Follow us
@icicleai

THE OHIO STATE UNIVERSITY

WILD-OV: Weakly-supervised Inference for Localization and Detection in Ecological and Agricultural Visual Scenes Using Open Vocabulary Models

Anirudh Potlapally, Pratham Sharma, Nawras Alnaasan, Hari Subramoni

{potlapally.2, sharma.1460, alnaasan.1, subramoni.1}@osu.edu

http://icicle.ai

Source: Kline, Jenna, Samuel, Stevens, Guy, Maalouf, Camille Rondeau, Saint-Jean, Dat Nguyen, Ngoc, Majid, Mirmehdi, David, Guerin, Tilo, Burghardt, Elzbieta, Pastucha, Blair, Costelloe, Matthew, Watson, Thomas, Richardson, and Ulrik Pagh Schultz, Lundquist. "MMLA: Multi-Environment, Multi-Species, Low-Altitude Aerial Footage Dataset". *arXiv preprint arXiv:2504.07744* (2025).

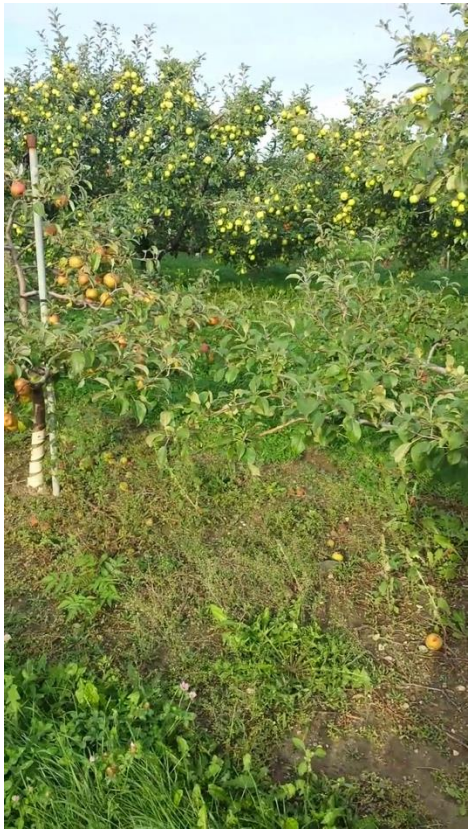
Motivation

- **New vantage points** (e.g., UAVs, UGVs) enable scalable, remote monitoring of complex agro-ecological scenes.
- **High-resolution imagery** reveals fine-grained object interactions—critical for plant phenology, species monitoring, and yield estimation.
- **Manual annotation bottleneck**: Systematic labeling of object/pixel-level data is labor-intensive, error-prone, and unscalable.
- **Low-supervision approaches** (few-/zero-shot) and **open-vocabulary detectors** (OWLv2, Grounding DINO, Florence-2) offer promising alternatives.

Accurate Labelling: Non-trivial task

"Labeling a single image takes up to 30 minutes, which translates to roughly 18 work-hours per 1000 instances"

Source: Hani, Nicolai, Pravakar, Roy, and Volkan, Isler. "MinneApple: A Benchmark Dataset for Apple Detection and Segmentation". *IEEE Robotics and Automation Letters* 5, no. 2 (2020): 852–858.



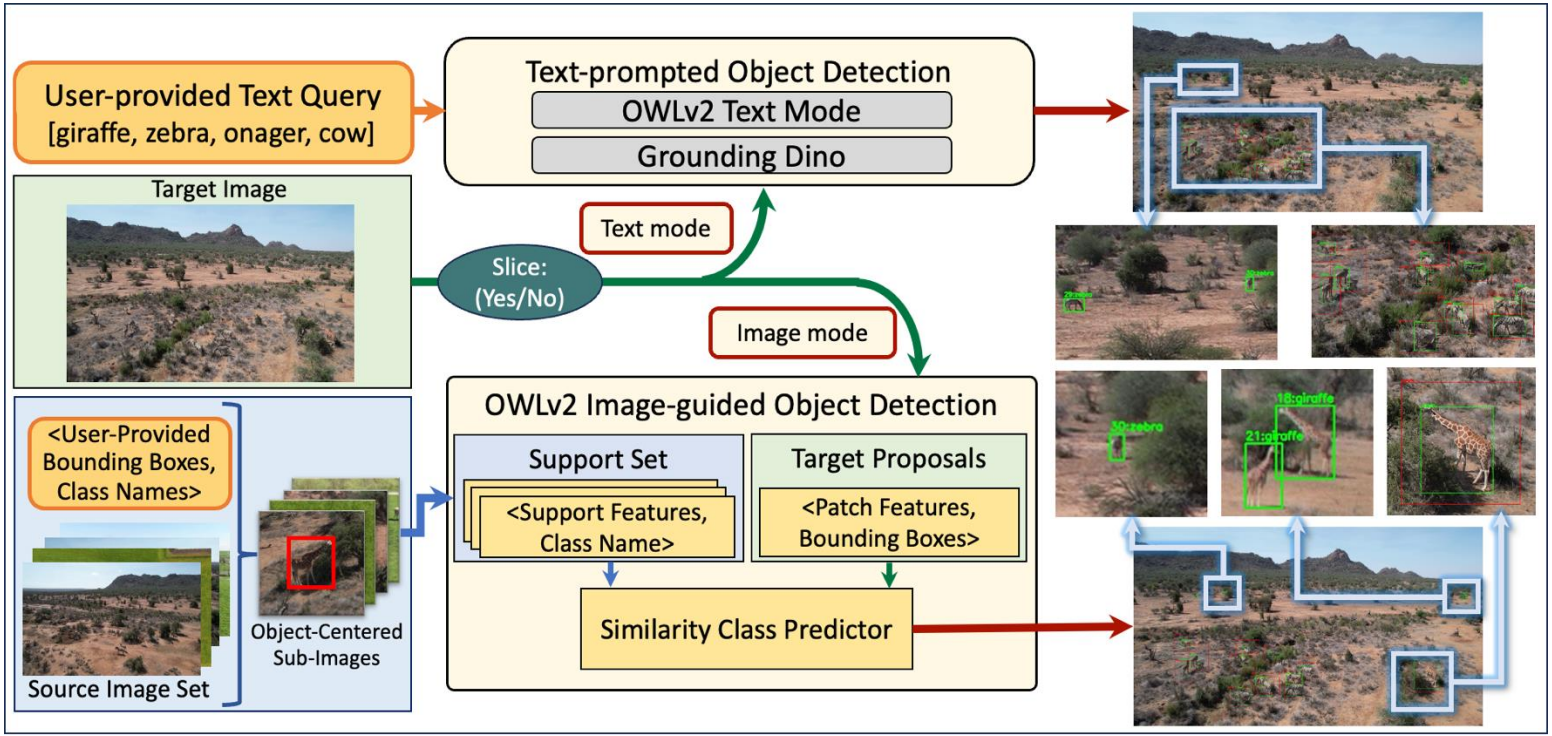
Key Challenges

- **Tiny and dense objects** in high-res imagery yield weak objectness and low detection recall.
- **Domain shift** from natural images to aerial, under-canopy, and biological scenes breaks generalization of pretrained models.
- **Image-guided inference** requires resolution-matched slicing and feature extraction for robust patch-to-patch comparison.
- **Need for systematic evaluation:** Real-world deployment depends on holistic benchmarking across detectors, slicing strategies, and support-image conditioning.

WILD-OV: A Unified Open-Vocabulary Detection Framework for Ecology and Agriculture

- Supports **few- or zero-shot object detection** using state-of-the-art **vision–language models** in both **text- and image-conditioning** modes.
- Effective in detecting objects across real-world ecological and agricultural scenarios.
- Enhances **small-object detection** with **Slicing Aided Hyper Inference (SAHI)**.
- Provides a modular, user-friendly pipeline that lets domain experts choose between full-frame or sliced inference and swap between models with ease.

WILD-OV: Schematic Overview



Schematic overview of WILD-OV

Text-based: Queries are embedded and matched to image features; Image-based: Sample boxes provide region features matched to proposals.

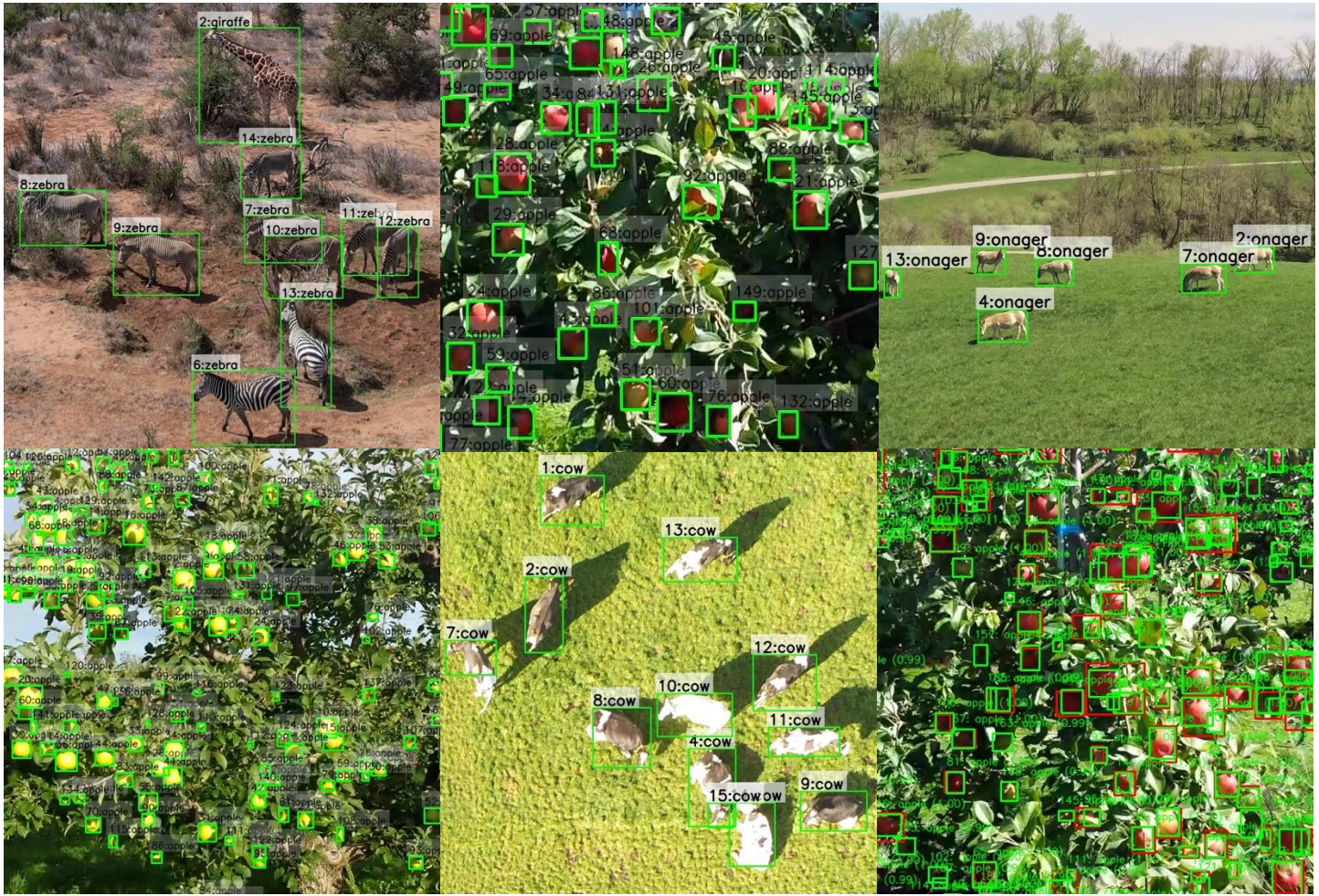

2:giraffe
14:zebra
8:zebra
9:zebra
7:zebra
10:zebra
12:zebra
13:zebra
6:zebra
13:onager
9:onager
8:onager
7:onager
4:onager
1:cow
13:cow
2:cow
7:cow
12:cow
10:cow
8:cow
11:cow
4:cow
9:cow

Table 1: Precision, Recall, and F1 Comparison of Aerial Animal Detection

Metrics	Megadetector		GroundingDino		OWLv2-Text		OWLv2-Imge	
	No SAHI	SAHI	No SAHI	SAHI	No SAHI	SAHI	No SAHI	SAHI
Precision	0.0020	0.3203	0.9975	0.8424	0.8718	0.7272	0.9262	0.8020
Recall	0.0618	0.9756	0.3685	0.8667	0.8680	0.9704	0.8612	0.9616
F1	0.0039	0.4822	0.5382	0.8544	0.8699	0.8314	0.8614	0.8551

Table 2: Precision, Recall, and F1 Comparison of Apple Detection

Metrics	GroundingDino		OWLv2-Text		OWLv2-Imge	
	No SAHI	SAHI	No SAHI	SAHI	No SAHI	SAHI
Precision	0.4883	0.4883	0.5660	0.3688	0.0931	0.4196
Recall	0.1027	0.1027	0.6847	0.9354	0.2114	0.9144
F1	0.1697	0.1697	0.6197	0.5290	0.1293	0.5752
IoU.50	0.0320	0.0320	0.2590	0.3400	0.3740	0.3640

References

1. Matthias Minderer , Alexey A. Gritsenko, and Neil Houlsby. "Scaling Open-Vocabulary Object Detection." In Thirty-seventh Conference on Neural Information Processing Systems.2023.
2. Shilong Liu, , Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. "Grounding DINO: Marrying DINO with Grounded Pre-Training for Open-Set Object Detection." (2024).
3. Akyon, F., Onur Altinuc, S., & Temizel, A. (2022). Slicing Aided Hyper Inference and Fine-Tuning for Small Object Detection. In *2022 IEEE International Conference on Image Processing (ICIP)* (pp. 966–970). IEEE.