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

THE OHIO STATE UNIVERSITY

Anirudh Potlapally, Pratham Sharma, Nawras Alnaasan, Hari Subramoni

{potlapally.2, sharma.1460, alnaasan.1, subramoni.1}@osu.edu, panda@cse.ohio-state.edu

ICICLE DEMOCRATIZING AI

SUMMARY

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.

RESEARCH 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.

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.

SCHEMATIC OVERVIEW OF WILD-OV

Text-based (top): User queries are passed to OWLv2 or Grounding DINO to extract text embeddings and compare them with visual features for object localization.

Image-based (bottom): Exemplar bounding boxes in source images are used by OWLv2 to extract region-level features and match similar proposals in the target image. When slicing-aided hyperinference is enabled, detection operates on overlapping patches to improve small-object recall.

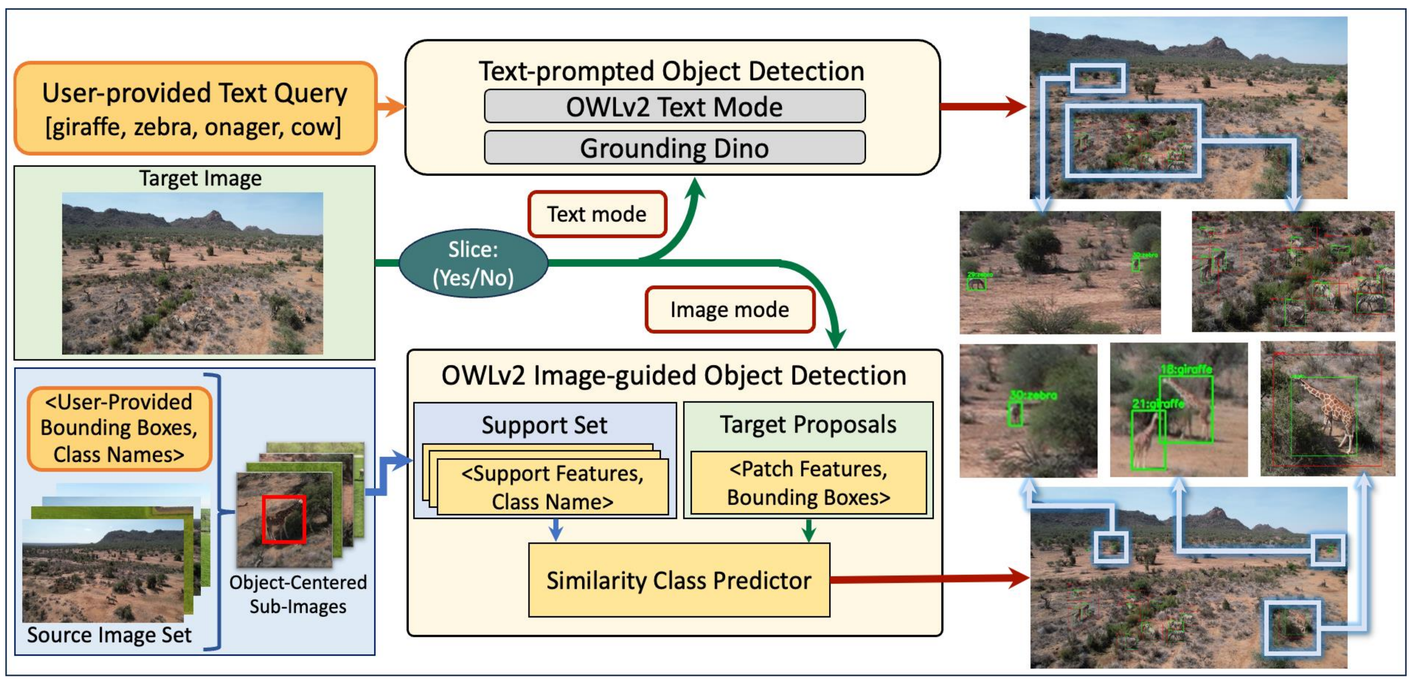

Figure 1: Schematic overview of WILD-OV: Text-based: Queries are embedded and matched to image features; Image-based: Sample boxes provide region features matched to proposals.

In this example, both the text-based and image-based variants of OWLv2 were applied independently to the same scene. It is observed that each variant was able to detect several additional animals in the background that human annotators had overlooked, highlighting the potential of open-vocabulary models to complement manual labeling. Red boxes denote the human-provided ground truth, while green boxes indicate the detections produced by the models.

PERFORMANCE ANALYSIS

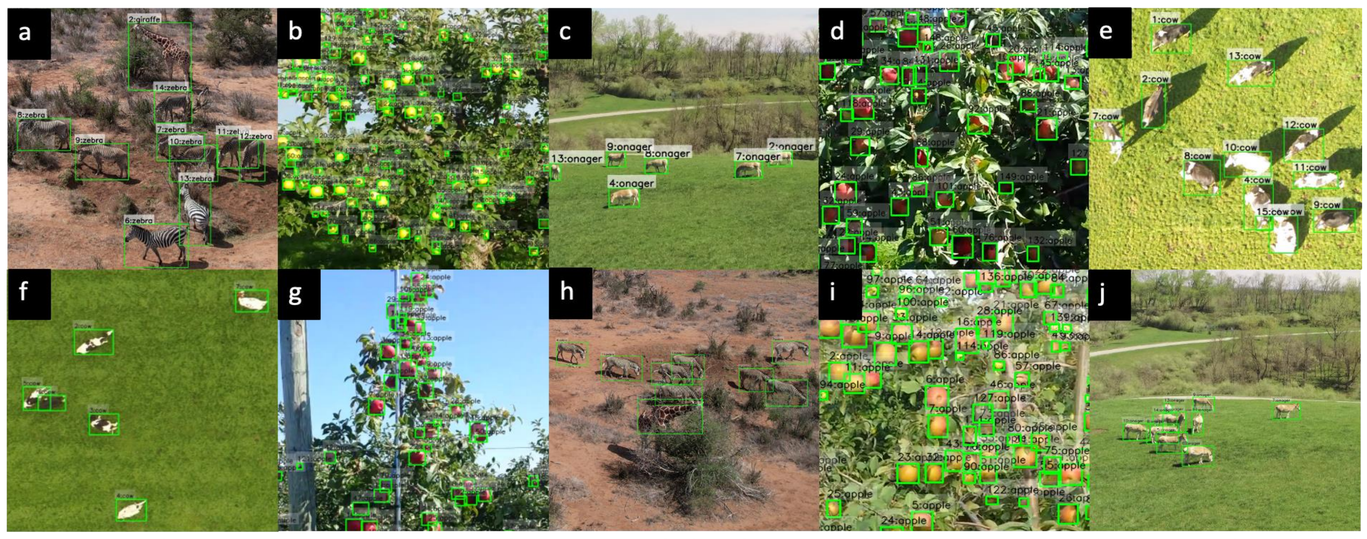

Figure 2: The image depicts the performance of the framework on our target datasets. The model accurately detects and labels different object categories even in complex backgrounds and high occlusion. The results demonstrate the model's ability to generalize well across diverse settings without task-specific training or fine-tuning.

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

MAJOR CONTRIBUTIONS

- Introduced **WILD-OV**, a unified pipeline combining text- and image-guided open-vocabulary detectors with slicing-aided inference.
- Achieved **strong performance without extra training**, uncovering objects often missed by human annotators.
- **OWLv2's dual modes** provide flexibility: text prompts generalize to unseen classes, while image prompts detect small or occluded instances.
- Validated on **ecological (aerial animals)** and **agricultural (orchard apples)** datasets, showing broad applicability.
- Provides domain experts with a **modular, practical toolkit** to accelerate annotation and analysis.

ONGOING WORK

- Expand to **new crop types** and **diverse wildlife datasets**.
- Test detection of **rare/unseen species** with text-guided prompts.
- Benchmark additional VLMs (**Florence-2, Paligemma, BioCLIP**) against OWLv2 and Grounding DINO.
- Explore **lightweight fine-tuning** (LoRA, adapters) for specialized tasks.
- Optimize slicing for **faster, real-time edge deployment**.
- Develop **user-friendly interfaces** to simplify adoption by non-ML experts.

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.