Towards benchmarking Western Bluebird detection in the wild

Estela Monserrat Arriaga Santana[1*] Julian Rosas Scull[1*] Ibeth P. Alarcón[2]
Bibiana Montoya[2] Aylin Sosa Mejía[2] Hugo Jair Escalante[3,4]
[1]Universidad Nacional Autónoma de México
[2]Universidad Autónoma de Tlaxcala
[3]The University of Texas at El Paso
[4]Instituto Nacional de Astrofísica, Óptica y Electrónica
{monse_arriaga,julian.rosas}@ciencias.unam.mx, ibethalarcon1@gmail.com,
bibianac.montoyal@uatx.mx, aylinsm25@gmail.com, hescalantebal@utep.edu
*Equal contribution

Abstract

Bird monitoring in natural environments is challenging due to the small size of some species of birds relative to the scene, background clutter, variability in illumination, and the observers' viewpoint. Progress is further limited by the scarcity of large-scale, realistic datasets, which are essential for understanding behavioral patterns. To address this gap, we introduce a new benchmark dataset for Sialia Mexicana (Western Bluebird) detection and segmentation, comprising over 6,000 labeled images from 41 recording sessions. The dataset features high-resolution (4K) in-the-wild images in which birds occupy only a small fraction of the image. We evaluated a range of methods, including supervised detectors (YOLOv8, YOLOv26, Faster R-CNN, RT-DETR), open-vocabulary models (OWL-ViT, Grounding DINO, YOLO-World) in zero-shot and fine-tuned settings, and segmentation approaches (YOLOv8-Seg, Mask R-CNN, Grounded-SAM). Our results show that the supervised detectors remain the most reliable overall, with Faster R-CNN achieving the highest detection mAP and RT-DETR offering the best precision–recall trade-off. Open-vocabulary models perform poorly in zero-shot settings; however, fine-tuning substantially improves their performance, with YOLO-World becoming competitive with supervised methods and achieving the highest precision, F1-score, and mAP@0.5. For segmentation, supervised methods significantly outperform Grounded-SAM and SAM 3: Mask R-CNN achieves the highest mask mAP, while YOLOv8-Seg provides the best precision and fastest inference. Overall, our findings highlight the difficulty of zero-shot bird detection in cluttered ecological scenes, suggest a more challenging regime for precise localization than is commonly observed in other wildlife and open-vocabulary benchmarks, and underscore the importance of domain adaptation in small-object settings.

1. Introduction

Computer vision has become increasingly relevant for ecological monitoring; however, bird detection in natural environments remains challenging due to small object size, background clutter, illumination variability, and frequent occlusion. Addressing this challenge is important for enabling robust analysis of bird behavior, including migration and breeding patterns, in unconstrained settings.

In this work, we focus on the Western Bluebird (*Sialia mexicana*), a species approximately 19 cm in length that exhibits frequent multi-individual interactions driven by competition for nesting sites. Such behaviors are difficult to capture and annotate under natural conditions, making manual inspection of video recordings both time-consuming and error-prone.

To address this, we extend a novel video dataset of Western Bluebirds, originally collected for behavioral ecology research [1], and repurpose it for the systematic evaluation of computer vision methods. The recordings were obtained during a 2024 field experiment on nest-site competition, using fixed cameras in natural environments.

Unlike standard bird datasets, which typically emphasize close-up classification or aerial monitoring, our setting captures small objects within high-resolution frames, often involving multiple interacting individuals and significant real-world variability such as occlusion and illumination changes. This makes the task particularly challenging and enables a more realistic evaluation of detection methods.

Building on this setup, we conduct a systematic comparison of representative models across supervised, transformer-based, and open-vocabulary detection paradigms.

Our contributions are threefold: (1) we present a curated and manually annotated dataset derived from ecologi-

cal video recordings of bird behavior; (2) we provide a comparative benchmark of supervised, transformer-based, and open-vocabulary models for bird detection, with segmentation as complementary analysis; and (3) we show that fixed-camera bird detection under natural conditions remains a challenging small-object regime in which supervised adaptation is more reliable than zero-shot open-vocabulary detection.

2. Related Work

Camera-trap and ecological imaging datasets have enabled large-scale wildlife monitoring and automated analysis [17, 22]. However, models trained in one ecological setting often struggle to generalize across regions and environments due to changes in species composition, illumination, and background clutter [2]. In bird monitoring, existing datasets are often based on aerial imagery [9, 10, 25] or fine-grained classification benchmarks such as CUB-200-2011, NABirds, and Birdsnap [3, 23, 24], where birds appear large and centered in the frame. By contrast, our dataset focuses on fixed-camera bird detection under challenging natural conditions, where birds occupy only a small fraction of high-resolution images and appear with clutter, motion blur, and viewpoint variation.

Object detection has evolved from two-stage methods such as Faster R-CNN [19] to one-stage detectors such as YOLO [21] [18], and more recently to transformer-based approaches including DETR and RT-DETR [4, 27]. Open-vocabulary detectors such as OWL-ViT and YOLO-World [6, 15], as well as grounding-based models like Grounding DINO [14], extend detection beyond fixed label sets through vision-language alignment. In parallel, segmentation models such as SAM [11] [5] enable prompt-based mask generation. Our benchmark builds on these directions by comparing supervised, transformer-based, and open-vocabulary methods in a fixed-camera setting.

3. A dataset for monitoring Western Bluebird

The source dataset was collected in a controlled behavioral experiment conducted at the **La Malinche National Park (Tlaxcala, Mexico)** by staff of **Estación Científica La Malinche**, part of the *Centro Tlaxcala de Biología de la Conducta, Universidad Autónoma de Tlaxcala*. Recordings captured the activity of 30 breeding pairs of Western bluebirds that interacted with a potential competitor for the nesting site, from a different species (Chipping Sparrow o House Finch), which was placed inside a wire mesh box in a forested area. While this setup is controlled in terms of camera placement and viewpoint, bird behavior occurs naturally, resulting in realistic ecological variability [1].

3.1. Source videos

Source videos were recorded using fixed-position cameras directed toward nest boxes, ensuring a consistent field of view within each recording, see [1] for details. Figure 1 shows sampled frames from the collection. All videos were captured at high resolution (3840×2160). Each experimental trial lasted approximately 13 minutes and was repeated multiple times per nest. Only the first 7 minutes of each of the 41 recordings were used, corresponding to the period immediately after a potential intruder from another species, placed in a wire mesh box, is uncovered, when bird activity is highest. Cameras were placed at different orientations across sessions, resulting in variations in viewpoint, scale, and background composition (Figure 1).

Figure 1. Sample images from the dataset.

3.2. Dataset and Annotations

Frames were extracted from the recorded videos and manually annotated using LabelMe [20]. Each image contains bounding box annotations for individual bird instances, and a subset includes instance segmentation masks for segmentation experiments.

The first version of the dataset consists of **6,016 images** and **7,637 annotated bird instances**, with an average of **1.27 instances per image**. The number of birds per image ranges from one to four: 4,694 with one, 1,043 with two, 259 with three, and 20 with four.

All images are in high-resolution (3840×2160), but birds occupy a very small portion of the frame. The median bounding box area is approximately 2.6×10^4 pixels (less than 0.3% of the image), making the dataset one of the most challenging for small-object detection.

The dataset is organized into 41 independent recording sessions, each corresponding to a distinct experimental trial. Due to the recording setup, all images contain at least one bird instance. Figure 2 shows representative annotated examples.

4. Considered methods

We benchmark bird detection in fixed-camera recordings under a unified experimental setup, with segmentation in-

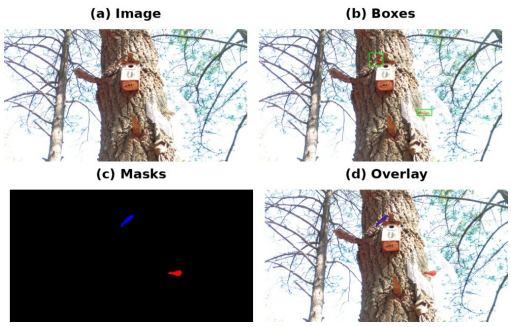

Figure 2. Example from the proposed bird detection dataset. (a) Original image captured under challenging real-world conditions. (b) Ground-truth bounding box annotations. (c) Instance-level segmentation masks. (d) Combined visualization of annotations.

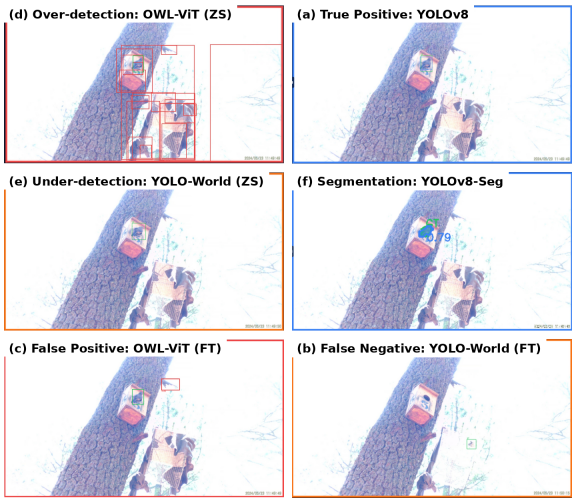

Figure 3. Qualitative results on the test set. (a) True positive from YOLOv8. (b) False negative from YOLO-World (FT). (c) False positive from OWL-ViT (FT). (d) Over-detection from OWL-ViT (ZS). (e) Under-detection from YOLO-World (ZS). (f) Successful segmentation from YOLOv8-Seg. Ground truth is shown in green, correct predictions in blue, and incorrect predictions in red.

cluded as a secondary analysis. We evaluate supervised detectors (YOLOv8, YOLOv26, Faster R-CNN, and RT-DETR) and open-vocabulary models (OWL-ViT, Grounding DINO, and YOLO-World). Supervised models are fine-tuned on the training split using COCO-pretrained weights [13]. Open-vocabulary models are evaluated in zero-shot (ZS) mode using a set of semantically related text prompts, including `bird`, `a bird`, `a flying bird`, `a perched bird`, and `bird in a cage`. When supported, models are also evaluated in a fine-tuned (FT) setting. OWL-ViT and YOLO-World are evaluated in both ZS and FT regimes, whereas Grounding DINO is evaluated only in the zero-shot setting.

Fine-tuned models are trained for up to 100 epochs with early stopping based on validation performance (patience = 20), providing a consistent training budget across methods. For each input image, models predict bounding boxes with confidence scores, and open-vocabulary methods additionally condition predictions on the text query. To complement detection, we evaluate segmentation using two supervised baselines, YOLOv8-Seg and Mask R-CNN, together with Grounded-SAM as a prompt-based zero-shot pipeline that combines Grounding DINO [14] with SAM [11], and SAM 3 [5] as an open-vocabulary zero-shot model.

We report standard COCO-style metrics. For detection, we use mAP@0.5:0.95, mAP@0.5, Precision, Recall, F1-score, and FPS. For segmentation, we report mask mAP@0.5:0.95, mask mAP@0.5, Precision, Recall, F1-score, mIoU, Dice, and FPS. This setup enables a consistent comparison of supervised, transformer-based, and open-vocabulary methods in a challenging small-object ecological setting.

5. Experiments and results

5.1. Experimental Setup

The dataset is split at the recording level to avoid temporal leakage across training, validation, and test sets. In total, it contains 41 recording sessions and 6,016 annotated images, divided into 35 training videos (4,435 images), 5 validation videos (575 images), and 1 held-out test video (1,006 images). Validation set is used for model selection and early stopping, final results are reported on the unseen test split.

All experiments are conducted on the Kaggle platform using NVIDIA T4 GPUs. When supported by the implementation, training is performed with two T4 GPUs; otherwise, a single T4 GPU is used. A fixed image resolution is adopted during both training and inference to ensure consistency across models.

5.2. Detection Results

Table 1 summarizes detection performance on the test set. Among supervised detectors, Faster R-CNN achieves the highest mAP@0.5:0.95, indicating superior localization under stricter IoU thresholds, while RT-DETR provides the best overall balance with the highest recall and strong precision and F1-score, as well as the highest mAP@0.5 among supervised methods. YOLO-based models offer an efficient trade-off: YOLOv8 performs competitively, while YOLOv26 improves precision and recall, achieving the highest F1-score and fastest inference speed, although with lower mAP@0.5:0.95 than Faster R-CNN and RT-DETR. Open-vocabulary models are unstable in the zero-shot set-

Task	Paradigm	Model	mAP	mAP@0.5	Precision	Recall	F1	mIoU	Dice	FPS
Detection	Supervised / FT	YOLOv8	0.3804	0.8428	0.8977	0.7338	0.8076	–	–	13.49
	Supervised / FT	YOLOv26	0.3633	0.8772	0.9704	0.8852	0.9259	–	–	**13.78**
	Supervised / FT	Faster R-CNN	**0.4662**	0.9026	0.4500	0.8987	0.5997	–	–	7.36
	Supervised / FT	RT-DETR	0.4170	0.9340	0.9120	**0.9210**	0.9170	–	–	5.50
	Open-vocabulary / ZS	OWL-ViT	0.0070	0.0189	0.0376	0.4609	0.0695	–	–	3.34
	Open-vocabulary / FT	OWL-ViT	0.1264	0.2770	0.2002	0.2546	0.2241	–	–	4.18
	Open-vocabulary / ZS	Grounding DINO	0.0230	0.0880	0.0880	0.0920	0.0900	–	–	1.18
	Open-vocabulary / ZS	YOLO-World	0.1560	0.4540	0.9000	0.0080	0.0170	–	–	12.37
	Open-vocabulary / FT	YOLO-World	0.4467	**0.9415**	**0.9810**	0.8968	**0.9370**	–	–	5.19
Segmentation	Supervised / FT	YOLOv8-Seg	0.1820	0.6653	**0.7991**	0.7136	0.7539	0.6459	0.7823	**13.83**
	Supervised / FT	Mask R-CNN	**0.6398**	**0.9154**	0.6701	**0.9441**	**0.7838**	**0.8488**	**0.9166**	7.75
	Prompt-based / ZS	Grounded-SAM	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.91
	Open-vocabulary / ZS	SAM 3	0.2229	0.3186	0.2282	0.7252	0.3472	0.1793	0.1931	0.29

Table 1. Benchmark results (test set). For detection, mAP denotes box AP; for segmentation, mAP denotes mask AP.

ting: OWL-ViT over-detects, YOLO-World under-detects, and Grounding DINO remains below supervised baselines. After fine-tuning, YOLO-World becomes competitive, achieving the highest mAP@0.5, precision, and F1-score, though still below Faster R-CNN in mAP@0.5:0.95, while OWL-ViT improves but remains weaker. Overall, recent real-time detectors such as YOLOv26 improve efficiency, but supervised models remain more reliable for precise localization, and domain adaptation is key for open-vocabulary methods.

5.3. Segmentation Results

Bottom rows in Table 1 report segmentation performance on the test set. Supervised methods clearly outperform zero-shot baselines: Mask R-CNN achieves the highest mask mAP, recall, mIoU, and Dice, while YOLOv8-Seg provides the highest precision and fastest inference. Among zero-shot methods, Grounded-SAM fails completely, whereas SAM 3 [5] attains competitive recall (0.7252) but low precision (0.2282) and the slowest inference (0.29 FPS), indicating it detects most birds but produces many spurious masks without domain-specific supervision. Overall, these results show that supervised pixel-level localization remains more reliable in cluttered natural scenes, and that despite promising recall, open-vocabulary segmentation is not yet suitable for precision-critical ecological applications without fine-tuning.

6. Conclusion

We introduced a benchmark for bird detection and segmentation in fixed-camera ecological recordings, capturing a challenging small-object regime where targets occupy only a small fraction of high-resolution frames and appear under occlusion, clutter, and illumination changes. Our results show that this setting remains challenging across modern vision paradigms. Supervised detectors provide the most consistent overall performance, while open-vocabulary models are unreliable in the zero-shot setting and often exhibit unstable precision–recall trade-offs. Fine-tuning substantially improves these models, in some cases making them competitive with supervised approaches, but their performance remains less consistently robust across evaluation criteria. A similar pattern is observed in segmentation, where supervised methods perform strongly, whereas prompt-based and open-vocabulary segmentation fails without task-specific adaptation.

Compared with results commonly reported on wildlife and open-vocabulary benchmarks, ours induces a markedly different ranking across model families. On aerial and wildlife detection benchmarks, YOLO-based and transformer-based detectors often outperform two-stage methods such as Faster R-CNN [16, 28], whereas in our benchmark Faster R-CNN provides the strongest performance on the primary localization metric and RT-DETR is stronger mainly in recall and overall balance. A similar shift is observed in the open-vocabulary setting, where models from the same families achieve substantially stronger results on general-purpose large-vocabulary benchmarks such as LVIS [7, 8, 12, 26], yet zero-shot OWL-ViT, Grounding DINO, and YOLO-World perform very poorly on our dataset [6, 14, 15]. Together, these findings suggest the main challenge of this benchmark is not category recognition alone, but precise localization of very small targets in cluttered ecological scenes under substantial domain shift. More broadly, the near-collapse of ZS open-vocabulary models, together with the sharper trade-offs observed even after adaptation, indicates that fixed-camera ecological recordings capture a failure regime that remains underrepresented in standard detection benchmarks. We hope this dataset and benchmark support future work on small-object perception, domain adaptation, and robust detection and segmentation methods for ecological applications.

References

[1] Ibeth Paola Alarcón Vásconez. Efectos parentales pre-eclosión mediados por cambios conductuales: Alteraciones en los patrones de incubación por la presencia de una especie competidora y posibles consecuencias en el fenotipo de la descendencia. Tesis de maestría en ciencias biológicas, Universidad Autónoma de Tlaxcala, Tlaxcala, México, 2025. Centro Tlaxcala de Biología de la Conducta (CTBC). 1, 2

[2] Sara Beery, Dan Morris, and Siyu Yang. Efficient pipeline for camera trap image review. In *CVPRW*, 2019. 2

[3] Travis Berg et al. Birdsnap: Large-scale fine-grained categorization of birds. In *CVPR*, 2014. 2

[4] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, 2020. 2

[5] Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, Jie Lei, Tengyu Ma, Baishan Guo, Arpit Kalla, Markus Marks, Joseph Greer, Meng Wang, Peize Sun, Roman Rädle, Triantafyllos Afouras, Effrosyni Mavroudi, Katherine Xu, Tsung-Han Wu, Yu Zhou, Liliane Momeni, Rishi Hazra, Shuangrui Ding, Sagar Vaze, Francois Porcher, Feng Li, Siyuan Li, Aishwarya Kamath, Ho Kei Cheng, Piotr Dollár, Nikhila Ravi, Kate Saenko, Pengchuan Zhang, and Christoph Feichtenhofer. Sam 3: Segment anything with concepts. 2025. 2, 3, 4

[6] Tianheng Cheng et al. Yolo-world: Real-time open-vocabulary object detection. *arXiv preprint arXiv:2401.17270*, 2024. 2, 4

[7] Shenghao Fu, Qize Yang, Qijie Mo, Junkai Yan, Xihan Wei, Jingke Meng, Xiaohua Xie, and Wei-Shi Zheng. Llmdet: Learning strong open-vocabulary object detectors under the supervision of large language models. *arXiv preprint arXiv:2501.18954*, 2025. 4

[8] Agrim Gupta, Piotr Dollar, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation. In *CVPR*, 2019. 4

[9] Madeline C. Hayes et al. Drones and deep learning produce accurate monitoring of seabird colonies. *Ornithological Applications*, 123(3):duab022, 2021. 2

[10] Suk-Ju Hong et al. Application of deep-learning methods to bird detection using uav imagery. *Animals*, 13(5):902, 2023. 2

[11] Alexander Kirillov et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023. 2, 3

[12] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, Kai-Wei Chang, and Jianfeng Gao. Grounded language-image pre-training. In *CVPR*, 2022. 4

[13] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *ECCV*, 2014. 3

[14] Shilong Liu et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. *arXiv preprint arXiv:2303.05499*, 2023. 2, 3, 4

[15] Matthias Minderer, Alexey Gritsenko, Austin Stone, Maxim Neumann, and Dirk Weissenborn. Simple open-vocabulary object detection with vision transformers. In *ECCV*, 2022. 2, 4

[16] Chao Mou, Tengfei Liu, Chengcheng Zhu, and Xiaohui Cui. Waid: A large-scale dataset for wildlife detection with drones. *Applied Sciences*, 13(18):10397, 2023. 4

[17] Mohammad Sadegh Norouzzadeh et al. Automatically identifying, counting, and describing wild animals in camera-trap images with deep learning. *PNAS*, 115(25):E5716–E5725, 2018. 2

[18] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *CVPR*, 2016. 2

[19] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NIPS*, 2015. 2

[20] Bryan C. Russell, Antonio Torralba, Kevin P. Murphy, and William T. Freeman. Labelme: A database and web-based tool for image annotation. *IJCV*, 77(1–3):157–173, 2008. 2

[21] Ranjan Sapkota et al. Yolo26: Key architectural enhancements and performance benchmarking for real-time object detection. *arXiv preprint arXiv:2509.25164*, 2025. 2

[22] Alexandra Swanson et al. Snapshot serengeti, high-frequency annotated camera trap images of 40 mammalian species in an african savanna. *Scientific Data*, 2:150026, 2015. 2

[23] Grant Van Horn et al. Building a bird recognition app and large scale dataset. In *CVPRW*, 2015. 2

[24] Catherine Wah et al. The caltech-ucsd birds-200-2011 dataset. Technical report, Caltech, 2011. 2

[25] Ben G. Weinstein et al. A general deep learning model for bird detection in high-resolution airborne imagery. *Ecological Applications*, 32(3):e2558, 2022. 2

[26] Lewei Yao, Renjie Pi, Jianhua Han, Xiaodan Liang, Hang Xu, Wei Zhang, Zhenguo Li, and Dan Xu. Detclipv3: Towards versatile generative open-vocabulary object detection. *arXiv preprint arXiv:2501.06066*, 2025. 4

[27] Yian Zhao et al. Detrs beat yolos on real-time object detection. In *CVPR*, 2024. 2

[28] Yu Zhou and Yan Wei. Uav-detr: An enhanced rt-detr architecture for efficient small object detection in uav imagery. *Sensors*, 25(15):4582, 2025. 4