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

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

[1]Wake Forest University, [2]City University of Hong Kong, [3]Xidian University, [4]Colby-Sawyer College

Introduction

The Challenge: Palms are vital to tropical ecosystems, but accurately mapping them in dense, natural forests with overlapping crowns and complex backgrounds is extremely difficult.

Our Solution: We developed PRISM, an end-to-end pipeline that automatically detects, segments, and maps individual palms from large-scale UAV orthomosaic images.

Key Features:

- Built on a new dataset with annotations from ecologically diverse sites in Ecuador.
- Integrates efficient object detectors with zero-shot segmentation models for precise localization.
- Includes model calibration and saliency maps to ensure results are both trustworthy and interpretable for ecologists.

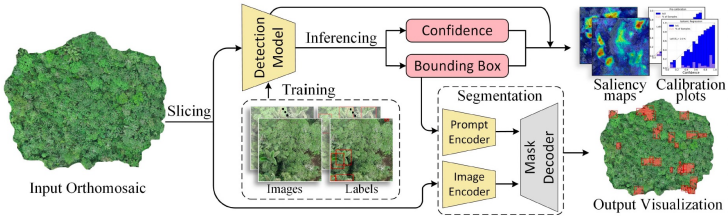
The PRISM Pipeline

Input: The pipeline processes large, high-resolution orthomosaic images created from thousands of aerial drone photos.

Detection: A fine-tuned **YOLOv10** model processes slices of the orthomosaic to efficiently detect palms, generating bounding boxes and confidence scores.

Segmentation: The detected bounding boxes are used as prompts for a **SAM 2** model, which performs zero-shot segmentation to create precise masks for each palm.

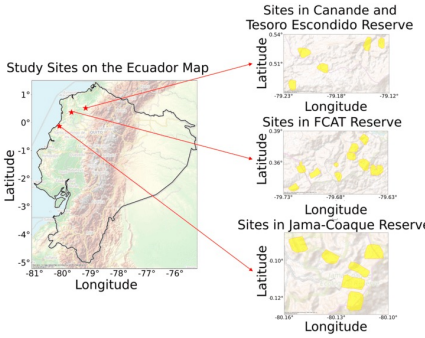
Mapping & Output: The final output consists of precise georeferenced coordinates for each palm, along with visualizations of the bounding boxes and segmentation masks for analysis.



The PALMS Dataset

Key Statistics:

- **Source:** Data collected from 21 ecologically diverse sites across four reserves in Ecuador.
- **Training Data:** Includes 1,500 image patches manually annotated with **8,830 bounding boxes**.
- **Validation Data:** For geographic accuracy, **5,026 palm crown centers** were manually georeferenced across four reserves.



PRISM Code

PALMS Data

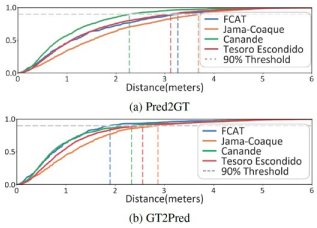
Results & Analysis

A. Detection Performance & Efficiency

YOLOv10 achieves a competitive 61.73% mAP while running at 177.04 FPS on an RTX 4090 GPU. YOLO-based methods excel in detecting partially visible palms, while DETR-based approaches can sometimes merge detections of overlapping crowns.

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	**0.8566 ± 0.0123**	0.8541 ± 0.0129	0.6354 ± 0.0137	0.5736 ± 0.0130
RT-DETR	222.5	65.5	151.49 ± 0.70	**0.8869 ± 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	**169.5**	53.2	114.96 ± 0.30	0.8763 ± 0.0176	0.7976 ± 0.0209	**0.8741 ± 0.0109**	0.6762 ± 0.0146	0.6162 ± 0.0122
YOLOv10	169.8	**31.6**	**177.04 ± 1.14**	0.8716 ± 0.0121	0.7968 ± 0.0089	0.8626 ± 0.0129	**0.6794 ± 0.0112**	**0.6173 ± 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

B. Localization Accuracy & Generalization



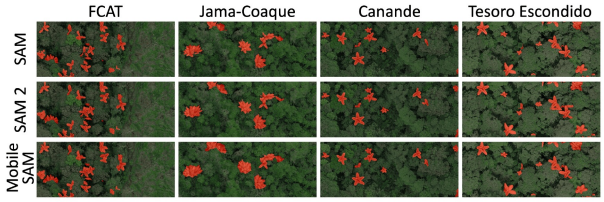
Site	Area (ha)	Counts	Pred2GT		GT2Pred	
			Ratio	Median (m)	Ratio	Median (m)
FCAT	21.62	471	0.9361	1.10	0.8854	0.77
Jama-Coaque	111.93	952	0.9348	1.50	0.8151	1.14
Canande	101.20	1,273	0.8956	0.82	0.7667	0.72
Tesoro Escondido	86.76	2,330	0.8981	1.09	0.9253	0.91

Pred2GT Ratio (Precision): Proportion of predictions matched to a ground truth palm.

GT2Pred Ratio (Recall): Proportion of ground truth palms matched by a prediction.

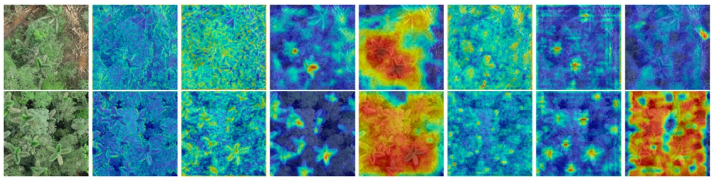
The model achieves a **sub-1.5-meter median localization shift** across all sites, confirming robust performance despite occlusions or partial visibility.

C. Zero-Shot Segmentation

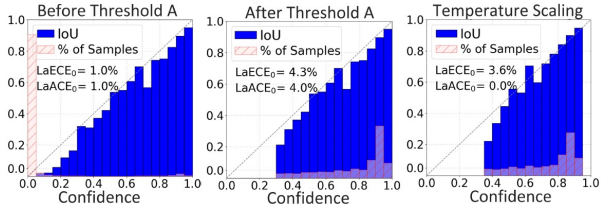


When compared to other variants, SAM 2 offers more balanced results and handles the areas near palm boundaries more effectively. It avoids the fragmentation sometimes seen in the original SAM and the background over-inclusion of Mobile SAM.

D. Trustworthiness & Interpretability

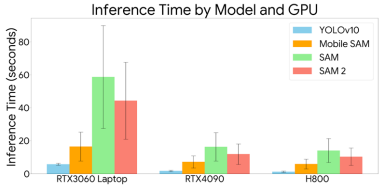


The Hierarchical Feature Learning analysis by Grad-CAM confirms the model learns a meaningful progression: early layers focus on low-level edges and textures; intermediate layers integrate spatial context; and deep layers exhibit focused activation over entire palm crowns.



Calibration techniques like Temperature Scaling are used to better align the model's confidence scores with its actual performance (IoU), making the predictions more trustworthy.

Real-Time Performance



- The YOLOv10 detection model achieves real-time speeds, processing a full, raw image in **5.70 seconds** on a mid-range RTX 3060 Laptop and just **1.21 seconds** on an H800 GPU.
- While segmentation time varies with the number of palms detected, the detection speed confirms PRISM is suitable for real-time operational requirements on a UAV.

Conclusion & Future Work

PRISM is an effective, real-time pipeline for automated palm detection and segmentation in complex tropical forests using UAV imagery. The **PALMS dataset** is a new resource for biodiversity monitoring, containing 8830 bounding boxes and 5026 georeferenced palm centers.

Future work includes deploying PRISM on edge devices for real-time field validation.

WAKE FOREST UNIVERSITY
CityU 香港城市大學 City University of Hong Kong
西安电子科技大学 XIDIAN UNIVERSITY
Colby·Sawyer College
CENTRO DE INNOVACIÓN CIENTÍFICA AMAZÓNICA
THE ANDREW SABIN FAMILY CENTER for ENVIRONMENT AND SUSTAINABILITY