Detection and Geographic Localization of Natural Objects in the Wild: A Case Study on Palms

Kangning Cui[1,2] , **Rongkun Zhu**[3] , **Manqi Wang**[2] , **Wei Tang**[1] , **Gregory D Larsen**[2] , **Victor P Pauca**[2] , **Sarra Alqahtani**[2] , **Fan Yang**[2] , **David Segurado**[2] , **David A. Lutz**[4] , **Jean-Michel Morel**[1] and **Miles R Silman**[2]

[1]Department of Mathematics, City University of Hong Kong
[2]Department of Computer Science & Department of Biology, Wake Forest University
[3]Department of Computer Science, Xidian University
[4]School of Arts & Sciences, Colby-Sawyer College
{cuij, zhur, paucavp, sarra-alqahtani, yangfan, silmanmr}@wfu.edu

Abstract

Palms are ecologically and economically indicators of tropical forest health, biodiversity, and human impact that support local economies and global forest product supply chains. While palm detection in plantations is well-studied, efforts to map naturally occurring palms in dense forests remain limited by overlapping crowns, uneven shading, and heterogeneous landscapes. We develop PRISM (Processing, Inference, Segmentation, and Mapping), a flexible pipeline for detecting and localizing palms in dense tropical forests using large orthomosaic images. Orthomosaics are created from thousands of aerial images and spanning several to hundreds of gigabytes. Our contributions are threefold. First, we construct a large UAV-derived orthomosaic dataset collected across 21 ecologically diverse sites in western Ecuador, annotated with 8,830 bounding boxes and 5,026 palm center points. Second, we evaluate multiple state-of-the-art object detectors based on efficiency and performance, integrating zero-shot SAM 2 as the segmentation backbone, and refining the results for precise geographic mapping. Third, we apply calibration methods to align confidence scores with IoU and explore saliency maps for feature explainability. Though optimized for palms, PRISM is adaptable for identifying other natural objects, such as eastern white pines. Future work will explore transfer learning for lower-resolution datasets (0.5–1m). Data and code can be found at github.com/Zippppo/PRISM.

1 Introduction

Palms (family Arecaceae) are vital to tropical ecosystems, serving as essential resources for pollinators and frugivores and influencing the evolution of dependent fauna [Eiserhardt *et al.*, 2011; Zambrana *et al.*, 2007]. Beyond their ecological roles, palms are deeply integrated in the livelihoods of rural and indigenous communities, providing food, construction materials, fuel, and medicine while supporting sustain-

Figure 1: Palm Distribution Comparison. The first three images from previous studies [2021; 2022] feature evenly spaced palms or clear backgrounds, while the last represents our case with natural spacing, occlusions, and complex backgrounds in tropical forests.

able non-timber forest product markets. Particularly in regions like the Amazon, palms support subsistence practices and enhance resilience to socio-economic and environmental changes, reflecting the intricate links between biodiversity, human well-being, and sustainability [Pitman *et al.*, 2014; Malhi *et al.*, 2014; van der Hoek *et al.*, 2019; Terborgh, 1986]. While their distinctive star-shaped crowns make palms well-suited for automated mapping [Sutherland *et al.*, 2013; Wagner *et al.*, 2020], the absence of scalable, precise tools to monitor these species across vast forests limits timely conservation efforts and support for dependent communities.

Orthomosaic images are a valuable tool in remote sensing, with applications across various fields such as construction site monitoring, agricultural planning, environmental impact assessments, property development, and surveying, where accurate spatial data is essential [Kucharczyk and Hugenholtz, 2021]. These images are created by stitching together hundreds or thousands of geolocated aerial frames into spatially coherent maps, enabling consistent analysis across landscapes. By providing a top-down view of large geographical areas, orthomosaics allow for precise measurement and detailed analysis of land features. Unlike standard aerial imagery, which contains perspective distortions, orthomosaics offer an in-scale representation of the ground. Low-flying drones enable efficient data collection without the need for extensive ground surveying [Olson and Anderson, 2021].

However, orthomosaics are not error-free and can be challenging to process. They often contain noise, artifacts, and stitching errors due to limited camera perspectives, variable lighting, wind-induced movement, and dynamic cloud cover

during image acquisition [Zhang *et al.*, 2023b]. Additionally, resolution loss is a frequent byproduct of the stitching process. Typically stored in GeoTIFF format, orthomosaics can be exceptionally large, particularly in biodiversity conservation, forest health, and sustainable management applications, where a single file may range from 1 to 200 GB. As a result, image analysis must be performed in small patches that fit within the memory constraints of processing hardware.

Object-based detection techniques, widely used in remote sensing, are well-established in computer vision. However, detecting and localizing naturally occurring objects in orthomosaics is inherently challenging [Tagle Casapia *et al.*, 2019; Cui *et al.*, 2024; Qin *et al.*, 2021; Jintasuttisak *et al.*, 2022]. Unlike structured plantations, where trees are arranged in grid patterns with ample spacing (see Figure 1), palms in tropical forests grow irregularly, often obscured by dense vegetation that distorts key diagnostic features. The distribution of palm species is highly imbalanced, with only a few dominant species present in sufficient abundance to train reliable detection models [Li *et al.*, 2016; Freudenberg *et al.*, 2019]. UAV-based imagery also introduces substantial lighting variability, as shadows from thick canopies and shifting sun angles create significant fluctuations in brightness and contrast across detection targets, further complicating model training and evaluation [Gibril *et al.*, 2021]. These challenges are amplified by the lack of high-quality labeled datasets, whose creation often requires extensive fieldwork and expert annotation [Muscarella *et al.*, 2020; Hidalgo Pizango *et al.*, 2022].

Lastly, manipulation of geographical information data requires domain expertise that is not readily available to computational scientists. A flexible pipeline that can process spatially referenced survey imagery and provide precise geographic target coordinates would support valuable downstream applications, leveraging state-of-the-art machine learning methods. These applications include monitoring palm abundance and distribution over time and across vast areas, and developing adaptive and modular survey methods. Additionally, balancing accuracy with interpretability and calibration is essential for trustworthy ecological decision-making [Selvaraju *et al.*, 2017; Kuzucu *et al.*, 2025]. This study therefore introduces a novel end-to-end pipeline that addresses a major need in environmental monitoring, is computationally efficient and reliable, and can support variable-resource processing in field applications.

Our work presents the following main contributions:

1. We **construct** and provide the **PA**lm **L**ocalization in **M**ulti-**S**cale **(PALMS) dataset**. PALMS contains data from extensive fieldwork across 21 sites in western Ecuador, spanning a rainfall gradient from the Choco's wettest forests to the dry edge of the Sechura desert, which captures corresponding gradients in palm species composition and canopy characteristics. For training, we annotated 1,500 image patches from 2 sites with 8,830 bounding boxes. For validation, we manually marked 5,026 palm crown centers from 4 reserves to direct compare detected georeferenced centers with ground truth.

2. We **develop the PRISM pipeline**, a unified and modular pipeline for natural object detection, segmentation, and

counting. PRISM integrates object detection with zero-shot segmentation to generate georeferenced palm coordinates. The modular design allows easy interchange of detection/segmentation models, while calibration analysis and saliency maps enhance trustworthiness and interpretability for ecological applications.

3. We apply PRISM to the PALMS dataset and **validate generalization** across four reserves with diverse environmental characteristics and species compositions. PRISM accurately localizes palm centers with strong ground-truth alignment. Benchmarking results show high inference efficiency across diverse hardware settings, supporting large-scale ecological monitoring.

2 Related Work

2.1 Palm Detection and Localization

Advancements in UAV technology, image stitching, and machine learning have driven significant progress in palm detection, segmentation, and localization from orthomosaic imagery. However, most studies have focused on commercially valuable species, such as oil and date palms, given their economic importance [Li *et al.*, 2016; Gibril *et al.*, 2021; Zheng *et al.*, 2021; Jintasuttisak *et al.*, 2022; Putra and Wijayanto, 2023]. For instance, Li *et al.* used CNNs with a sliding window for oil palm counting in Malaysia, while Gibril *et al.* developed a U-Net variant for enhanced date palm segmentation in UAE. Zheng *et al.* proposed a Faster R-CNN variant with refined feature extraction and a hybrid class-balanced loss to monitor individual oil palm growth. More recently, YOLO-based approaches have been adopted: Jintasuttisak *et al.* applied YOLOv5 for detecting date palms from UAV imagery over UAE farmlands, while Putra and Wijayanto employed YOLOv3 to detect and count oil palm trees for sustainable agricultural monitoring in Indonesia.

In contrast, the detection and localization of naturally occurring palms in tropical forests is largely underexplored. Tagle Casapia *et al.* pioneered palm crown identification using random forest, showing machine learning's potential for individual palm counting. Ferreira *et al.* applied a fully convolutional neural network with morphological operations to refine palm species segmentation. Wagner *et al.* used U-Nets and high-resolution multispectral imagery from the GeoEye satellite to map canopy palms over a large region in Amazon.

2.2 Object Detection and Zero-Shot Segmentation

Object detection, a core computer vision task, identifies and localizes objects via bounding boxes [Zou *et al.*, 2023] and underpins advanced applications such as image segmentation and object tracking [Wang *et al.*, 2022; Ma *et al.*, 2024; Li *et al.*, 2024]. The field is dominated by methods using You Only Look Once (YOLO) and Detection Transformer (DETR).

The YOLO family frames detection as a regression task balancing speed and accuracy, with overlapping predictions typically resolved via non-maximum suppression (NMS). YOLOv8 [2023] improves feature fusion and adopts an adopts anchor-free detection head optimized for accuracy and speed. YOLOv9 [2024b] introduces programmable gradient information and the generalized efficient layer aggregation
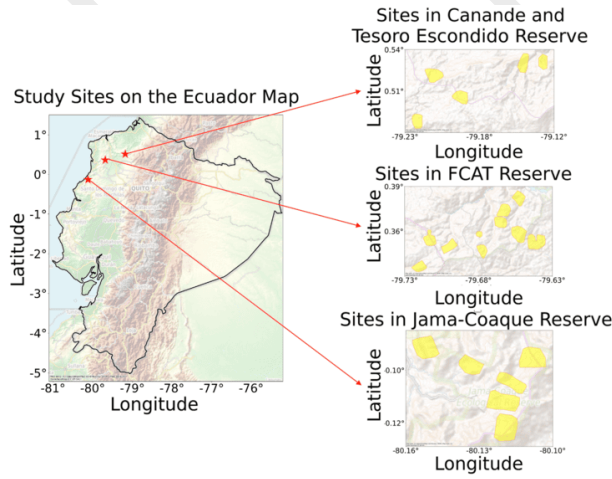
Figure 2: Geographic Locations of Study Sites. The left panel shows a map of Ecuador with red stars marking the study regions. The right panels zoom in on 21 study areas within four ecological sites.

network to address information loss. YOLOv10 [2024a] removes NMS through consistent dual assignments in training and one-to-one inference matching, coupled with a refined CSPNet backbone and a lightweight head to reduce computational cost. YOLO11 [2024] further enhances performance with a refined CSP bottleneck, hybrid attention, and adaptive anchors with extended IoU loss.

DETR [2020] directly predicts object sets using learned queries, bypassing the need for post-processing such as NMS. DINO [2022] enhances DETR with contrastive denoising and hybrid query initialization, while DDQ-DETR [2023c] introduces dense query assignment for improved one-to-one inference matching. RT-DETR [2024] optimizes DETR for real-time use via a hybrid encoder and multi-scale feature fusion.

Segment Anything Models (SAMs) are advanced segmentation models capable of segmenting any object in images using prompts such as points, boxes, or text [2023; 2024; 2023a]. Trained on the SA-1B (1 billion masks, 11 million images), SAM enables zero-shot inference and often surpasses fine-tuned methods in accuracy and efficiency [2023]. Its architecture features a ViT for image encoding, a prompt encoder to process input prompts, and a mask decoder that fuses features from both to generate segmentation masks. SAM 2 [2024], trained on the SA-V dataset (50.9k videos, 642.6k masks), enhances video segmentation and object tracking by refining multi-scale feature extraction. Mobile SAM [2023a] optimizes SAM for mobile use by simplifying the image encoder and using decoupled distillation, enhancing speed without compromising segmentation quality.

3 PALMS Dataset

In this section, we introduce the PALMS (**PA**lm **L**ocalization in **M**ulti-**S**cale) dataset. The following paragraphs detail the study sites, data collection, and manual labeling process.

Study Sites. Data in this study (See Figure 2) come from western Ecuador's Choco tropical forest, including *Fundación para la Conservación de los Andes Tropicales Re-*

serve and adjacent Reserva Ecológica Mache-Chindul park (FCAT; 00°23'28" N, 79°41'05" W), *Jama-Coaque Ecological Reserve* (00°06'57" S, 80°07'29" W), *Canande Reserve* (0°31'34" N 79°12'47" W), and *Tesoro Escondido Reserve* (0°33'16" N 79°10'31" W). FCAT is a high diversity humid tropical forest at elevation ∼500m, receiving ∼3000 mm yr^{-1} precipitation with persistent fog during drier period. Jama-Coaque ranges from the boundary of the tropical moist deciduous/tropical moist evergreen forest at the lower elevations (∼1000 mm precipitation yr^{-1}, ∼250 m asl) to fog-inundated wet evergreen forests above 580m to 800m. Canande (350–500 m elevation) and Tesoro Escondido (∼200 m elevation) are lowland everwet Choco forests, both receiving 4000–5000 mm yr^{-1} precipitation with no month experiencing drought stress or precipitation below 100 mm. These forests host several palm species with exposed canopy crowns, including the economically important *Iriartea deltoidea*, *Socratea exorrhiza*, and *Oenocarpus bataua*, with lesser amounts of *Attalea colenda* and *Astrocaryum standleyanum*, and species composition varying across study sites [Browne and Karubian, 2016; Lueder *et al.*, 2022].

Data Collection. We collected UAV imagery in two stages, capturing 8,845 photos across 21 areas spanning 1,995 hectares, with a ground sampling distance under 6 cm. In June 2022, the first stage covered 95 hectares and produced 387 photos, while the second stage in February 2023 surveyed 1,900 hectares and captured 8,458 photos. Both missions used a DJI Phantom 4 RTK drone equipped with a 1" CMOS sensor and GS RTK for mission planning. The first mission flew at 90 meters above ground level with 70% sidelap and 80% frontlap, while the second operated at 150 meters. The collected images were processed for subsequent analysis. To enable landscape-level forest analysis, we created orthomosaics and conducted post-processing steps using Agisoft Metashape 2.0, including noise removal, edge trimming, and the generation of digital surface and terrain models.

Manual Labels. To enable fine-tuned palm detection and zero-shot segmentation, we curated a dataset of 1,500 images (800×800 pixels) from two FCAT reserve sites, capturing varied quality and palm density typical of natural forests. Manual annotation of palm crowns and isolated leaves faced challenges from vegetation overlap and orthomosaic distortions. Three trained experts initially labeled bounding boxes, followed by iterative refinement using a YOLOv8 detector trained on these annotations to identify and correct labeling errors or omissions. For landscape-scale validation, a trained expert annotated 5,026 georeferenced palm centers on four orthomosaics using ArcGIS Pro 3.3.1, with predicted coordinates from PRISM to refine annotations. This hybrid human-model workflow addressed the inherent complexity of labeling in natural forest environments, ensuring robust annotations for both detection and geospatial validation.

4 PRISM Pipeline

PRISM (Processing, Inference, Segmentation, and Mapping) is an end-to-end pipeline that process orthomosaic images to generate georeferenced palm coordinates alongside bounding boxes and segmentation masks for visualization. Fine-tuned
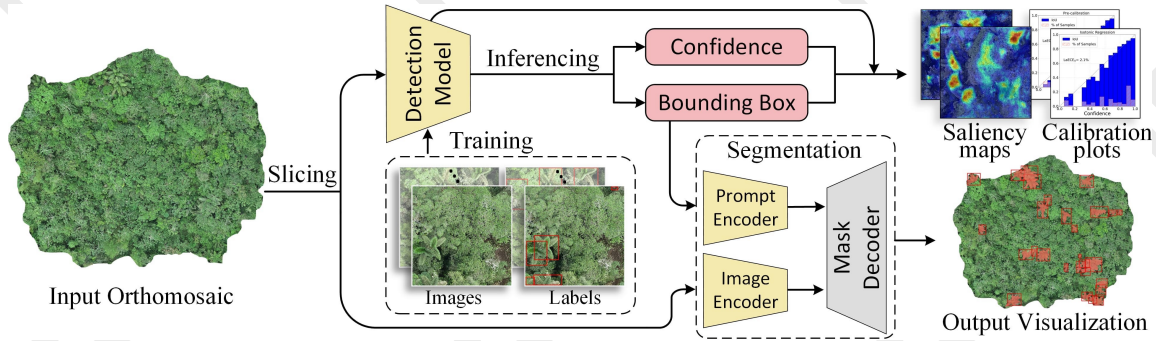
Figure 3: PRISM Pipeline Overview. The detection model, trained on the PALMS dataset, processes orthomosaic slices to generate confidence scores and bounding boxes. These bounding boxes are refined and serve as prompts, along with the sliced input images, for zero-shot segmentation. The bounding boxes and confidence scores are further utilized for saliency map generation and calibration analysis.

on the PALMS dataset, PRISM addresses challenges specific to dense rainforest environments, such as irregular palm distributions and overlapping crowns. By integrating fine-tuned detection models with zero-shot segmentation, PRISM ensures adaptability to a wide range of environmental tasks. Figure 3 illustrates the PRISM pipeline, with the details of each model component discussed in the following sections.

Detection Model. We selected YOLOv10 [2024a] for its speed and performance. YOLOv10 introduces a consistent dual assignment strategy for NMS-free training, which combines one-to-many and one-to-one label assignments for enriched supervision while eliminating post-processing NMS. This approach uses a unified matching metric:

$$m(\alpha, \beta) = s \cdot p^\alpha \cdot \text{IoU}(\hat{b}, b)^\beta,$$

where p is the classification score, $\hat{b}$ and b are the predicted and ground truth (GT) bounding boxes, and s is the spatial prior, with α and β balancing classification and IoU. Architecturally, YOLOv10 adopts an efficiency-accuracy-oriented design, integrating lightweight classification heads, spatial-channel decoupled downsampling, rank-guided block design, and advanced features such as large-kernel convolutions and partial self-attention, achieving robust performance with faster inference and fewer parameters. Trained on PALMS dataset, the model processes orthomosaic patches to output bounding boxes and confidence scores for further processing.

Segmentation Model. We chose SAM 2 [2024] for zero-shot segmentation due to its superior segmentation quality and improved speed compared to its predecessor. SAM 2 uses a hierarchical image encoder for multi-scale feature extraction and an optimized architecture that reduces computational overhead while maintaining high precision. Its efficient prompt encoding and memory attention mechanism enable rapid mask refinement during inference. During inference, bounding boxes from the detection model undergo NMS to remove duplicates, as patches are cropped from the orthomosaic with a stride. The remaining bounding boxes and their surrounding image regions serve as input prompts for SAM 2 to generate segmentation masks. Final outputs include visualized bounding boxes and masks, alongside georeferenced coordinates derived from NMS-cleaned bounding box centers, which are subsequently used to quantify counting accuracy.

Calibration. Calibration ensures model trustworthiness by aligning predicted confidences with empirical IoU, critical for ecological monitoring where overconfident errors can mislead conservation decisions. We evaluate four calibration methods to improve palm detection reliability [2025]: (1) Linear regression (LR) fits a linear logit-probability mapping; (2) Isotonic regression (IR) fits a monotonic function: $\min_{\hat{y}_1, \hat{y}_2, ..., \hat{y}_n} \sum_{i=1}^n (y_i - \hat{y}_i)^2$, subject to $\hat{y}_i \leq \hat{y}_j, \forall i < j$; (3) Temperature scaling (TS) divides logits by a learned temperature T before sigmoid; (4) Platt scaling (PS) applies logistic regression to map logits to calibrated probabilities.

Interpretability via Saliency Maps. Grad-CAM [2017] enhances interpretability by highlighting regions most influential to palm predictions. For a class score y, it computes gradients of y with respect to feature maps A^k, yielding weights $\alpha_k = \frac{1}{Z} \sum_{i,j} \frac{\partial y}{\partial A_{ij}^k}$, where Z is the number of spatial positions. The heatmap: $L_{\text{map}} = \text{ReLU} \left(\sum_k \alpha_k A^k \right)$ is then generated to highlight regions influencing predictions, such as palm crown or leaf patterns. These visualizations help ecologists assess whether predictions rely on meaningful cues like crown shapes rather than irrelevant artifacts.

5 Experimental Results

This section evaluates the proposed pipeline across various tasks and study sites. We numerically compare the detection performance and computational efficiency of different models, and visually compare the results of different zero-shot segmentation models. Calibration metrics are examined to measure model trustworthiness by analyzing performance before and after calibration. Saliency maps are analyzed to trace attention shifts during inference. Next, we apply PRISM to orthomosaics with distribution shift to examine the counting performance in practical scenarios. Finally, real-time simulations are conducted to assess the detection and segmentation speed, as well as the computational demands across devices.

5.1 Experimental Setting

The detection models were trained on PALMS dataset that split into 80% for training, 10% for validation, and 10% for testing. Data splitting was randomly performed five times to ensure robust evaluation, with training and testing data
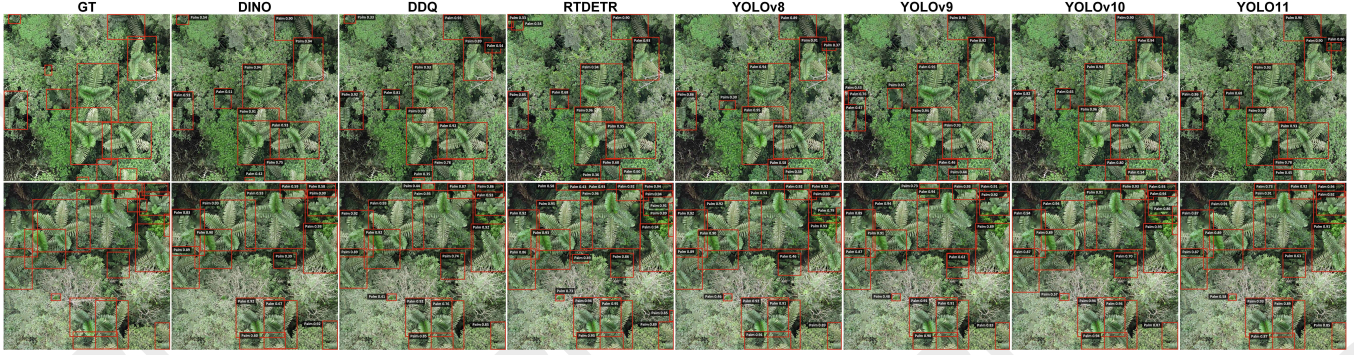
Figure 4: Comparison of Palm Detection Performance. Several models are compared in detecting palms, including small, occluded, and boundary-adjacent cases. All models perform well on large palms, even in occluded scenarios. DETR-based models excel at detecting small palms, while YOLO-based models perform better for partially visible palms on boundaries.

Model	GFLOPS ↓	Params (M) ↓	FPS ↑	Precision ↑	Recall ↑	AP_{50} ↑	AP_{75} ↑	mAP ↑
DINO	1920.3	218.2	18.98 ± 0.95	0.7629 ± 0.0177	0.8494 ± 0.0071	0.8169 ± 0.0166	0.5455 ± 0.0150	0.5102 ± 0.0101
DDQ	1232.6	218.6	19.18 ± 0.96	0.7825 ± 0.0124	$\mathbf{0.8566 \pm 0.0123}$	0.8541 ± 0.0129	0.6354 ± 0.0137	0.5736 ± 0.0130
RT-DETR	222.5	65.5	151.49 ± 0.70	$\mathbf{0.8869 \pm 0.0230}$	0.7598 ± 0.0310	0.8416 ± 0.0181	0.6198 ± 0.0181	0.5769 ± 0.0145
YOLOv8	226.7	61.6	174.92 ± 0.86	0.8729 ± 0.0165	0.7997 ± 0.0203	0.8667 ± 0.0141	0.6777 ± 0.0137	0.6148 ± 0.0128
YOLOv9	$\mathbf{169.5}$	53.2	114.96 ± 0.30	0.8763 ± 0.0176	0.7976 ± 0.0209	$\mathbf{0.8741 \pm 0.0109}$	0.6762 ± 0.0146	0.6162 ± 0.0122
YOLOv10	169.8	$\mathbf{31.6}$	$\mathbf{177.04 \pm 1.14}$	0.8716 ± 0.0121	0.7968 ± 0.0089	0.8626 ± 0.0129	$\mathbf{0.6794 \pm 0.0112}$	$\mathbf{0.6173 \pm 0.0090}$
YOLO11	194.4	56.8	170.40 ± 0.95	0.8721 ± 0.0095	0.7896 ± 0.0127	0.8684 ± 0.0108	0.6677 ± 0.0180	0.6115 ± 0.0109

Table 1: Comparative Analysis of Computational Efficiency and Detection Accuracy Among Models for PRISM. The table presents average performance metrics with standard deviations obtained from five random sampling experiments. FPS values were measured on an NVIDIA RTX 4090 24 GB GPU. Bold values indicate the best performance for each metric.
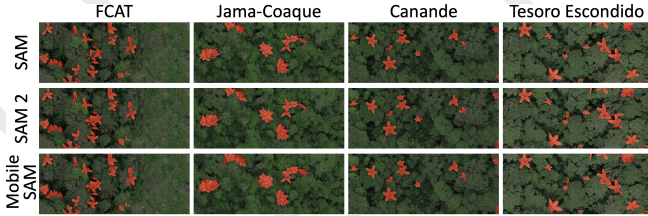

Figure 5: Zero-shot segmentation of SAM variants under distribution shifts. Rows correspond to SAM variants, while columns represent four distinct reserves. The box prompts were derived from the detection model trained on geographically distinct data.

drawn from different ecological sites, and validation data mixed from the two sites. YOLO variants and RT-DETR were trained for 100 epochs, while DINO and DDQ used 30 epochs. Data augmentations, including hue, saturation, and brightness adjustments, along with rotations, scaling, translations, and flips, were employed in training. For inference on landscape orthomosaics, a stride of 400 was used to ensure complete coverage, and NMS was applied to eliminate duplicate predictions. Training was conducted on an RTX 4090 GPU, and testing was performed on multiple hardware platforms: RTX 3060 Laptop, RTX 4090 and H800 GPUs.

Detection was evaluated using precision, recall, AP_{50}, AP_{75}, and mAP. Consistency was assessed by reporting the standard deviation of these metrics across five random experiments. Computational efficiency was evaluated using GFLOPS and parameter counts (Params) to measure model complexity, and FPS to assess inference speed, illustrating

their suitability for real-world applications.

We evaluated the calibration performance of detectors by

$$\text{LaECE}_0 = \sum_{j=1}^{J} \frac{D_j \, |p_j - \text{IoU}_j|}{D}, \text{LaACE}_0 = \sum_{i=1}^{D} \frac{|p_i - \text{IoU}_i|}{D},$$

where $J = 25$ denotes the number of confidence bins, D and D_j represent the total detections and detections in bin j, and p_j and IoU_j are the average confidence and IoU within bin j. For LaACE_0, p_i and IoU_i correspond to per-detection confidence and IoU [Kuzucu *et al.*, 2025]. We evaluated the counting performance using retrieval ratios between predicted and GT centers within $d = 5$ meters and report the median shift of successful matches to quantify spatial deviation. For real-time analysis, we assess the inference speed on raw images that were directly taken from UAVs.

5.2 Detection and Segmentation Performance

Table 1 compares the detection performance when training and testing on different geographical regions within the FCAT reserve, introducing natural distribution shifts. YOLOv10 demonstrates superior efficiency with 177.04 FPS and 31.6M parameters while achieving competitive accuracy, particularly in higher-quality metrics like AP75 (67.94%) and mAP (61.73%). It marginally outperforms other YOLO variants by 0.17% AP_{75} and 0.11% mAP. DINO's lower AP_{50} suggests reduced detection capability under shifted distributions. DDQ attains the best recall (85.66%) and a solid AP_{50}, but its lower FPS limits real-time applicability. RT-DETR achieves the highest precision (88.69%) but low recall (75.98%), indicating numerous missed palms. These results position

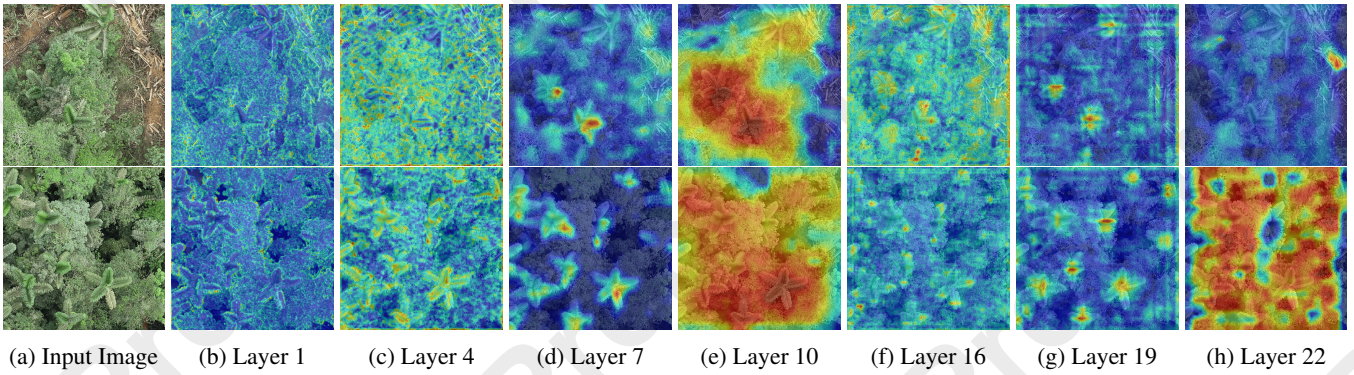| (a) Input Image | (b) Layer 1 | (c) Layer 4 | (d) Layer 7 | (e) Layer 10 | (f) Layer 16 | (g) Layer 19 | (h) Layer 22 |

Figure 6: Feature Saliency in YOLOv10 via GradCAM Visualization. Early layers (b-c) extract boundary and edge details, mid-level layers (d-e) integrate spatial context, and deeper layers (f-h) specialize in scale-specific object detection.

Metric	Model	Uncalibrated	IR	LR	PS	TS
	DINO	4.10%	**2.90%**	3.50%	4.00%	3.70%
	DDQ	2.40%	2.30%	**1.60%**	1.80%	1.90%
	RT-DETR	8.40%	**3.50%**	4.10%	3.60%	5.90%
$LaECE_0$	YOLOv8	3.60%	**3.10%**	4.90%	4.30%	4.00%
	YOLOv9	3.30%	**2.40%**	2.90%	2.70%	3.20%
	YOLOv10	4.30%	4.70%	4.40%	4.20%	**3.60%**
	YOLO11	3.20%	**2.20%**	3.80%	3.40%	3.50%
	DINO	0.16%	0.27%	0.15%	**0.12%**	0.77%
	DDQ	0.92%	0.61%	**0.41%**	0.42%	0.55%
	RT-DETR	8.43%	2.95%	2.84%	2.80%	**1.84%**
$LaACE_0$	YOLOv8	**0.67%**	0.95%	0.92%	0.95%	1.41%
	YOLOv9	**0.15%**	0.19%	0.42%	0.44%	1.01%
	YOLOv10	3.97%	4.12%	4.06%	4.17%	**0.02%**
	YOLO11	0.78%	0.40%	**0.02%**	0.06%	1.11%

Table 2: Comparison of Calibration Performance. The best calibration for each line is highlighted in bold. IR performs best $LaECE_0$, while performance for $LaACE_0$ varies by model.

Figure 7: Impact of Thresholding and Calibration on YOLOv10's Confidence Calibration. Calibration plots show confidence (x-axis) versus mean IoU (y-axis), with $LaECE_0$ and $LaACE_0$ marked.

YOLOv10 as optimal for speed-critical deployments, with DDQ reserved for high-recall scenarios.

Detection analysis under challenging conditions reveals model-specific capabilities (see Figure 4). All models detect large palms reliably, even when overlapping, but exhibit lower confidence on small palms. DETR-based methods demonstrate superior small-object detection performance (first-row top-left/bottom, second-row center), whereas YOLO-based methods excel in occluded scenarios with partial leaf visibility. In such cases, YOLO variants successfully detect individual palms while DETR-based approaches frequently merge detections (first-row bottom, second-row top-right). The sliding-window processing of orthomosaics in PRISM ensures that partially visible palms near edges are typically captured in adjacent patches, thereby compensating for missed edge detections.

We evaluated zero-shot segmentation performance using images from four reserves with distinct ecosystems, where bounding boxes generated by the detection model (trained on geographically distinct regions) introduce distribution shifts when used as prompts during inference, partly due to variations in palm species across sites. Figure 5 demonstrates robust zero-shot segmentation performance despite occasional errors from imperfect bounding box predictions. SAM oc-
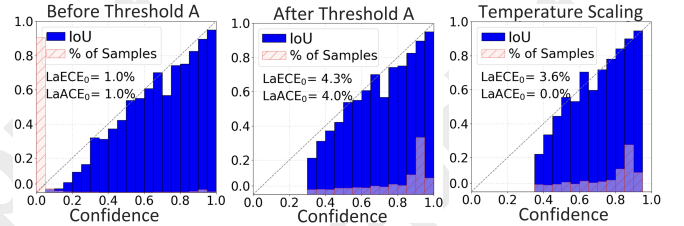
casionally produces fragmented palm leaves due to partial segmentation and background misclassification, while Mobile SAM exhibits background over-inclusion. SAM 2 offers more balanced results, handling areas near palm boundaries more effectively. Considering both segmentation quality and computational efficiency (analyzed in §5.5), SAM 2 is the optimal choice for the segmentation backbone in PRISM.

5.3 Calibration Analysis and Saliency Maps

To enhance confidence reliability in palm detection, we evaluated calibration performance before and after applying calibration techniques. Confidence scores should align with IoU, and LRP-based thresholding (A) is used to filter out low-confidence detections [Kuzucu *et al.*, 2025]. As shown in Table 2, IR achieves the best $LaECE_0$ for most models, while YOLOv10 benefits most from TS and DDQ from LR. For $LaACE_0$, the optimal method varies – TS improves RT-DETR and YOLOv10, whereas LR is most effective for DDQ and YOLO11. Notably, no single method is universally best, and in some cases, lowering $LaECE_0$ increases $LaACE_0$. Figure 7 further illustrates the effect of LRP-based thresholding and calibration on YOLOv10's confidence distribution relative to IoU. Initially, low-confidence predictions (< 0.05) dominate, skewing calibration metrics. LRP filtering refines the distribution by removing unreliable detections. Post-calibration, both $LaECE_0$ and $LaACE_0$ further decrease, indicating improved alignment between confidence scores and IoUs.

Grad-CAM saliency maps illustrate YOLOv10's hierarchical feature extraction across layers. Early layers (Figure 6(b-
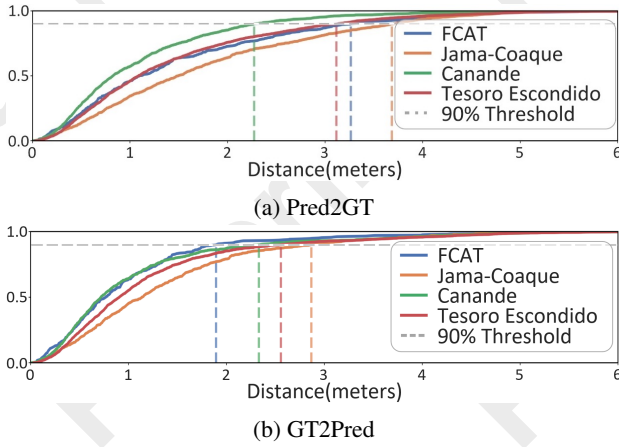
(a) Pred2GT

(b) GT2Pred

Figure 8: Bidirectional Localization Shift Analysis: Cumulative distribution of Pred2GT and GT2Pred alignment.

c)) capture fine-grained details, such as palm leaf boundaries and edge contrasts, progressively refining features from edges to full palm structures. Mid-level layers (Figure 6(d-e)) expand spatial context, enhancing palm localization. Deeper layers specialize in object scales: layer 16 emphasizes small features like individual leaves, layer 19 focuses on medium-sized palm crowns, and layer 22, designed for larger palms, shows limited relevance due to their absence in the scene.

5.4 Counting Performance

Table 3 shows site-dependent palm counting performance. Pred2GT ratios (proportion of predictions matched to GT) remains stable (0.8956–0.9361), while GT2Pred ratios (proportion of GT matched to predictions) vary markedly (0.7667–0.9253). Tesoro Escondido demonstrates strong bidirectional alignment with near-symmetric ratios, indicating balanced performance in localization and detection. FCAT, despite being a distinct sub-region, shows training-data familiarity through its high GT2Pred ratio and the smallest GT2Pred median shift. However, its elevated Pred2GT shift (1.10m) suggests overfitting to site-specific features, yielding "precise but misplaced" predictions. Canande's moderate Pred2GT ratio masks severe recall failures (GT2Pred: 0.767), where unmatched GT palms are likely true misses rather than localization errors. Jama-Coaque, with a low GT2Pred ratio (0.815) and the largest median shifts (Pred2GT: 1.50m, GT2Pred: 1.14m), reflects systematic challenges in both detection and localization. All sites maintain sub-1.5m median shifts, confirming robust localization despite occlusions or partial visibility.

Figure 8 illustrates these trends: Canande's smaller 90th percentile Pred2GT shifts (Figure 8a) align with its precise-but-conservative predictions, while FCAT and Tesoro Escondido's dense canopies enable "proximal counting", where predictions align with clustered GT instances (e.g., overlapping crowns). This phenomenon partially inflates match rates in dense regions despite minor localization inaccuracies. FCAT's strong GT2Pred performance highlights training benefits, whereas Jama-Coaque's weaknesses underscore the need for targeted improvements in low-detection settings.

Site	Area (ha)	Counts	Pred2GT		GT2Pred	
			Ratio	Median (m)	Ratio	Median (m)
FCAT	21.62	471	0.9361	1.10	0.8854	0.77
Jama-Coaque	111.93	952	0.9348	1.50	0.8151	1.14
Canande	101.20	1,273	0.8956	0.82	0.7667	0.72
Tesoro Escondido	86.76	2,330	0.8981	1.09	0.9253	0.91

Table 3: Counting performance across sites. Pred2GT and GT2Pred ratios and median distances quantify bidirectional alignment.

GPU	YOLOv10	Mobile SAM	SAM	SAM 2
RTX3060 Laptop	5.70±0.58	16.49±8.76	58.66±31.21	44.29±23.33
RTX4090	1.68±0.41	7.18±3.67	16.31±8.58	11.87±6.24
H800	1.21±0.37	5.92±3.00	14.05±7.27	10.33±5.31

Table 4: Inference time per image (seconds) across hardware configurations. Mean processing times (± standard deviation) are shown.

5.5 Real-Time Simulation

We evaluated the computational feasibility for real-time onboard processing by simulating the detection-segmentation pipeline on raw UAV imagery using three GPU configurations as shown in Table 4. Testing across 20 images with heterogeneous palm densities demonstrated robust detection performance: YOLOv10 achieves real-time inference speeds (1.21–5.70 s/image) with low temporal variance, confirming that mid-range hardware (an RTX3060 Laptop) remains capable of sustaining real-time detection tasks. Segmentation times, however, show substantial variability, as run-time scales linearly with detected palm count. Although segmentation provides valuable auxiliary visualization, its computational cost makes it optional for latency-critical deployments. These results confirm detection as the time-determining component, with YOLOv10's stability and speed meeting real-time UAV operational requirements.

6 Conclusion

We presented PRISM for automated palm detection and segmentation using UAV imagery, validated on the western Ecuador's ecologically diverse reserves. PALMS dataset (21 sites, 8,830 bounding boxes, 5,026 georeferenced palm centers) captures climatic and species diversity critical for biodiversity monitoring. The modular pipeline achieves real-time processing across GPUs, with sub-1.5m median localization shifts under environmental distribution shifts, while calibration analysis and saliency maps ensure trustworthiness and interpretability. Future work will deploy PRISM on edge devices for UAV-integrated field validation. PRISM's design addresses challenges specific to palms – occlusion, irregular spacing, and lighting variability – which are common to detecting trees in structurally complex environments (e.g., *eastern white pines*). This robustness allows adaptation to other ecologically critical tree species in wild ecosystems, and to lower-resolution satellite imagery (0.5–1m), which offers significantly reduced data acquisition costs (vs. frequent aerial/ground surveys) for scalable ecological monitoring. This positions PRISM as a useful tool for diverse stakeholders, enabling ecologists, conservation practitioners, and even local communities to apply advanced image analysis for better forest management and biodiversity protection.

Acknowledgments

Contribution Statement

Kangning Cui (corresponding author) and Rongkun Zhu contributed equally to this work.

References

[Browne and Karubian, 2016] Luke Browne and Jordan Karubian. Diversity of palm communities at different spatial scales in a recently fragmented tropical landscape. *Botanical Journal of the Linnean Society*, 182(2):451–464, 2016.

[Carion *et al.*, 2020] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020.

[Cui *et al.*, 2024] Kangning Cui, Zishan Shao, Gregory Larsen, Victor Pauca, Sarra Alqahtani, David Segurado, et al. Palmprobnet: A probabilistic approach to understanding palm distributions in ecuadorian tropical forest via transfer learning. In *Proceedings of the 2024 ACM Southeast Conference*, pages 272–277, 2024.

[Eiserhardt *et al.*, 2011] Wolf L Eiserhardt, Jens-Christian Svenning, W Daniel Kissling, and Henrik Balslev. Geographical ecology of the palms (arecaceae): determinants of diversity and distributions across spatial scales. *Annals of Botany*, 108(8):1391–1416, 2011.

[Ferreira *et al.*, 2020] Matheus Pinheiro Ferreira, Danilo Roberti Alves de Almeida, Daniel de Almeida Papa, Juliano Baldez Silva Minervino, Hudson Franklin Pessoa Veras, Arthur Formighieri, Caio Alexandre Nascimento Santos, Marcio Aurélio Dantas Ferreira, Evandro Orfano Figueiredo, and Evandro José Linhares Ferreira. Individual tree detection and species classification of amazonian palms using uav images and deep learning. *Forest Ecology and Management*, 475:118397, 2020.

[Freudenberg *et al.*, 2019] Maximilian Freudenberg, Nils Nölke, Alejandro Agostini, Kira Urban, Florentin Wörgötter, and Christoph Kleinn. Large scale palm tree detection in high resolution satellite images using u-net. *Remote Sensing*, 11(3):312, 2019.

[Gibril *et al.*, 2021] Mohamed Barakat A Gibril, Helmi Zulhaidi Mohd Shafri, Abdallah Shanableh, Rami Al-Ruzouq, Aimrun Wayayok, and Shaiful Jahari Hashim. Deep convolutional neural network for large-scale date palm tree mapping from uav-based images. *Remote Sensing*, 13(14):2787, 2021.

[Hidalgo Pizango *et al.*, 2022] C Gabriel Hidalgo Pizango, Eurídice N Honorio Coronado, Jhon del Águila-Pasquel, Gerardo Flores Llampazo, Johan de Jong, et al. Sustainable palm fruit harvesting as a pathway to conserve amazon peatland forests. *Nature Sustainability*, 5(6):479–487, 2022.

[Jintasuttisak *et al.*, 2022] Thani Jintasuttisak, Eran Edirisinghe, and Ali Elbattay. Deep neural network based date palm tree detection in drone imagery. *Computers and Electronics in Agriculture*, 192:106560, 2022.

[Jocher and Qiu, 2024] Glenn Jocher and Jing Qiu. Ultralytics yolo11. https://github.com/ultralytics/ultralytics, 2024.

[Jocher *et al.*, 2023] Glenn Jocher, Ayush Chaurasia, and Jing Qiu. Ultralytics yolov8. https://github.com/ultralytics/ultralytics, 2023.

[Kirillov *et al.*, 2023] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023.

[Kucharczyk and Hugenholtz, 2021] Maja Kucharczyk and Chris H Hugenholtz. Remote sensing of natural hazard-related disasters with small drones: Global trends, biases, and research opportunities. *Remote Sensing of Environment*, 264:112577, 2021.

[Kuzucu *et al.*, 2025] Selim Kuzucu, Kemal Oksuz, Jonathan Sadeghi, and Puneet K Dokania. On calibration of object detectors: Pitfalls, evaluation and baselines. In *European Conference on Computer Vision*, pages 185–204. Springer, 2025.

[Li *et al.*, 2016] Weijia Li, Haohuan Fu, Le Yu, and Arthur Cracknell. Deep learning based oil palm tree detection and counting for high-resolution remote sensing images. *Remote Sensing*, 9(1):22, 2016.

[Li *et al.*, 2024] Yijie Li, Hewei Wang, and Aggelos Katsaggelos. CPDR: Towards Highly-Efficient Salient Object Detection via Crossed Post-decoder Refinement. In *35th British Machine Vision Conference 2024, BMVC 2024, Glasgow, UK, November 25-28, 2024*. BMVA, 2024.

[Lueder *et al.*, 2022] Sarah Lueder, Kaushik Narasimhan, Jorge Olivo, Domingo Cabrera, Juana Jurado, Lewis Greenstein, and Jordan Karubian. Functional traits, species diversity, and species composition of a neotropical palm community vary in relation to forest age. *Frontiers in Ecology and Evolution*, 10:678125, 2022.

[Ma *et al.*, 2024] Xiaowen Ma, Zhenkai Wu, Rongrong Lian, Wei Zhang, and Siyang Song. Rethinking remote sensing change detection with a mask view, 2024.

[Malhi *et al.*, 2014] Yadvinder Malhi, Toby A Gardner, Gregory R Goldsmith, Miles R Silman, and Przemyslaw Zelazowski. Tropical forests in the anthropocene. *Annual Review of Environment and Resources*, 39:125–159, 2014.

[Muscarella *et al.*, 2020] Robert Muscarella, Thaise Emilio, Oliver L Phillips, Simon L Lewis, Ferry Slik, William J Baker, et al. The global abundance of tree palms. *Global Ecology and Biogeography*, 29(9):1495–1514, 2020.

[Olson and Anderson, 2021] Daniel Olson and James Anderson. Review on unmanned aerial vehicles, remote sensors, imagery processing, and their applications in agriculture. *Agronomy Journal*, 113(2):971–992, 2021.

[Pitman *et al.*, 2014] Nigel CA Pitman, Juan Ernesto Guevara Andino, Milton Aulestia, Carlos E Cerón, David A Neill, Walter Palacios, et al. Distribution and abundance of tree species in swamp forests of amazonian ecuador. *Ecography*, 37(9):902–915, 2014.

[Putra and Wijayanto, 2023] Yoga Cahya Putra and Arie Wahyu Wijayanto. Automatic detection and counting of oil palm trees using remote sensing and object-based deep learning. *Remote Sensing Applications: Society and Environment*, 29:100914, 2023.

[Qin *et al.*, 2021] Zhenwang Qin, Wensheng Wang, Karl-Heinz Dammer, Leifeng Guo, and Zhen Cao. Ag-yolo: A real-time low-cost detector for precise spraying with case study of palms. *Frontiers in Plant Science*, 12:753603, 2021.

[Ravi *et al.*, 2024] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024.

[Selvaraju *et al.*, 2017] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[Sutherland *et al.*, 2013] William J Sutherland, Robert P Freckleton, H Charles J Godfray, Steven R Beissinger, Tim Benton, Duncan D Cameron, et al. Identification of 100 fundamental ecological questions. *Journal of ecology*, 101(1):58–67, 2013.

[Tagle Casapia *et al.*, 2019] Ximena Tagle Casapia, Lourdes Falen, Harm Bartholomeus, Rodolfo Cárdenas, Gerardo Flores, Martin Herold, Eurídice N Honorio Coronado, and Timothy R Baker. Identifying and quantifying the abundance of economically important palms in tropical moist forest using uav imagery. *Remote Sensing*, 12(1):9, 2019.

[Terborgh, 1986] John Terborgh. Community aspects of frugivory in tropical forests. In *Frugivores and seed dispersal*, pages 371–384. Springer, 1986.

[van der Hoek *et al.*, 2019] Yntze van der Hoek, Sara Álvarez Solas, and María Cristina Peñuela. The palm mauritia flexuosa, a keystone plant resource on multiple fronts. *Biodiversity and Conservation*, 28(3):539–551, 2019.

[Wagner *et al.*, 2020] Fabien H Wagner, Ricardo Dalagnol, Ximena Tagle Casapia, Annia S Streher, Oliver L Phillips, Emanuel Gloor, and Luiz EOC Aragão. Regional mapping and spatial distribution analysis of canopy palms in an amazon forest using deep learning and vhr images. *Remote Sensing*, 12(14):2225, 2020.

[Wang *et al.*, 2022] Hewei Wang, Bolun Zhu, Yijie Li, Kaiwen Gong, Ziyuan Wen, Shaofan Wang, and Soumyabrata Dev. Sygnet: a svd-yolo based ghostnet for real-time driving scene parsing. In *2022 IEEE International Conference on Image Processing (ICIP)*, pages 2701–2705. IEEE, 2022.

[Wang *et al.*, 2024a] Ao Wang, Hui Chen, Lihao Liu, Kai Chen, Zijia Lin, Jungong Han, and Guiguang Ding. Yolov10: Real-time end-to-end object detection, 2024. arXiv preprint arXiv:2405.14458.

[Wang *et al.*, 2024b] Chien-Yao Wang, I-Hau Yeh, and Hong-Yuan Mark Liao. Yolov9: Learning what you want to learn using programmable gradient information, 2024. arXiv preprint arXiv:2402.13616.

[Zambrana *et al.*, 2007] Narel Y Paniagua Zambrana, Anja Byg, Jens-Christian Svenning, Monica Moraes, Cesar Grandez, and Henrik Balslev. Diversity of palm uses in the western amazon. *Biodiversity and Conservation*, 16:2771–2787, 2007.

[Zhang *et al.*, 2022] Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M Ni, and Heung-Yeung Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. *arXiv preprint arXiv:2203.03605*, 2022.

[Zhang *et al.*, 2023a] Chaoning Zhang, Dongshen Han, Yu Qiao, Jung Uk Kim, Sung-Ho Bae, Seungkyu Lee, and Choong Seon Hong. Faster segment anything: Towards lightweight sam for mobile applications, 2023. arXiv preprint arXiv:2306.14289.

[Zhang *et al.*, 2023b] Jiguang Zhang, Shilin Xu, Yong Zhao, Jiaxi Sun, Shibiao Xu, and Xiaopeng Zhang. Aerial orthoimage generation for uav remote sensing. *Information Fusion*, 89:91–120, 2023.

[Zhang *et al.*, 2023c] Shilong Zhang, Xinjiang Wang, Jiaqi Wang, Jiangmiao Pang, Chengqi Lyu, Wenwei Zhang, Ping Luo, and Kai Chen. Dense distinct query for end-to-end object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7329–7338, 2023.

[Zhao *et al.*, 2024] Yian Zhao, Wenyu Lv, Shangliang Xu, Jinman Wei, Guanzhong Wang, Qingqing Dang, Yi Liu, and Jie Chen. Detrs beat yolos on real-time object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16965–16974, 2024.

[Zheng *et al.*, 2021] Juepeng Zheng, Haohuan Fu, Weijia Li, Wenzhao Wu, Le Yu, Shuai Yuan, Wai Yuk William Tao, Tan Kian Pang, and Kasturi Devi Kanniah. Growing status observation for oil palm trees using unmanned aerial vehicle (uav) images. *ISPRS Journal of Photogrammetry and Remote Sensing*, 173:95–121, 2021.

[Zou *et al.*, 2023] Zhengxia Zou, Keyan Chen, Zhenwei Shi, Yuhong Guo, and Jieping Ye. Object detection in 20 years: A survey. *Proceedings of the IEEE*, 111(3):257–276, 2023.