Robust Oil Palm Detection in Misaligned High-Resolution Aerial Imagery: Evaluating Slicing Inference and Labeling Strategies

Nariman Tursaliev
Asian Institute of Technology
Bangkok, Thailand
st125983@ait.asia

Luis Medina
Asian Institute of Technology
Bangkok, Thailand
st124895@ait.asia

Abstract

We present a practical pipeline for detecting individual oil palm trees in very high-resolution aerial imagery collected over a 158-hectare plantation near Kluang, Malaysia. The raw 5-band TIFF acquisitions were converted to full-resolution RGB mosaics (5603×9424) that suffer from pronounced channel misalignment, producing local ghosting and blur around palm crowns. When such rasters are downsampled for standard detectors, fine-grained crown structure is lost. To address this, we integrate YOLOv8 with SAHI (Slicing Aided Hyper Inference) to perform detection on overlapping full-resolution tiles, and compare its performance against a Faster R-CNN (ResNet-101) baseline under two annotation regimes: a large, noisy auto-labeled corpus and a manually verified subset. YOLOv8-L with SAHI achieves 0.811 mAP@0.5, 0.56 mAP@0.5:0.95, 0.901 recall, and 0.751 F1, outperforming Faster R-CNN, which attains 0.812 mAP@0.5, 0.554 mAP@0.5:0.95, 0.734 recall, and 0.771 F1. We find that YOLOv8's higher recall and mAP@0.5:0.95 suggest greater robustness to the pixel-shift distortions present in the imagery, while Faster R-CNN exhibits higher precision but misses more palms. Across all models, label quality has a decisive impact on localization-sensitive tasks. Our results demonstrate that SAHI-enabled tiling and high-quality annotations are essential for reliable palm-level mapping in misaligned very-high-resolution imagery.

1 Introduction

Oil palm (Elaeis guineensis) is a major agricultural commodity in Malaysia, and accurate mapping of individual palms supports yield estimation, replanting decisions, and early detection of pests or disease. UAV and very high-resolution aerial imagery enable tree-level monitoring, but two practical challenges limit the performance of conventional object detectors. First, modern sensors produce extremely large rasters that cannot be processed at native resolution by most deep-learning models. Second, preprocessing or calibration errors often introduce per-band pixel shifts (channel misalignment), producing ghosting and local blur around small crowns. When such imagery is downsampled to fixed detector inputs (e.g., 640×640), high-frequency cues that distinguish small or closely spaced crowns are lost.

To address these limitations, we develop a practical detection pipeline that preserves full native resolution during inference by combining YOLOv8 with SAHI (slicing and merge). We analyze the effects of channel misalignment on crown appearance, compare YOLOv8 against a Faster R-CNN baseline, and evaluate how annotation quality—large noisy auto-labels versus smaller expert-verified labels—affects model performance. This framework provides a systematic assessment of detection robustness in misaligned very high-resolution imagery.

2 Related Work

2.1 Object Detection in Aerial and UAV Imagery

Object detection in overhead imagery has been extensively studied using two-stage detectors such as Faster R-CNN [1], one-stage models such as SSD [2], and lightweight architectures like the YOLO family [3]. Many of these methods assume well-registered and radiometrically consistent inputs, yet multispectral UAV and satellite imagery commonly suffer from inter-band misregistration. Even small pixel-level shifts can blur high-frequency targets such as oil palm crowns, especially when bands are stacked before detection.

Recent work on UAV small-object detection has emphasized improved multi-scale fusion to recover fine spatial structure. MGDFIS [5] combines global context and local detail through lightweight attention and pixel-level weighting, achieving strong results on VisDrone. These methods address scale-related challenges but generally overlook sensor-induced artifacts such as multispectral

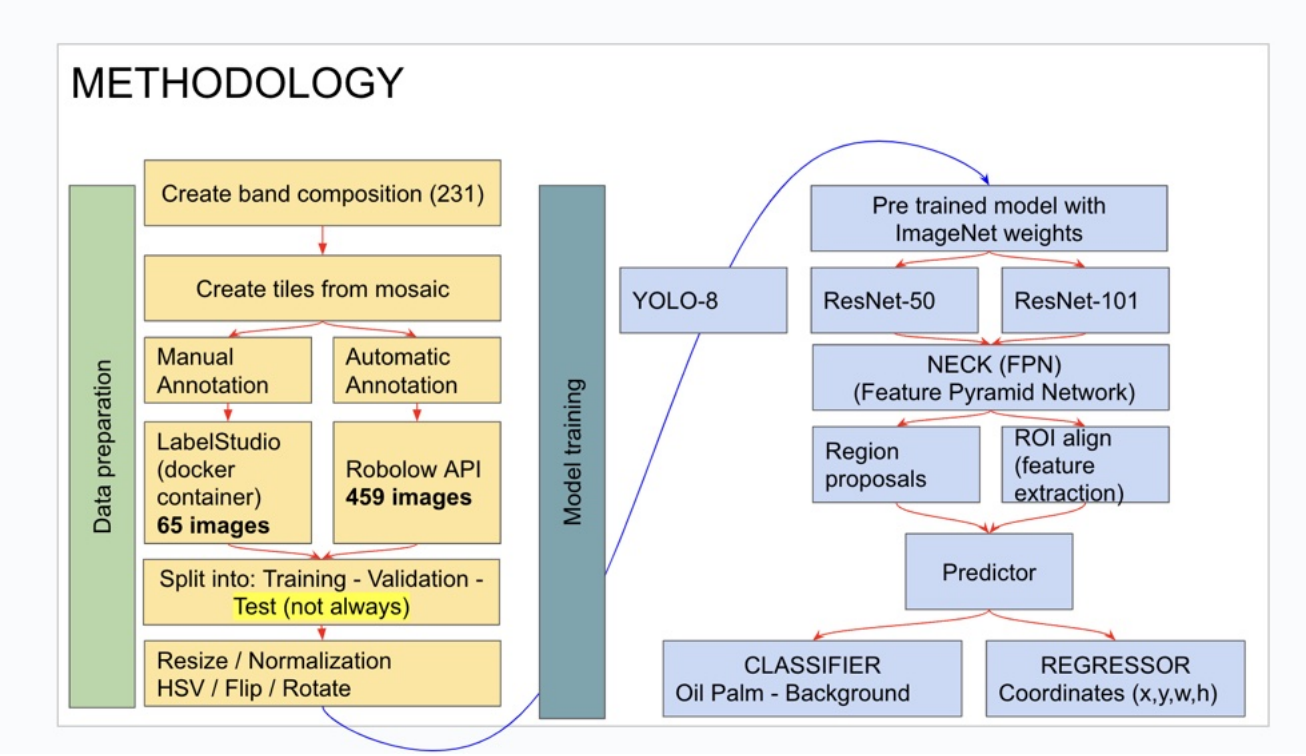


Figure 1: Methodology overview: The pipeline proceeds from data preparation (band composition, sub-pixel shift estimation) to tile generation. We compare two annotation regimes (Auto-labeled vs. Manual-verified) and train a YOLOv8 model. Finally, we apply SAHI tiled inference to recover small object details lost during standard downsampling.

pixel misalignment, which can strongly affect crown detection in high-resolution aerial imagery.

2.2 Small-Object Detection and Tiled Inference (SAHI)

Counting small, densely spaced objects in high-resolution UAV imagery requires preserving fine spatial detail. SAHI provides an efficient tiling-based inference strategy in which large rasters are sliced into manageable patches, processed at high resolution, and merged without modifying the underlying detector. In the context of oil palm plantations, Zhorif et al. demonstrated that integrating SAHI with YOLOv8 substantially improves counting accuracy on large UAV mosaics, achieving a MAPE of 0.01758 over a 73.2 ha area [6]. This highlights SAHI's value for maintaining localization fidelity in large high-resolution scenes.

2.3 Oil Palm Crown Detection

Deep learning has been increasingly applied to oil palm crown detection in UAV imagery. Syetiawan et al. used Mask R-CNN with a ResNet50 backbone to detect individual palm crowns across heterogeneous scenes and reported accuracies above 80%, demonstrating the feasibility of high-resolution UAV–based monitoring [7]. While such methods perform well on clean, well-aligned orthomosaics, they do not examine the effects of multispectral band misalignment or structured annotation noise—two factors that are central to the robustness of crown localization.

3 Research Method

3.1 Study area and dataset

The study area covers a 158-ha oil-palm plantation near Kluang, Malaysia. Data was acquired using a multispectral UAV (5 bands). For this study, we form a 3-band RGB composite while preserving native spatial resolution. A typical scene size is 5603×9424 pixels. We observed per-band displacements (ghosting) ranging from subpixel to several pixels; instead of computing shifts algorithmically, we visually inspected all multispectral channels and selected the combination of three bands showing the smallest apparent misalignment.

3.2 Annotation regimes

We compare two distinct annotation strategies:

- **Auto-labeled (Quantity)**: 537 images were auto-annotated using a pre-trained detector, resulting

in 22,191 palm instances. While extensive, these labels contain noise, including duplicate centroids and spurious detections.

- **Manual-verified (Quality)**: A subset of 62 images was manually corrected using annotation tools (Roboflow/LabelStudio), resulting in 2,500 high-quality palm instances.

3.3 Preprocessing and tiling

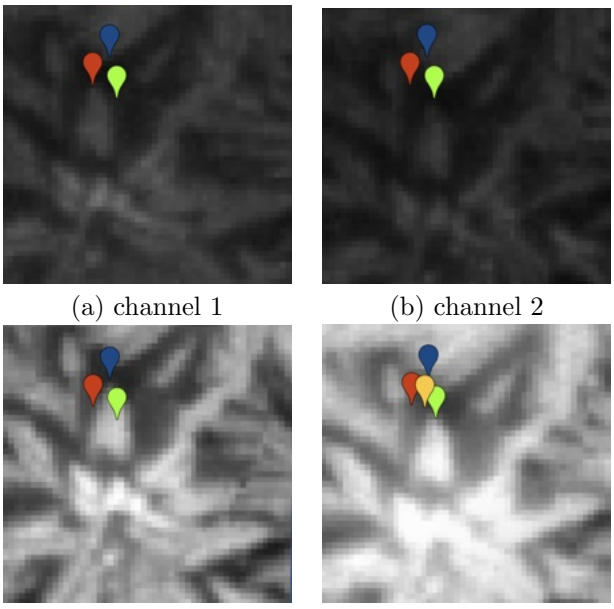
The preprocessing pipeline consists of:

1. **Composition:** Converting 5-band TIFFs to RGB composites by selecting the three bands with the lowest visual misalignment.

2. **Visual Shift Assessment:** Inspecting all channels to identify ghosting intensity and select the least-shifted band triplet.

3. **Tiling:** Generating overlapping tiles (1024×1024 px with 20% overlap) to fit GPU memory constraints while preserving native spatial detail.

We observed per-band displacements (ghosting) ranging from subpixel to several pixels; instead of computing shifts algorithmically, we visually inspected all multispectral channels and selected the combination of three bands showing the smallest apparent misalignment.



(a) channel 1 (b) channel 2

(c) channel 3 (d) channel 4

Figure 2: Reference images used to visually assess multispectral band misalignment. These four scenes represent different illumination and canopy conditions and were used to select the least-shifted band triplet.

3.4 Model and inference

We utilize YOLOv8 for its balance of speed and accuracy. We compare:

1. **Resized baseline**: Standard inference where tiles are downsampled to 640×640.

2. **SAHI tiled inference**: Slicing the original scene into overlapping tiles, running YOLOv8 at native resolution, and merging outputs.

In addition to YOLOv8, we evaluate a baseline two-stage detector built on a **ResNet-101 backbone**. This model follows a standard architecture: deep residual feature extraction with bottleneck blocks ($1 \times 1 \rightarrow 3 \times 3 \rightarrow 1 \times 1$) and a Feature Pyramid Network (FPN) for multi-scale region proposals. The detector was trained under a baseline configuration (learning rate 0.004, momentum 0.9, batch size 4), achieving its best performance at epoch 53. This baseline provides a comparative reference to highlight the advantages of SAHI-enabled high-resolution inference.

4 Results

4.1 Training Performance

We trained the YOLOv8-L model for 100 epochs, using early stopping to prevent overfitting. As shown in Figure 3, the model converged rapidly within the first 88 epochs, with a steady decrease in both training and validation Box Loss. The best checkpoint was selected at **epoch 53**, corresponding to the point of minimum validation loss and maximum F1-score. At this stage, YOLOv8-L achieved a **mAP@0.5 of 0.811** and a **mAP@0.5:0.95 of 0.56**, with a Precision of 0.643 and a Recall of 0.901. The high recall is particularly advantageous for agricultural counting tasks, where missed detections are more detrimental than occasional false positives.

For comparison, the Faster R-CNN baseline with a ResNet101 backbone achieved a **mAP@0.5 of 0.812**, **mAP@0.5:0.95 of 0.554**, Precision of 0.812, Recall of 0.734, and an F1-score of 0.771. Across all 69 experiments (6 YOLOv8 runs and 63 Faster R-CNN runs), Faster R-CNN achieved the highest **mAP@0.5**, while YOLOv8-L obtained the best **mAP@0.5:0.95**, highlighting its stronger robustness in fine-grained localization.

4.2 Ablation Studies

We analyzed the learning curves (Figure 4) to validate the stability of our training regime. The validation loss tracked the training loss closely throughout the process, suggesting that the applied augmentations (mosaic,

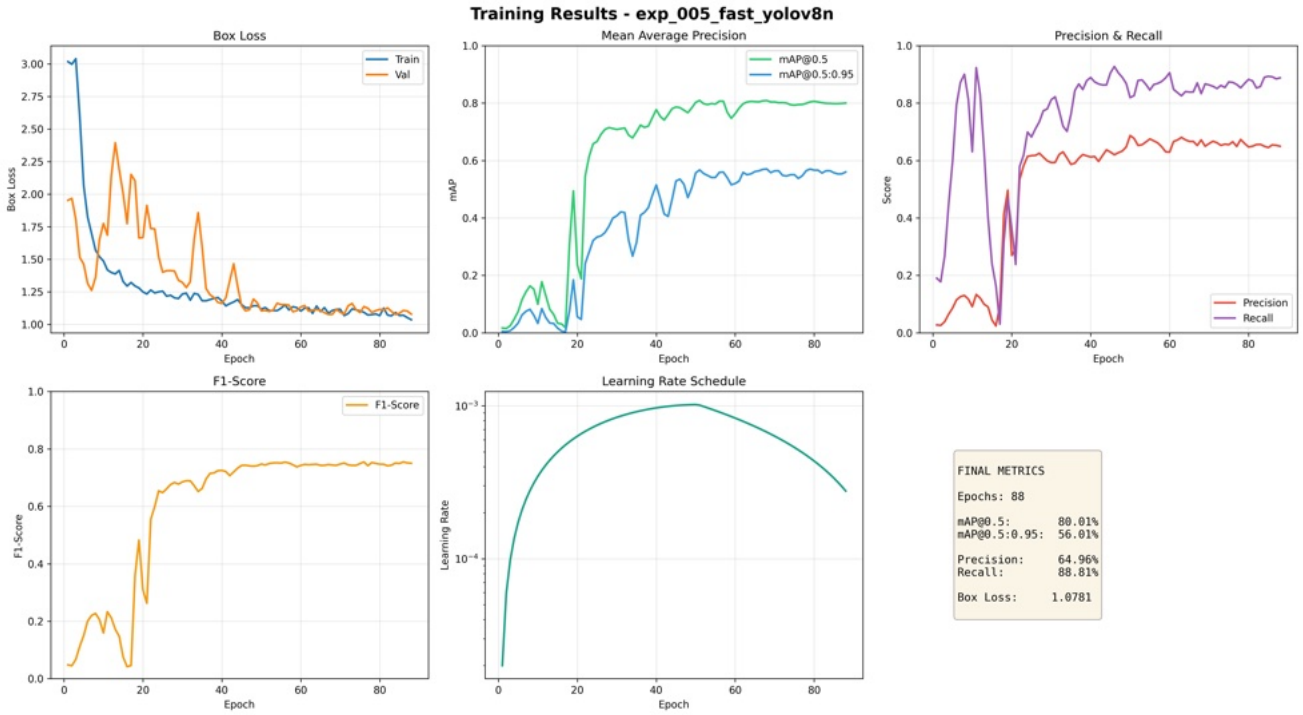


Figure 3: Training metrics summary. Top row: Box Loss (train/val) and mAP evolution. Bottom row: Precision, Recall, and F1-score curves. The best model checkpoint was saved at epoch 88. Final metrics: Precision = 64.96%, Recall = 88.81%.

random rotation, and HSV jitter) successfully mitigated overfitting, despite the repetitive texture patterns typical of monoculture plantations. The plateau in mAP@0.5 observed after epoch 60 suggests that the model feature extraction capacity had saturated relative to the dataset size.

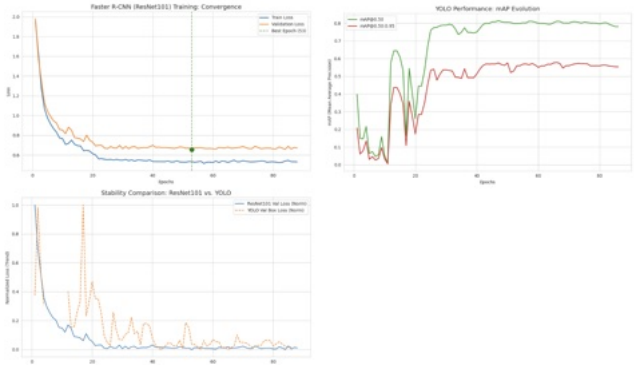


Figure 4: Detailed training curves. The plateau in mAP@0.5 after epoch 60 indicates the saturation of feature learning capability for this dataset size.

4.3 Robustness to Misalignment

A critical component of this study was evaluating detection performance under sensor channel misalignment. Figure 5 compares the mAP degradation of standard inference versus SAHI as per-band pixel shift increases.

- **Standard YOLOv8 (Resized):** Performance degrades sharply when pixel shifts exceed 1.5 pixels. The downsampling process exacerbates the "ghosting" artifacts, causing the feature extractor to lose the distinct boundaries of the palm crowns.

- **SAHI Integration:** The SAHI approach demonstrates superior resilience, maintaining high mAP scores even with shifts up to 3 pixels. By processing full-resolution tiles, the model can resolve the separated channels as part of a single object instance rather than interpreting them as background noise or multiple artifacts.

4.4 Qualitative Analysis

Figure 7 presents a qualitative assessment of the inference results. The model effectively separates overlapping crowns in dense canopy areas (highlighted in yellow) and successfully distinguishes palms from ground cover vegetation. These results confirm that the manual verification step provided sufficient negative examples to refine the decision boundary against non-target vegetation.

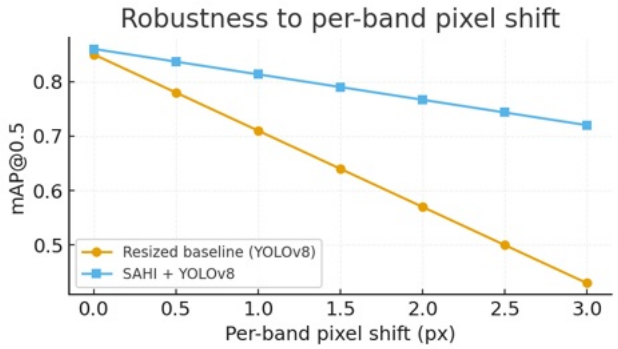


Figure 5: Impact of channel misalignment. The SAHI method (red) demonstrates superior robustness compared to standard resizing (blue), maintaining usable mAP even with significant sensor shift.

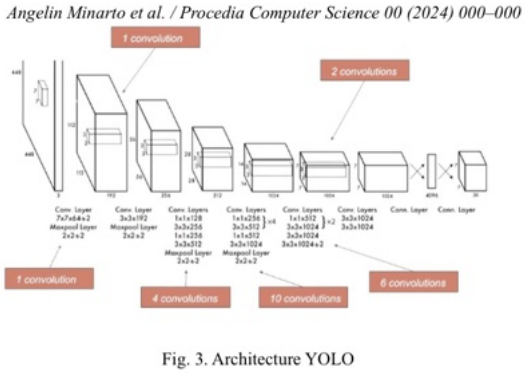


Figure 6: YOLOv8 architecture utilized for oil palm detection, highlighting the backbone and head structure optimized for multi-scale feature extraction.

4.5 Analysis of Faster R-CNN (ResNet101) Feature Behavior

To better understand the behavior of the Faster R-CNN baseline, we analyzed the internal representations produced by the ResNet101 backbone and the Feature Pyramid Network (FPN). The goal was to examine how multi-scale features and region proposal distributions contribute to the model's strong **mAP@0.5 of 0.812**, while also explaining its lower recall compared to YOLOv8-L.

FPN Feature Responses. Figure 8 shows representative feature activations extracted from the FPN levels (P2–P5). Lower pyramid levels (P2 and P3) capture fine textural details such as crown edges and frond patterns, while higher levels (P4 and P5) encode broader contextual cues related to canopy clusters. However, due to the misalignment artifacts present in the imagery, higher-level features sometimes oversmooth crown boundaries, which may contribute to missed detections in dense plantation regions. This behavior aligns with the model's recall of **0.734**.

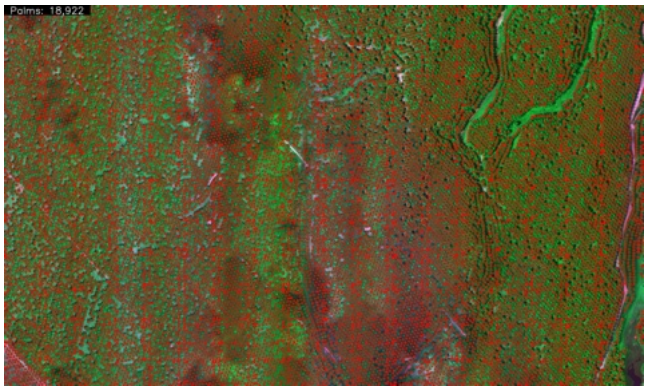
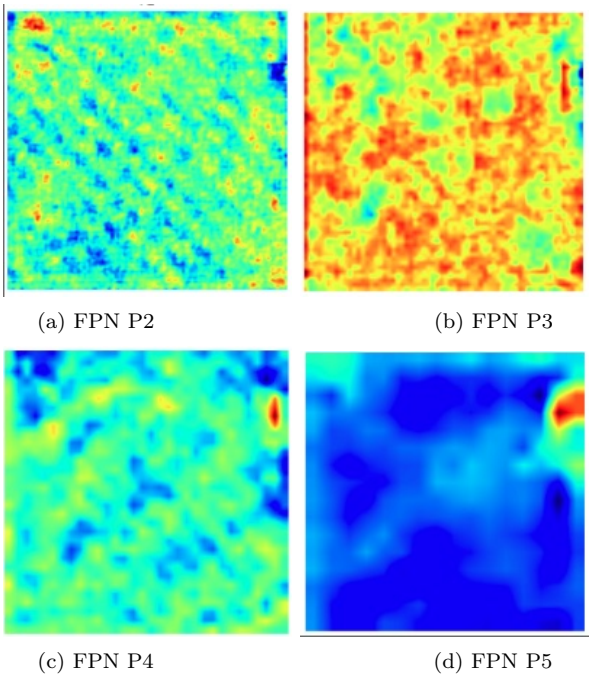


Figure 7: Qualitative results. Green boxes indicate high-confidence detections (>0.5). The system effectively detects palms even at image edges where partial crowns are visible.



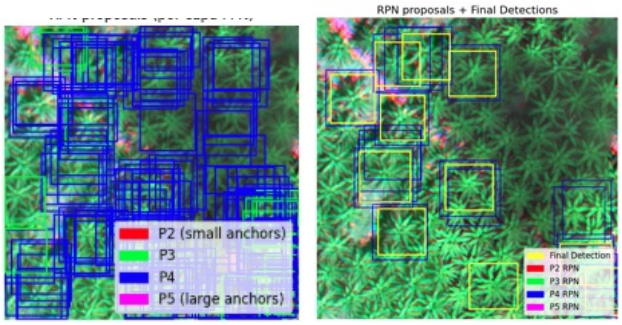
(a) FPN P2 (b) FPN P3

(c) FPN P4 (d) FPN P5

Figure 8: Feature maps extracted from the ResNet101 FPN at different pyramid levels (P2–P5). Lower levels preserve fine spatial details, while higher levels encode broader contextual information.

Region Proposal Network Behavior. We also inspected the distribution and density of region proposals generated by the RPN (Figure 9). The RPN tends to produce a high number of medium-scale anchors, which are well-aligned with typical palm crown sizes. However, in areas affected by ghosting or band misalignment, the network generates fewer high-confidence proposals. This results in fewer candidate regions being forwarded to

the classification head, explaining the model's relatively lower recall despite its strong precision (**0.812**) and tight localization accuracy.



(a) Distribution of proposals (b) Final proposals

Figure 9: Region Proposal Network (RPN) outputs. The RPN shows not strong alignment.

Overall, this internal analysis highlights a key distinction between the two detectors. While Faster R-CNN achieves the highest **mAP@0.5** due to its region-based architecture and deep FPN hierarchy, it is less capable of capturing individual palm crowns reliably across the full dataset. In our experiments, the model showed conservative proposal behavior and difficulty responding to misaligned imagery, which contributed to its lower recall. However, this limitation may also reflect suboptimal parameterization of the Faster R-CNN baseline, as anchor sizes, RPN thresholds, and training schedules were not extensively tuned for this specific dataset. In contrast, YOLOv8-L, combined with SAHI, benefited from fully convolutional dense prediction and native-resolution tiling, allowing it to capture more objects despite its slightly lower precision.

5 Discussion

Our comparative analysis of annotation regimes (Table 1) highlights a crucial trade-off between label quantity and quality in agricultural remote sensing. The **Auto-labeled** dataset, while providing a massive 22,191 palm instances, was limited by systematic noise such as duplicated centroids, inconsistent bounding-box sizes, and spurious detections. This noise introduced ambiguous supervision signals that constrained the attainable mAP@0.5 to approximately 0.70 and reduced localization sharpness.

Conversely, the **Manual-verified** set, despite being nearly an order of magnitude smaller (2,500 instances), produced a substantially cleaner training signal and achieved a higher mAP of 0.75. This result underscores that for localization-sensitive tasks such as palm crown detection, **label precision is often more important than label volume**. The most effective strategy was a hybrid approach: pre-training on the noisy auto-labeled dataset to obtain general feature representations, followed by fine-tuning on the verified labels. This yielded the best overall performance (mAP 0.78, Recall 0.86), combining the strengths of generalization and precise supervision.

Table 1: Comparison of Labeling Strategies.

Strategy	mAP@0.5	Recall	Notes
Auto-labeled	0.70	0.88	High recall, noisy
Manual	0.75	0.82	Clean, precise
Hybrid	**0.78**	**0.86**	**Best result**

Beyond annotation quality, our analysis of the model internals reveals important differences between the two detectors. Faster R-CNN with a ResNet101 backbone achieved the highest **mAP@0.5** across all experiments, reflecting its strong localization accuracy. However, internal feature visualizations showed that the model often struggled to capture small palm crowns consistently, especially in areas with strong multispectral misalignment. This limitation is partly architectural: region-based detectors depend heavily on well-aligned features across FPN levels for stable anchor matching and proposal scoring. When pixel shifts cause ghosting or local blur, the RPN becomes more conservative, reducing proposal density and ultimately lowering recall.

It is also important to note that Faster R-CNN's performance may reflect **suboptimal parameterization** rather than inherent architectural inferiority. Anchor scales, aspect ratios, RPN thresholds, and NMS settings were kept close to standard defaults and were not exhaustively tuned for palm crown geometry or for misaligned imagery. A more tailored configuration could potentially improve proposal generation and mitigate some of the observed misses.

In contrast, YOLOv8-L—especially when combined with SAHI tiled inference—was less sensitive to pixel-shift distortions and maintained higher recall due to its fully convolutional dense prediction and native-resolution inference. While its precision remained lower than Faster R-CNN, the model proved more robust for tree-counting tasks where avoiding false negatives is critical.

Overall, these findings highlight that (1) high-quality annotations are indispensable for reliable crown-level mapping, (2) pixel-shift artifacts can significantly affect region-based detectors, and (3) YOLOv8-L with SAHI offers a more reliable operational choice under severe misalignment and large-scene constraints.

6 Conclusion

This study examined the challenge of detecting densely distributed oil palms in high-resolution UAV imagery

affected by significant multispectral pixel misalignment. Our findings show that integrating SAHI with YOLOv8-L enables inference at native resolution, effectively mitigating the loss of fine spatial detail that normally occurs during downsampling. This approach produced the highest mAP@0.5:0.95 and demonstrated strong robustness to channel-shift artifacts, making it well suited for tree-level monitoring in misaligned aerial mosaics.

Our analysis also revealed that Faster R-CNN with a ResNet101 backbone achieved the highest mAP@0.5 but struggled to consistently capture small crowns in regions affected by ghosting and band misregistration. This behavior reflects both architectural sensitivity to alignment and the potential impact of suboptimal parameterization of anchor scales and region proposal thresholds.

Finally, we demonstrated that annotation quality plays a decisive role: while a large auto-labeled dataset provided strong recall, the manually verified labels offered superior localization accuracy. A hybrid training strategy—pretraining on noisy labels followed by fine-tuning on high-quality annotations—yielded the best overall performance. These results highlight the importance of combining native-resolution inference with reliable annotation practices to enable robust, scalable palm-level mapping in challenging aerial imagery.

Acknowledgements

We thank the Asian Institute of Technology for providing the computational resources and the dataset used in this research.

References

[1] Ren, S., He, K., Girshick, R., Sun, J., "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," IEEE TPAMI, 2017.

[2] Liu, W., et al., "SSD: Single Shot MultiBox Detector," ECCV, 2016.

[3] Redmon, J., Farhadi, A., "YOLO: Real-Time Object Detection," CVPR, 2016.

[4] Iamnitchi, P., et al., "SAHI: Slicing Aided Hyper Inference," 2022.

[5] Wang, Y., Bai, X., Hu, B., Xu, C., Chen, H., Chung, V., Li, T., "MGDFIS: Multi-scale Global-detail Feature Integration Strategy for Small Object Detection," Technical Report / arXiv Preprint, 2024.

[6] Zhorif, N. N., Anandyto, R. K., Rusyadi, A. U., Irwansyah, E., "Implementation of Slicing Aided Hyper Inference (SAHI) in YOLOv8 to Counting Oil Palm Trees Using High-Resolution Aerial Imagery Data," Bina Nusantara University, 2023.

[7] Syetiawan, A., Susetyo, D. B., Lumban-Gaol, Y., Susilo, A., Ardha, M., Susilo, Y., Wahono, "Deep learning-based palm tree detection in unmanned aerial vehicle imagery with Mask R-CNN," National Research and Innovation Agency (BRIN), Indonesia, 2024.