGB and synthetic depth.
- **A.5 – Late Fusion:** Dual frozen backbones (RGB + real depth) with feature concatenation.

2.4 Training Protocol

All models are trained for **100 epochs** with early stopping (patience=30). Each experiment is repeated across **five seeds** (42, 123, 456, 789, 101). A geometry-only augmentation policy is applied uniformly: translate=0.1, scale=0.5, horizontal flip=0.5; HSV jitter, mosaic, and mixup are disabled to ensure fair comparison across modalities.

For depth-related experiments (A.2, A.3, A.4a, A.4b), a **BatchNorm reset** procedure is applied: after loading pretrained weights, a forward pass over 100 real training images recalibrates the running statistics to the depth domain.

2.5 Late Fusion Architecture (A.5)

Experiment A.5 employs a dual-backbone architecture. Two separately pretrained YOLOv11n backbones—one for RGB (from A.1) and one for real depth (from A.2)—are frozen. Their multi-scale feature maps (P3, P4, P5) are concatenated channel-wise and fed into a trainable detection neck and head. Only the neck and head parameters are updated during training.

3 Results

3.1 Main Comparison

Table 1 summarizes the mean and standard deviation of detection metrics across five seeds for all six experiments, ranked by mAP50.

Table 1: Summary comparison of all experiments (mean ± std over 5 seeds). Evaluated on test set.

Rank	Input	mAP50	mAP50–95	Precision	Recall
1	A.3: RGB + Real Depth	**0.8403±0.016**	0.3687±0.010	0.8019±0.050	**0.7743±0.018**
2	A.1: RGB Only	0.8385±0.025	0.3645±0.011	0.8028±0.046	0.7605±0.057
3	A.4b: RGB + Synth. Depth	0.8233±0.012	**0.3676±0.007**	0.7959±0.034	0.7387±0.040
4	A.5: Late Fusion	0.8084±0.030	0.3176±0.016	0.7620±0.030	0.7829±0.012
5	A.2: Real Depth Only	0.7325±0.042	0.2915±0.012	0.7147±0.038	0.7267±0.013
6	A.4a: Synth. Depth Only	0.6533±0.036	0.2754±0.025	0.7176±0.040	0.6170±0.082

3.2 Per-Seed Results

Tables 2–7 report individual seed results for each experiment.

Table 2: A.1 – RGB Only: per-seed results.

Seed	mAP50	mAP50–95	Precision	Recall
42	0.8809	0.3662	0.7571	0.8313
123	0.8324	0.3697	0.8684	0.6913
456	0.8148	0.3657	0.7717	0.8047
789	0.8325	0.3751	0.8296	0.7417
101	0.8317	0.3458	0.7873	0.7333
Mean±Std	**0.8385±0.025**	**0.3645±0.011**	**0.8028±0.046**	**0.7605±0.057**

Table 3: A.2 – Real Depth Only: per-seed results.

Seed	mAP50	mAP50–95	Precision	Recall
42	0.6908	0.2742	0.7001	0.7238
123	0.7169	0.2987	0.6682	0.7479
456	0.8005	0.3054	0.7636	0.7238
789	0.7138	0.2881	0.6996	0.7238
101	0.7402	0.2910	0.7419	0.7143
Mean±Std	**0.7325±0.042**	**0.2915±0.012**	**0.7147±0.038**	**0.7267±0.013**

Table 4: A.3 – RGB + Real Depth (Early Fusion): per-seed results.

Seed	mAP50	mAP50–95	Precision	Recall
42	0.8386	0.3549	0.8019	0.8000
123	0.8344	0.3618	0.7939	0.7810
456	0.8337	0.3798	0.7488	0.7666
789	0.8681	0.3701	0.8838	0.7524
101	0.8267	0.3770	0.7809	0.7714
Mean±Std	**0.8403±0.016**	**0.3687±0.010**	**0.8019±0.050**	**0.7743±0.018**

Table 5: A.4a – Synthetic Depth Only: per-seed results.

Seed	mAP50	mAP50–95	Precision	Recall
42	0.6066	0.2333	0.7244	0.4952
123	0.6309	0.2789	0.6483	0.7048
456	0.6707	0.2762	0.7408	0.6190
789	0.6577	0.2981	0.7291	0.5897
101	0.7009	0.2906	0.7452	0.6762
Mean±Std	**0.6533±0.036**	**0.2754±0.025**	**0.7176±0.040**	**0.6170±0.082**

Table 6: A.4b – RGB + Synthetic Depth (Early Fusion): per-seed results.

Seed	mAP50	mAP50–95	Precision	Recall
42	0.8210	0.3636	0.7489	0.7669
123	0.8149	0.3704	0.8019	0.7143
456	0.8103	0.3614	0.7775	0.6857
789	0.8410	0.3793	0.8381	0.7397
101	0.8293	0.3632	0.8130	0.7868
Mean±Std	**0.8233±0.012**	**0.3676±0.007**	**0.7959±0.034**	**0.7387±0.040**

Table 7: A.5 – Late Fusion: per-seed results.

Seed	mAP50	mAP50–95	Precision	Recall
42	0.7610	0.2977	0.7106	0.7714
123	0.7955	0.3120	0.7594	0.7810
456	0.8279	0.3390	0.7911	0.8000
789	0.8347	0.3205	0.7803	0.7714
101	0.8229	0.3188	0.7684	0.7905
Mean±Std	**0.8084±0.030**	**0.3176±0.016**	**0.7620±0.030**	**0.7829±0.012**

3.3 Seed Variability Analysis

Table 8 compares the standard deviation of mAP50 across seeds for each experiment. Lower standard deviation indicates greater stability with respect to random initialization.

Table 8: Seed stability comparison (mAP50 standard deviation).

Experiment	mAP50 Std	Rank (most stable)
A.4b (RGB + Synth. Depth)	0.0122	1
A.3 (RGB + Real Depth)	0.0161	2
A.1 (RGB Only)	0.0249	3
A.5 (Late Fusion)	0.0304	4
A.4a (Synth. Depth Only)	0.0363	5
A.2 (Real Depth Only)	0.0419	6

Fusion configurations (A.3, A.4b) exhibit the lowest variance, suggesting that combining modalities regularizes the model and reduces sensitivity to initialization. Depth-only experiments show the highest variance, indicating less stable optimization landscapes.

3.4 V1 vs. V2 Protocol Comparison

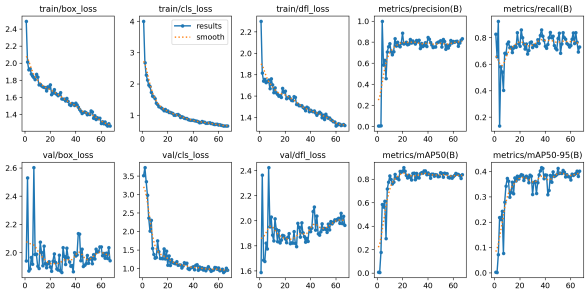
Table 9 compares results before (V1) and after (V2) enforcing uniform augmentation and BatchNorm reset.

Table 9: Effect of V2 protocol changes (uniform augmentation + BN reset).

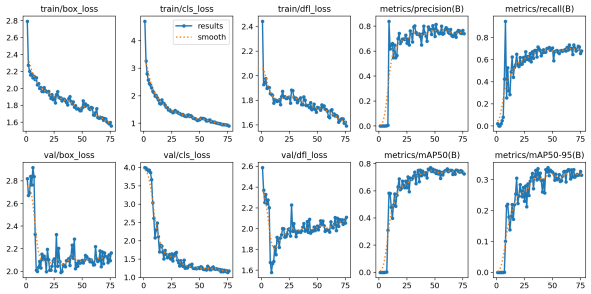
Experiment	V1 mAP50	V2 mAP50	Δ	Note
A.1 (RGB)	0.869	0.839	−3.5%	HSV disabled for cross-modality fairness
A.2 (Real depth)	0.748	0.733	−2.0%	BN reset added
A.3 (Real RGBD)	0.842	0.840	−0.2%	BN reset added
A.4a (Synth. depth)	0.708	0.653	−7.8%	Sensitive to augmentation changes
A.4b (RGB + synth.)	0.813	0.823	+1.2%	BN reset improves synthetic fusion
A.5 (Late fusion)	N/A	0.808	New	Dual frozen backbones + fusion layers

The V2 protocol reduces A.1 performance by 3.5% due to the removal of HSV augmentation, but this is necessary for a fair comparison since depth channels cannot benefit from color jitter. A.4b is the only experiment that improves under V2, suggesting BN reset is particularly beneficial when combining RGB with synthetic depth.

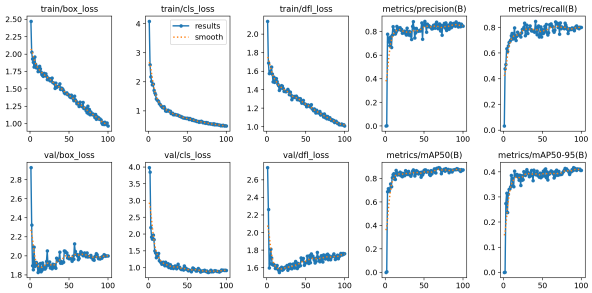
3.5 Training Curves and Diagnostic Plots



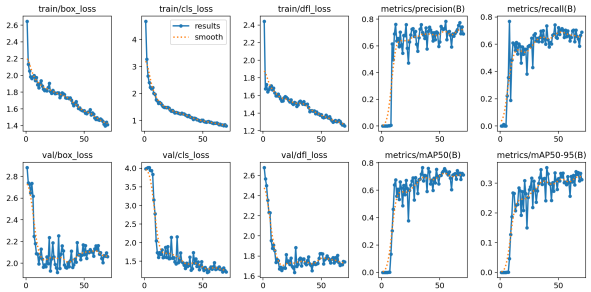
(a) A.1 – RGB Only (seed 42)



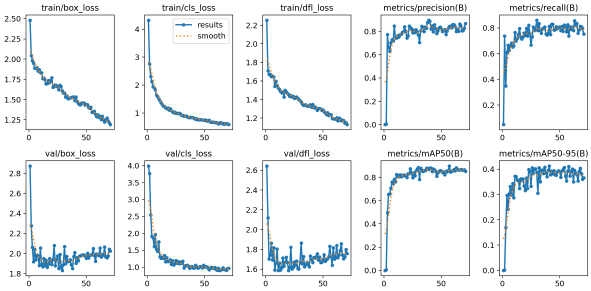
(b) A.2 – Real Depth Only (seed 456)



(c) A.3 – RGB + Real Depth (seed 789)

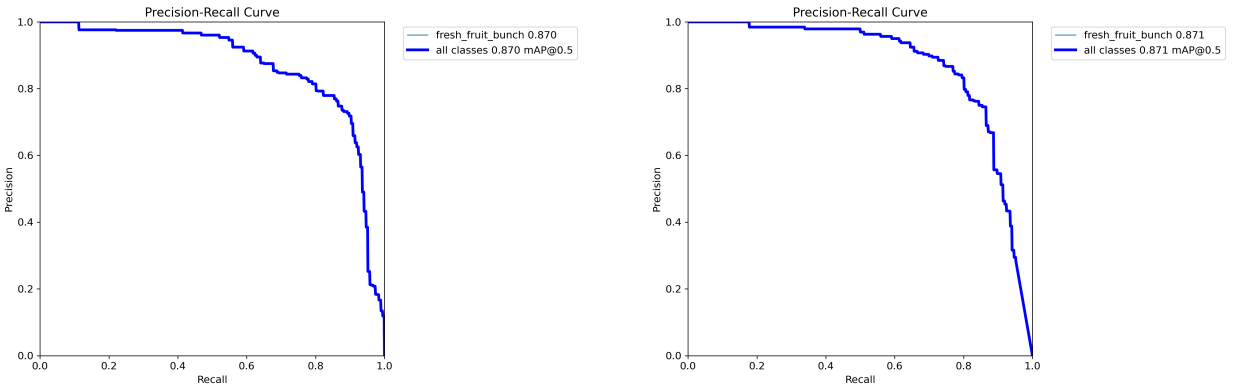


(d) A.4a – Synthetic Depth Only (seed 101)



(e) A.4b – RGB + Synth. Depth (seed 789)

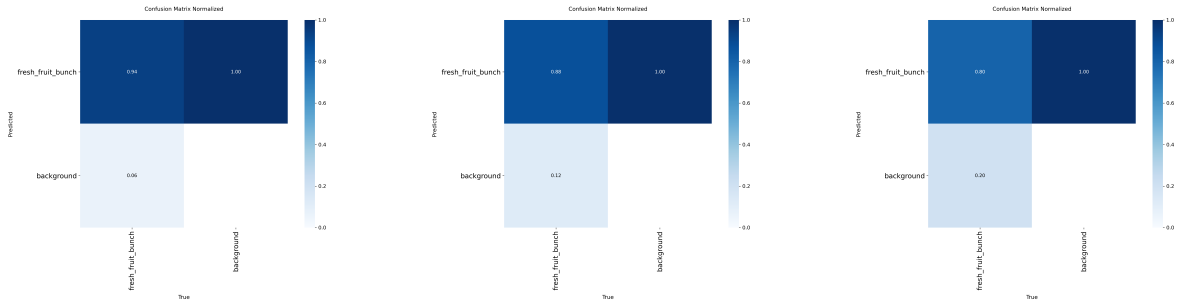
Figure 1: Training curves (loss, mAP50, mAP50–95) for representative seeds of each experiment.



(a) A.1 – RGB Only (seed 42)

(b) A.3 – RGB + Real Depth (seed 789)

Figure 2: Box Precision–Recall curves comparing the RGB baseline (A.1) with the best-performing fused model (A.3).



(a) A.1 – RGB Only

(b) A.3 – RGB + Real Depth

(c) A.4a – Synth. Depth Only

Figure 3: Normalized confusion matrices for selected experiments, illustrating detection accuracy and background confusion rates.

4 Discussion

RGB remains highly competitive. The RGB-only baseline (A.1) achieves 0.8385 mAP50, which is within the margin of error of the best result A.3 (0.8403). The difference of +0.2% is not statistically significant given the standard deviations involved (± 0.025 vs. ± 0.016). This suggests that for this dataset and model capacity, RGB features already capture sufficient discriminative information for FFB detection.

Depth fusion provides marginal gains. Early fusion with real depth (A.3) slightly improves both mAP50 and recall compared to A.1, and notably reduces cross-seed variance (std 0.016 vs. 0.025). The stabilizing effect may be the most practically relevant benefit of depth fusion, even if the absolute accuracy gain is small.

Depth-only detection is substantially weaker. Both A.2 (real depth, 0.7325) and A.4a (synthetic depth, 0.6533) fall well below the RGB baseline. This confirms that depth maps lack the textural and chromatic features critical for distinguishing FFBs from background vegetation. Real depth consistently outperforms synthetic depth by $\sim$8 percentage points, likely because sensor-

captured depth provides metrically accurate distance information, whereas monocular estimation introduces systematic errors.

Synthetic depth is viable when fused with RGB. A.4b (0.8233) performs respectably as the third-ranked configuration and exhibits the lowest seed variance (0.0122). This is practically significant: when no depth sensor is available, running Depth-Anything-V2 on RGB images and fusing the result can still contribute to a more stable—if slightly less accurate—detector.

Late fusion underperforms early fusion. A.5 (0.8084) ranks below both early-fusion variants (A.3, A.4b) and shows the second-highest variance (0.0304). Freezing the backbones limits the model's ability to learn cross-modal interactions, and the concatenated feature space may introduce redundancy without additional attention or gating mechanisms.

5 Conclusion

This study systematically evaluated six input configurations for oil palm FFB detection using YOLOv11n across five random seeds under a controlled experimental protocol. The key findings are:

1. **RGB + real depth early fusion (A.3)** achieves the highest mAP50 (0.8403) and the best recall (0.7743), but the improvement over RGB-only (A.1, 0.8385) is marginal.
2. **Depth information alone is insufficient** for competitive detection. Real depth (0.7325) outperforms synthetic depth (0.6533), but both lag significantly behind RGB.
3. **Multi-modal fusion improves stability**: both A.3 and A.4b exhibit lower cross-seed variance than single-modality baselines.
4. **Late fusion with frozen backbones** is inferior to early fusion, suggesting that joint feature learning across modalities is important.

Deployment recommendations:

- For maximum accuracy with depth hardware: use A.3 (RGB + real depth).
- For simplicity without depth sensor: use A.1 (RGB only).
- For improved stability without depth hardware: use A.4b (RGB + synthetic depth via Depth-Anything-V2).

Future work should investigate attention-based fusion mechanisms, larger model capacities (YOLOv11s/m), and domain-specific fine-tuning of the monocular depth estimator to reduce the gap between synthetic and real depth.