Towards Adaptable and Deployable Wildlife Detection from Aerial Imagery

by

Jayden Hsiao

A thesis
presented to the University of Waterloo
in fulfillment of the
thesis requirement for the degree of
Master of Applied Science
in
Systems Design Engineering

Waterloo, Ontario, Canada, 2026

© Jayden Hsiao 2026

Author's Declaration

This thesis consists of material all of which I authored or co-authored: see Statement of Contributions included in the thesis. This is a true copy of the thesis, including any required final revisions, as accepted by my examiners.

I understand that my thesis may be made electronically available to the public.

Statement of Contributions

Jayden Hsiao is the sole author of all chapters except Chapters 3–6. All chapters were written under the supervision of Dr. David A. Clausi and Dr. Lincoln Xu. Chapters 3 and 4 include material published as an arXiv preprint [99], with expanded methods, experiments, and discussion in this thesis. Chapter 5 incorporates work published in IAAI 2026 [42], and Chapter 6 includes material accepted for the ISPRS 2026 Toronto conference [43]. While some content overlaps in motivation, dataset description, system design, and selected figures, this thesis provides a more detailed account of system architecture and deployment.

The inclusion of previously disseminated material follows publisher policies. OpenWildlife is publicly available on arXiv. Copyright of the IAAI paper is held by AAAI; reproduction in this thesis is permitted under the AAAI Author Agreement with proper acknowledgement. Copyright of the ISPRS paper remains with the author; reproduction here is permitted under the Creative Commons Attribution licence with proper acknowledgement. All reused material is appropriately attributed, and no copyrighted content beyond the author's contributions is reproduced.

Chapter	Citation	Author's Contributions
3-4	M. Patel, J. N. Turnes, J. Hsiao, L. Xu, and D. A. Clausi, "OpenWildlife: Open-Vocabulary Multi-Species Wildlife Detector for Geographically-Diverse Aerial Imagery," June 2025. Available via arXiv [99].	Consolidated dataset, developed method, ran experiments, and visualized results. Co-first author with M. Patel and J. N. Turnes.
5	J. Hsiao, A. Kalia, Z. Zhang, H. Sun, M. Patel, D. A. Clausi, L. Xu, B. Segal, and J. Heath, "Human-in-the-Loop Eider Duck Counting in Arctic Canada with an Open-Vocabulary Multi-Species Wildlife Detector," Proceedings of the AAAI Conference on Artificial Intelligence, March 2026 [41].	Customized annotation interface, conducted user study, and created figures. Co-first author with A. Kalia, Z. Zhang, H. Sun, and M. Patel.
6	J. Hsiao, M. Patel, D. A. Clausi, L. Xu, "DINOKey: Transformer-Based Keypoint Detection for Wildlife Monitoring in Aerial Imagery," accepted for presentation at ISPRS 2026, Toronto, Canada, Apr. 2026 [43].	Developed method, ran experiments, and created figures. Co-first author with M. Patel.

Abstract

Reliable wildlife monitoring is critical for biodiversity conservation, yet large-scale aerial surveys remain constrained by labour-intensive manual counting and limited model generalization. While camera trap–based approaches are used for species identification and behavioural monitoring at fixed locations, aerial imagery acquired from drones and aircraft enables large-area coverage and is commonly used to estimate population abundance, species distributions, and temporal trends through repeated surveys. However, these workflows remain heavily reliant on manual annotation. Although recent advances in computer vision enable automated wildlife detection, many existing approaches rely on species-specific training data, exhibit poor transferability across environments, and lack mechanisms to incorporate expert corrections into iterative model improvement. These limitations hinder scalability and long-term deployment in operational conservation workflows.

We propose OpenWildlife as an open-vocabulary, multi-species wildlife detection framework designed for RGB aerial imagery captured from drones or aircraft, together with a human-in-the-loop (HITL) annotation system that supports incremental model refinement. The framework adapts a language-grounded detection architecture to allow species specification through natural language prompts, enabling flexible identification across terrestrial and marine environments without retraining for each new category. Trained on 15 publicly available wildlife datasets, the model achieves up to 0.981 mAP50 (mean Average Precision at 50% overlap, a standard object detection metric) under fine-tuning and 0.597 mAP50 on seven datasets containing novel species in zero-shot settings, demonstrating generalization to unseen species across diverse vertebrate groups, including mammals and birds.

To support practical deployment, the detection model is integrated into a HITL workflow that combines automated pre-labelling, regional human correction, and incremental fine-tuning. While these components are individually well-established, the proposed system integrates them into a unified workflow tailored to aerial wildlife imagery. This design avoids exhaustive full-image annotation in dense scenes and enables expert feedback to directly improve subsequent model iterations. A case study conducted with the Arctic Eider Society on high-resolution aerial surveys of eider ducks in Arctic Canada demonstrates practical impact: the system achieves 77.6% recall with a 22.2% counting error while reducing annotation time by 87.4% compared to fully manual labelling, demonstrating its applicability for semi-automated population abundance estimation from aerial surveys.

These results demonstrate that combining open-vocabulary detection with human-in-the-loop learning provides a scalable and adaptable approach for wildlife monitoring, enabling efficient and consistent large-area surveys across diverse species and environments.

Acknowledgements

I am deeply grateful to everyone who has supported me throughout this journey.

First, I extend my sincere thanks to my graduate supervisors, Prof. David Clausi and Prof. Lincoln Xu. I am especially appreciative of Prof. Clausi for encouraging me to pursue opportunities beyond my remote sensing research and for fostering a culture where I felt empowered to do my best work. Your guidance, feedback, and mentorship have been invaluable in shaping both this thesis and my development as a researcher. I also thank my thesis readers, Prof. Yuhao Chen and Dr. Chantel Markle, for their expertise in computer vision and ecological conservation, which greatly enhanced the quality and rigour of this work.

I am fortunate to have been part of the Vision and Image Processing Research Group and Systems Design Engineering community. My colleagues, particularly Javier, Lily, Muhammed, Fernando, and Kiernan, provided collaboration, advice, and companionship that sustained me throughout this process. I am equally grateful to the Arctic Eider Society partners, Becky Segal and Joel Heath, for their openness, support, and willingness to engage with this research, which enriched both the practical and ethical dimensions of my work. I also thank my undergraduate research assistants Aryan, Leo, and Naman for their energetic support and willingness to learn.

On a personal note, I am thankful for the friends who have made this journey joyful and meaningful. To "The Drunken Noodles" and the wider communities I have met through dance, running, badminton, Magic the Gathering, interactive art, book club, and poetry: your friendship, curiosity, and support have made all the difference.

I am also grateful for my family for celebrating every small milestone, supporting me through challenges, and taking an interest in my research despite its distance from their own work. Thank you as well for travelling to conferences with me, turning professional milestones into shared memories that I will treasure. Your encouragement and presence have been a constant source of strength.

Finally, I acknowledge the role of financial and technical support in making this research possible. Funding from the University of Waterloo Engineering Excellence Master's Fellowship and computing resources provided by the Digital Research Alliance of Canada enabled me to pursue this degree and carry out the work presented in this thesis.

Land Acknowledgement

I acknowledge that much of my work takes place on the traditional territories of the Neutral, Anishinaabeg, and Haudenosaunee peoples. The University of Waterloo's main campus is situated on the Haldimand Tract, land granted to the Six Nations that spans six miles on each side of the Grand River. This research, which focuses on developing scalable and adaptable methods for wildlife monitoring from aerial imagery, occurs on lands with rich ecological and cultural histories.

I recognize that Indigenous communities have long maintained deep knowledge of local species, ecosystems, and seasonal patterns, and that effective conservation depends not only on technological innovation but also on ethical engagement with these knowledge systems. I am committed to designing tools that can support collaborative monitoring efforts, reduce labour burdens in data collection, and respect local stewardship practices.

By situating this work within the context of Indigenous lands and knowledge, I aim to ensure that technological advances in biodiversity monitoring are applied responsibly, acknowledging both the ecological and cultural significance of the landscapes in which these systems operate.

Dedication

To my family, whose sacrifices, encouragement, and constant presence gave me the confidence to pursue this degree.

Table of Contents

Author's Declaration ii

Statement of Contributions iii

Abstract iv

Acknowledgements v

Land Acknowledgement vi

Dedication vii

List of Figures xii

List of Tables xiii

1 Introduction 1

1.1 Motivation 1

1.2 Thesis Contributions 4

1.3 Thesis Organization 4

2 Background 6

2.1 Deep Learning Foundations 6

2.1.1 Neural Networks and Representation Learning . . . 6
2.1.2 Convolutional Neural Networks . . . 7
2.1.3 Attention Mechanisms and Transformers . . . 8
2.1.4 Closed-set and Open-set Prediction . . . 9
2.1.5 Learning with Limited Supervision . . . 10
2.1.6 Foundation Models . . . 11
2.1.7 Object Detection . . . 12
2.2 AI for Biodiversity Monitoring . . . 17
2.2.1 Biodiversity Monitoring . . . 17
2.2.2 Population Counting from Aerial Imagery . . . 18
2.3 Human-in-the-Loop Learning . . . 20

3 Dataset **22**
3.1 Overview . . . 22
3.2 Cross-Dataset Consolidation . . . 22

4 OpenWildlife Model **27**
4.1 Overview . . . 27
4.2 Architecture . . . 27
4.2.1 Caption Generation . . . 28
4.2.2 Feature Extraction and Fusion . . . 29
4.2.3 Language-Guided Query Selection . . . 29
4.2.4 Cross-Modality Decoder . . . 30
4.2.5 Training Losses . . . 30
4.3 Experimental Protocol . . . 31
4.3.1 Implementation and Training . . . 31
4.3.2 Caption Generation . . . 32
4.3.3 Comparisons with Closed-Set SOTA and Literature . . . 32

4.4 Results . . . 33
4.4.1 Closed-Set SOTA Comparison . . . 33
4.4.2 Literature Comparison . . . 34
4.4.3 Detection of Novel Labels . . . 36
4.4.4 Detection in Novel Environments . . . 39
4.4.5 Limitations . . . 39

5 Deployment **42**
5.1 Deployment-specific Dataset . . . 43
5.2 Workflow Implementation . . . 44
5.3 Experiments . . . 44
5.4 Results . . . 46
5.4.1 Accuracy and Efficiency . . . 46
5.4.2 Generalization Across Images . . . 46
5.4.3 Qualitative Analysis . . . 48
5.5 Summary . . . 48

6 Conclusion **50**
6.1 Summary . . . 50
6.2 Future Work . . . 51
6.2.1 Ecological Impact Assessment . . . 51
6.2.2 Multi-Modal Sensor Integration . . . 52
6.2.3 Enhanced Caption Generation and Alignment . . . 53
6.2.4 Active Learning Integration . . . 54
6.2.5 Keypoint-Native Detection for Wildlife Monitoring . . . 55

References **57**

Appendices **75**

A Dataset Metadata **76**

B Human-in-the-loop Interface Details **82**

List of Figures

1.1 Traditional versus proposed (OpenWildlife) wildlife monitoring workflows. 3

3.1 Global distribution of the aerial wildlife datasets used in this work. 23
3.2 Example imagery from the consolidated dataset 26

4.1 OpenWildlife architecture diagram. 28
4.2 Comparison of OW with closed-set detectors and MM-Grounding-DINO. . 34
4.3 Qualitative detection results on novel species and in novel environments. . 38
4.4 Limitations of novel species detection. 41

5.1 Visual characteristics of eider ducks in aerial surveys. 43
5.2 The human-in-the-loop pipeline built within LabelStudio. 45
5.3 Qualitative improvement through human-in-the-loop refinement. 48
5.4 Comparison of fully manual annotation and the human-in-the-loop workflow. 49

B.1 Example of the bubble clustering algorithm used to improve UI performance. 83
B.2 Enhanced model settings page, containing options for model versioning. . . 84
B.3 Settings dialog to edit model training and testing parameters. 84
B.4 Software design diagram of prediction and training workflows. 85

List of Tables

3.1 Dataset partitioning for open-set continuous pre-training and evaluation. . 25

4.1 Key training hyperparameters for OW model pre-training and fine-tuning. 31
4.2 Training loss weights and Hungarian matching cost settings for the OW model. 31
4.3 Sample keyword and sentence captions for representative datasets. 33
4.4 Results of open-set continuous fine-tuning. 35
4.5 Novel species evaluation. 36
4.6 Finetune evaluation. 37
4.7 Novel environment evaluation: OW models 39
4.8 Novel environment evaluation: Finetune 40

5.1 Performance comparison of model variants on the held-out test set. 47
5.2 Annotation time comparison: manual labelling versus HITL. 47
5.3 Generalization performance after HITL on non-overlapping images. 47

A.1 The 14 datasets used for open-set continuous fine-tuning. 76
A.2 The seven datasets used for testing on novel detection categories. 79
A.3 The three datasets used for evaluation in novel environments. 81

Chapter 1

Introduction

1.1 Motivation

Reliable wildlife monitoring is fundamental to biodiversity conservation, population management, and evidence-based environmental policy. Estimates of species abundance and distribution inform decisions ranging from habitat protection to harvest regulation and climate adaptation strategies. Biodiversity indicators, including species richness, abundance, and population dynamics, are widely used to assess ecosystem health and guide conservation planning [73, 91, 13, 83]. As ecological systems face accelerating pressures from habitat loss, climate change, and human activity, the need for timely, accurate, and scalable monitoring has become increasingly urgent.

Aerial imagery has emerged as a powerful tool for surveying wildlife across large and remote areas. High-resolution sensors mounted on crewed aircraft or uncrewed aerial systems enable rapid coverage of expansive landscapes that would otherwise be logistically inaccessible [118, 66, 25]. However, translating imagery into actionable ecological data remains a major bottleneck. Counting and annotating animals in aerial imagery is labour-intensive, time-consuming, and subject to observer variability. In dense scenes or challenging environmental conditions, manual annotation can require substantial effort while still producing inconsistent results [65, 60, 11].

Recent advances in computer vision have enabled automated wildlife detection and counting from imagery. In particular, deep learning–based object detection models have demonstrated strong performance across a wide range of visual recognition tasks [63, 75]. These approaches leverage hierarchical representation learning to automatically extract

relevant visual features from raw image data [8]. When trained on sufficiently large and well-annotated datasets, modern detection architectures can localize and classify multiple objects within an image with high accuracy [14].

Despite these advances, many existing wildlife detection systems rely on species-specific training data and are optimized for narrowly defined scenarios [116, 109, 5]. Training a new model for each species or survey context requires substantial annotation effort and technical expertise. Furthermore, models trained in one environment often exhibit significant performance degradation when applied to new geographic regions, sensor types, or species [51, 6]. This limitation reflects a broader challenge in computer vision: models trained under closed-set assumptions typically struggle when confronted with previously unseen categories or distribution shifts [31, 114]. As a result, the practical deployment of automated detection systems in real-world conservation workflows remains limited.

Another persistent limitation is the disconnect between model predictions and human corrections. In many operational workflows, automated detections are generated by a model and subsequently reviewed by experts, who correct or refine the predicted labels for reporting purposes [115, 51, 93]. However, these human corrections are often not incorporated into subsequent model updates. As a result, expert labelling effort does not accumulate into progressively improved models, and the system fails to benefit from feedback collected during annotation. Interactive machine learning and active learning research have shown that integrating human corrections directly into the training process can significantly improve model performance and labelling efficiency [115, 3, 57]. The absence of such feedback mechanisms in many wildlife monitoring pipelines therefore represents a key barrier to scalable, long-term monitoring.

At the same time, recent developments in large-scale pre-training and foundation models offer new opportunities for addressing these challenges. Models trained on massive and diverse datasets can learn general-purpose representations that transfer across tasks and domains [9]. In particular, multimodal models that jointly learn visual and textual representations enable open-vocabulary recognition, allowing categories to be specified through natural language rather than fixed label sets [103, 136]. These capabilities provide a promising pathway toward wildlife monitoring systems that can adapt to new species and environments with minimal retraining. When combined with human-in-the-loop learning, such systems can support iterative refinement, allowing expert feedback to directly improve future predictions. Leveraging these advances in a way that is robust, adaptable, and deployable in real conservation contexts motivates the work presented in this thesis. Figure 1.1 summarizes the conceptual shift from traditional wildlife monitoring pipelines to the proposed OpenWildlife workflow, highlighting the role of open-vocabulary detection and iterative human-in-the-loop refinement.

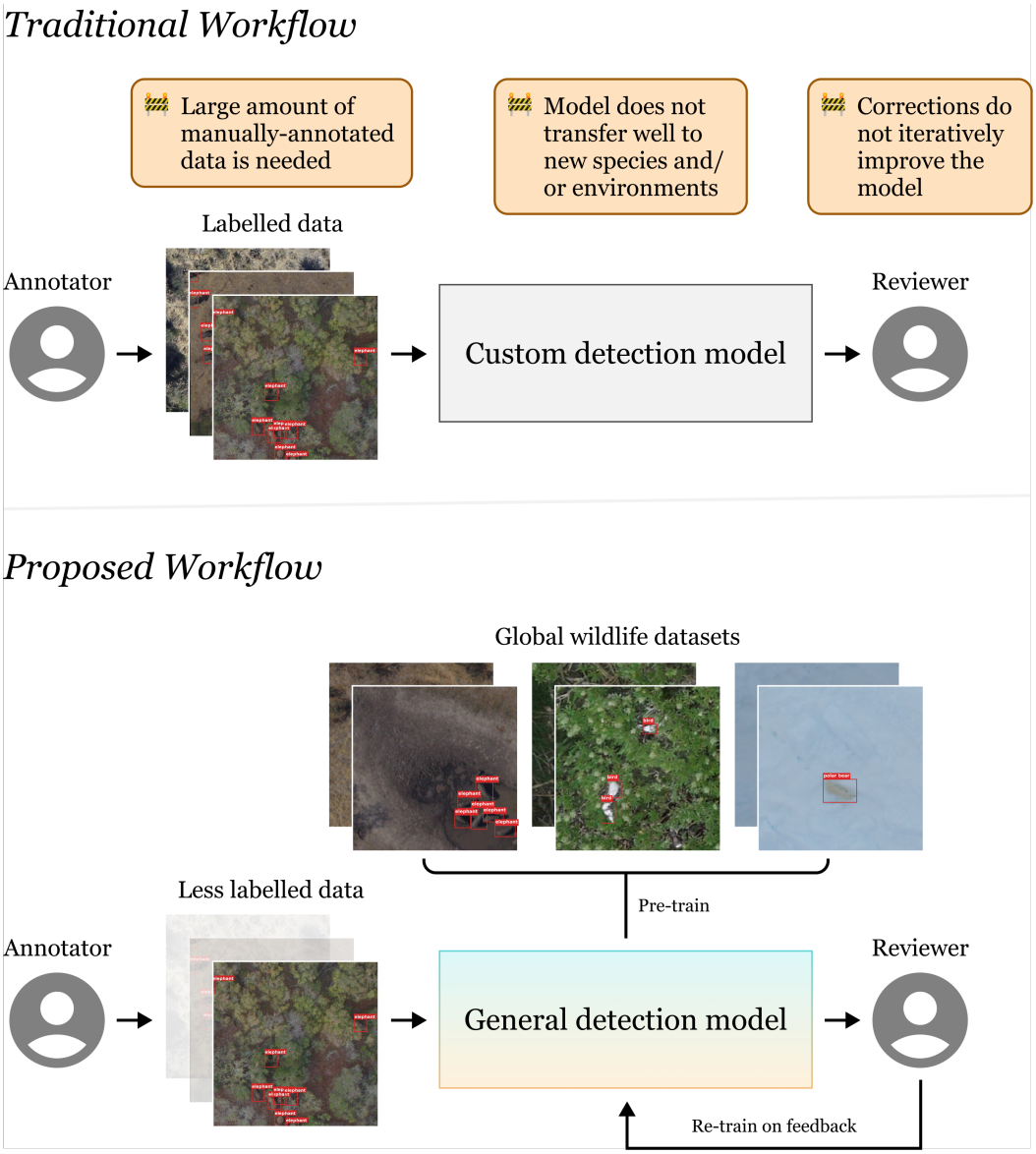


Figure 1.1: Comparison between traditional wildlife monitoring workflows and the proposed OpenWildlife pipeline. Traditional approaches rely on large manually labelled datasets to train species-specific models and do not incorporate expert corrections into model updates. OpenWildlife instead leverages pre-training on global wildlife datasets for open-vocabulary detection and integrates human corrections through iterative refinement.

1.2 Thesis Contributions

This thesis introduces OpenWildlife, an open-vocabulary, multi-species wildlife detection framework designed for aerial imagery, together with a human-in-the-loop annotation system for incremental refinement. The central contributions are as follows:

1. A language-enabled wildlife detection model that supports open-vocabulary species specification through natural language prompts, enabling flexible identification across terrestrial and marine environments without retraining for each new category.
2. A unified benchmark constructed from publicly available wildlife datasets, with dataset splits explicitly designed to evaluate cross-species and cross-environment generalization.
3. A human-in-the-loop annotation workflow that integrates automated predictions, regional human correction, and incremental fine-tuning, allowing expert feedback to directly improve subsequent model iterations.
4. A real-world case study with the Arctic Eider Society demonstrating the practical impact of the proposed system in high-resolution aerial surveys, including substantial reductions in annotation time while maintaining strong detection performance.

Together, these contributions establish a scalable and adaptable framework for wildlife monitoring that bridges advances in foundation models and language-guided detection with the practical requirements of ecological fieldwork.

1.3 Thesis Organization

The remainder of this thesis is organized as follows.

- Chapter 2 provides background on deep learning foundations, object detection, biodiversity monitoring, the use of artificial intelligence in ecological applications, and human-in-the-loop learning paradigms.
- Chapter 3 describes the construction of a unified benchmark for open-world aerial wildlife detection. It details the consolidation of diverse public datasets, the generation of aligned textual descriptions, and the design of dataset splits that explicitly test generalization across species and environments.

- Chapter 4 presents the OpenWildlife model architecture, training strategy, and experimental evaluation. Quantitative results are reported for both fine-tuned and zero-shot settings to assess cross-species generalization.
- Chapter 5 introduces the human-in-the-loop annotation interface and reports on a deployment case study with the Arctic Eider Society. The effectiveness of incremental refinement and its impact on annotation efficiency and counting accuracy are analyzed.
- Chapter 6 concludes the thesis with a summary of findings and directions for future research.

Chapter 2

Background

This chapter introduces the technical and domain-specific concepts that underpin the methods developed in this work. We first review core principles of deep learning, including neural networks, convolutional architectures, attention mechanisms, and transformer-based models, with emphasis on representation learning and open-vocabulary prediction. We then discuss object detection frameworks, training objectives, and evaluation metrics relevant to visual recognition tasks. The chapter next examines the application of artificial intelligence to biodiversity monitoring, focusing on aerial imagery–based wildlife population estimation and the challenges associated with large-scale ecological surveys. Finally, we review human-in-the-loop learning paradigms, which integrate expert feedback into model training and are central to the annotation and refinement workflow explored in this research.

2.1 Deep Learning Foundations

2.1.1 Neural Networks and Representation Learning

Neural networks are parametric function approximators that learn a mapping $f_\theta : \mathcal{X} \to \mathcal{Y}$ from an input space $\mathcal{X}$ (e.g., images) to an output space $\mathcal{Y}$ (e.g., class labels or bounding boxes), where θ denotes a high-dimensional set of learnable parameters [113]. The most basic architecture is the *fully connected network* (FCN), in which every unit in one layer is connected to every unit in the next. For an input vector $x \in \mathbb{R}^d$, a single FCN layer computes

$$h = \sigma(Wx + b), \tag{2.1}$$

where W is a weight matrix, b is a bias vector, and $\sigma(\cdot)$ is a non-linear activation function [63]. By stacking such layers, FCNs can approximate highly complex functions, a property formalized by the universal approximation theorem [40].

A defining characteristic of deep learning is *representation learning* [8]. Rather than relying on handcrafted features, neural networks learn hierarchical feature representations directly from raw inputs. Early layers typically encode low-level structures (e.g., edges, textures, and colour gradients), while deeper layers capture progressively more abstract concepts such as object parts and semantic categories. These learned representations form a latent feature space in which semantically meaningful structure is encoded geometrically.

This representation-learning paradigm naturally extends to settings involving multiple data *modalities*, where a modality refers to a distinct type of input such as images, text, audio, or structured metadata. In multimodal systems, separate encoder networks are typically used to map each modality into a continuous embedding space. An image encoder transforms pixel-level inputs into visual feature vectors, while a text encoder converts tokenized language into semantic embeddings. When trained jointly with *image-text pairs* (e.g. an image of an apple labelled "apple"), these encoders can produce aligned representations, such that semantically related inputs across modalities occupy nearby regions of the latent space [103, 81]. This cross-modal alignment enables tasks such as retrieval, grounding, and open-vocabulary prediction, where supervision or conditioning may originate from a different modality than the primary input [103, 67].

Model parameters are optimized via stochastic gradient descent (SGD) or related variants to minimize a task-specific loss function. However, the high capacity of deep networks introduces the risk of *overfitting*, where a model memorizes training data at the expense of generalization to unseen samples. Regularization strategies such as weight decay, dropout, data augmentation, and early stopping are commonly employed to mitigate this effect [56].

To evaluate generalization performance and mitigate overfitting, datasets are typically partitioned into separate training, validation, and test splits. The training set is used to optimize model parameters, the validation set is used for model selection during development, and the test set provides an unbiased estimate of final model performance on unseen data. Maintaining strict separation between these splits is essential to prevent information leakage and ensure reliable evaluation of learned representations.

2.1.2 Convolutional Neural Networks

While FCNs treat inputs as flat vectors, many data types, particularly images, exhibit strong spatial structure. *Convolutional neural networks* (CNNs) incorporate this struc-

ture directly into the architecture [62, 55, 63]. Instead of connecting every input element to every output element, CNNs restrict connections to local neighbourhoods and share weights across spatial locations. This reduces the number of parameters and encodes the assumption that meaningful visual patterns may appear anywhere in the image.

Let $f \in \mathbb{R}^{H \times W \times C}$ denote an input image with height H, width W, and C channels, and let $g \in \mathbb{R}^{k_h \times k_w \times C}$ denote a learnable convolutional kernel. Using the discrete three-dimensional convolution formulation, the convolution operation is defined as

$$(f * g)[i, j, k] = \sum_{m=0}^{k_h-1} \sum_{n=0}^{k_w-1} \sum_{o=0}^{C-1} f[i-m, j-n, k-o]\, g[m, n, o]. \tag{2.2}$$

In the context of image processing, the channel index is typically fully aggregated by the kernel, producing a two-dimensional output feature map. Intuitively, the kernel slides across the spatial dimensions of the image and computes a weighted sum over a local neighbourhood and all input channels. Since the same kernel weights are reused at every spatial location, the network can detect the same visual pattern regardless of its position in the image.

In practice, convolutional layers employ multiple kernels $\{g^{(1)}, \ldots, g^{(F)}\}$, each producing a distinct output feature map. These responses are typically followed by bias addition and a non-linear activation function, enabling the network to learn increasingly expressive representations.

The locality of the convolutional kernel defines a *receptive field*, which grows with network depth as successive layers compose spatially local operations. Early layers tend to capture low-level patterns such as edges and textures, while deeper layers aggregate increasingly abstract spatial structures. Pooling operations or strided convolutions are often used to reduce spatial resolution while increasing the effective receptive field, enabling hierarchical feature extraction. These inductive biases make CNNs particularly effective for visual recognition tasks.

2.1.3 Attention Mechanisms and Transformers

The progression from FCNs (Section 2.1.1) to CNNs (Section 2.1.2) reflects increasing incorporation of structural assumptions about the data. Transformers take a different approach. Rather than relying on dense connectivity as in FCNs or fixed local receptive fields as in CNNs, they use attention mechanisms to model relationships between elements directly and dynamically.

Self-attention allows every element in a representation to interact with every other element in a single layer. Given input embeddings projected into query (Q), key (K), and value (V) vectors, attention weights are computed as

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V, \tag{2.3}$$

where d_k is the dimensionality of the key vectors [124]. This mechanism enables global, content-dependent information exchange without requiring deep stacking to expand receptive fields.

Transformers are architectures built by stacking layers of self-attention and feed-forward networks [124]. Originally introduced for natural language processing, they have since been adapted to vision by treating image patches as tokens [27, 80]. Compared to CNNs, transformers impose weaker spatial inductive biases and instead rely on attention to learn both local and long-range structure from data. Recurrent neural networks (RNNs) also model dependencies across sequences, but they process inputs sequentially and are less amenable to parallel computation.

Cross-attention extends the attention mechanism to operate between two distinct modalities or feature sets. In this setting, queries are derived from one representation (e.g., text embeddings), while keys and values are derived from another (e.g., visual features) [81]. This mechanism enables multimodal conditioning, such as the use of textual descriptions to guide visual feature selection.

2.1.4 Closed-set and Open-set Prediction

Traditional machine learning systems are typically formulated under a *closed-set* assumption, in which all possible classes are known during training and inference is restricted to this fixed label space [31]. Under this setting, models are optimized to discriminate among predefined categories, and predictions outside this space are implicitly forced into one of the known classes. While effective in controlled benchmarks, this assumption is often unrealistic in real-world applications, where novel categories may appear at deployment time.

Open-set prediction relaxes this constraint by allowing the possibility of unknown classes at inference [114]. In addition to classifying known categories, the model must detect and appropriately handle samples that do not belong to the training distribution. This requires mechanisms for uncertainty estimation, calibrated confidence scoring, or explicit modelling of out-of-distribution examples.

Open-vocabulary prediction extends this idea further by enabling models to generalize to arbitrary categories specified at inference time, often through natural language descriptions [136]. Rather than learning a fixed classifier over discrete labels, models learn a shared embedding space in which visual and textual representations are aligned. Categories are defined by semantic descriptions, allowing flexible expansion beyond the training taxonomy without retraining the model.

As introduced in Section 2.1.1, central to open-vocabulary systems is the use of *text encoders* and *image encoders* trained to produce embeddings in a shared semantic space. Cross-modal alignment is typically achieved using contrastive objectives that encourage matching image–text pairs to have high similarity while separating mismatched pairs. Cosine similarity is commonly used as a normalized similarity metric:

$$\text{sim}(\mathbf{u}, \mathbf{v}) = \frac{\mathbf{u}^\top \mathbf{v}}{\|\mathbf{u}\|\|\mathbf{v}\|}, \tag{2.4}$$

which measures angular proximity between embedding vectors. This formulation facilitates scalable comparison across large vocabularies and enables relevance-based ranking or retrieval [45].

By grounding predictions in a continuous semantic embedding space rather than a fixed set of discrete labels, open-world recognition frameworks provide a flexible foundation for deployment in dynamic environments. Such models can incorporate semantic priors derived from language, adapt to emerging categories, and generalize beyond the constraints of closed-set supervision.

2.1.5 Learning with Limited Supervision

Many real-world machine learning problems involve limited or incomplete supervision, as collecting large, fully labelled datasets can be expensive, time-consuming, or impractical. This limitation arises in two common settings: few-shot or zero-shot learning, where labelled examples are scarce or absent for certain categories, and partial annotation scenarios, where only a subset of relevant instances are labelled.

In *few-shot learning*, the model adapts to new classes using a small number of labelled examples, leveraging transferable representations to achieve accurate predictions with minimal supervision [120]. In the context of wildlife monitoring, this may involve introducing a new species (e.g., a rarely observed marine mammal) using only a handful of annotated aerial images, enabling the model to extend its detection capabilities without large-scale

data collection. *Zero-shot learning* extends this paradigm to cases with no labelled examples, relying on auxiliary semantic information such as attributes or textual descriptions to align inputs with unseen categories in a shared representation space [133]. For example, a model may be prompted to detect a previously unseen species such as a specific seabird or terrestrial mammal based solely on its textual description (e.g., "large white bird with black wingtips"), without any annotated training instances. Both approaches require embeddings that capture semantically meaningful structure, enabling the model to distinguish and relate categories in a way that supports rapid adaptation to novel categories [47].

Incomplete supervision also arises in partially annotated datasets. In large or densely populated images, annotators may label only a subset of objects while leaving the remainder unannotated. Naïvely treating unlabelled regions as background during training can introduce incorrect supervisory signals, biasing the model and suppressing recall. Selective or masked training strategies address this issue by restricting loss computation to annotated regions or explicitly masking uncertain areas. These approaches are related to partial label learning and semi-supervised learning frameworks [79, 7], and in practice, masked loss functions allow models to benefit from corrected predictions without making assumptions about unlabelled regions.

2.1.6 Foundation Models

In recent years, deep learning has increasingly shifted toward large-scale pre-training on diverse and heterogeneous data. Models trained on massive corpora of images, text, audio, or structured signals can learn general-purpose representations that capture broad statistical regularities across domains. Rather than optimizing exclusively for a single downstream objective, these models are typically trained with self-supervised or weakly supervised objectives that encourage the emergence of reusable features.

This paradigm has led to the development of so-called *foundation models* [9]. A foundation model is trained once at large scale and subsequently adapted to many different tasks. Adaptation may take the form of full fine-tuning, parameter-efficient updates, prompt-based conditioning, or zero-shot inference. The central idea is that a sufficiently expressive model, exposed to sufficiently diverse data, can acquire representations that generalize beyond its original training distribution.

Foundation models now span multiple modalities. In natural language processing, large language models demonstrate strong few-shot and zero-shot capabilities. In computer vision, large-scale image or image–text pre-training has enabled open-vocabulary recognition and cross-modal retrieval. In scientific domains, models such as AlphaFold [46] illustrate

how large-scale training on structured data can yield highly accurate predictive systems. Across settings, the appeal of foundation models lies in their flexibility, transferability, and the reduced need for task-specific model design.

Despite these advantages, foundation models introduce important trade-offs. Their behaviour reflects biases and gaps in the data used for pre-training, and their scale can make them computationally expensive and difficult to interpret. Transfer performance depends on the alignment between the pre-training distribution and the target task, and mismatches can lead to degraded accuracy. Careful evaluation and adaptation are therefore essential when deploying such models in specialized domains.

2.1.7 Object Detection

Object detection defines the mapping introduced in Section 2.1.1 as predicting both the category and location of multiple objects within an input image. Formally, given an input image $x \in \mathcal{X}$, the model predicts a set of outputs $y \in \mathcal{Y}$, where each element of y consists of a class label and corresponding spatial coordinates (e.g., bounding boxes) for a detected object [75].

Detection Frameworks

Detection frameworks can be categorized into *anchor-based* and *anchor-free* paradigms. Anchor-based detectors define a fixed set of candidate boxes at multiple scales and aspect ratios, which are then refined to match objects in the scene [105, 77]. Anchor-free methods, in contrast, directly predict object locations and dimensions without relying on predefined priors, often improving efficiency and simplifying design choices [121, 142].

Another key distinction is between *dense prediction* and *set-based prediction*. Dense detectors produce predictions at every spatial location in the feature map, followed by non-maximum suppression (NMS) to remove duplicate detections [95]. NMS is a post-processing algorithm that iteratively selects the highest-confidence detection and suppresses overlapping predictions whose Intersection over Union (IoU) exceeds a predefined threshold. This procedure ensures that multiple predictions corresponding to the same object are consolidated into a single final detection. Set-based detectors, such as DETR [14], reformulate detection as a direct set prediction problem, leveraging learnable object queries to produce a fixed number of detections and removing the need for post-processing heuristics such as NMS.

As discussed in Section 2.1.3, transformer-based object detection integrates attention mechanisms to model global context and long-range dependencies. Deformable DETR further incorporates multi-scale reasoning by sampling features at adaptively selected spatial locations, improving efficiency and accuracy on small objects [143]. Deformable attention allows each object query to attend to a sparse set of informative features across scales, reducing computational cost while preserving performance.

Object queries serve as learnable detection slots that interact with image features via cross-attention. The initialization of these queries can impact performance: standard random initialization may be inefficient in sparse scenes, whereas semantic priors or learned embeddings can improve convergence and detection quality [76].

Detection Objectives and Losses

Object detection models are trained using a combination of classification and localization objectives that reflect the set-based prediction formulation. To ensure a one-to-one correspondence between predicted and ground-truth objects, *bipartite (Hungarian) matching* is performed. Let $\{\hat{y}_i\}_{i=1}^{N}$ denote the set of predicted objects and $\{y_j\}_{j=1}^{M}$ the set of ground-truth objects. The Hungarian algorithm finds the optimal assignment σ that minimizes the total matching cost:

$$\sigma^* = \arg\min_{\sigma \in \mathfrak{S}_N} \sum_{i=1}^{N} \mathcal{L}_{\text{match}}(\hat{y}_i, y_{\sigma(i)}), \tag{2.5}$$

where $\mathfrak{S}_N$ denotes the set of all permutations over N elements, and $\mathcal{L}_{\text{match}}$ combines classification and localization costs [14]. Unmatched predictions are assigned to a special background class.

Classification losses measure the discrepancy between predicted class probabilities $\hat{p}_c$ and the ground-truth class. In the standard formulation, the ground-truth label is represented as a one-hot vector, meaning $p_c = 1$ for the correct class and $p_c = 0$ otherwise. The categorical cross-entropy loss is defined as

$$\mathcal{L}_{\text{cls}} = -\sum_{c=1}^{C} p_c \log \hat{p}_c, \tag{2.6}$$

which, under one-hot encoding, reduces to the negative log-probability of the correct class. Confident but incorrect predictions are therefore penalized more strongly than uncertain ones.

In many detection settings, the number of background examples far exceeds foreground objects, leading to class imbalance. *Focal loss* addresses this issue by down-weighting well-classified examples and focusing training on harder, misclassified instances:

$$\mathcal{L}_{\text{focal}} = -\sum_{c=1}^{C}(1-\hat{p}_c)^{\gamma} p_c \log \hat{p}_c, \tag{2.7}$$

where γ controls how strongly easy examples are suppressed. When $\hat{p}_c$ is close to 1 (i.e., correctly predicted with high confidence), the modulating factor $(1-\hat{p}_c)^{\gamma}$ becomes small, reducing its contribution to the loss [71]. This mechanism helps stabilize training in dense detection scenarios.

Localization losses quantify the accuracy of predicted bounding boxes $\hat{b}$ relative to ground truth b. A bounding box is typically parameterized by its center coordinates, width, and height. The L1 loss measures the absolute difference between predicted and true box parameters:

$$\mathcal{L}_{\text{L1}} = \sum_{i} |\hat{b}_i - b_i|, \tag{2.8}$$

encouraging precise coordinate regression.

Beyond coordinate differences, overlap-based metrics capture spatial alignment more directly. The Intersection over Union (IoU) between two boxes is defined as

$$\text{IoU}(\hat{b}, b) = \frac{|\hat{b} \cap b|}{|\hat{b} \cup b|}, \tag{2.9}$$

where $\cap$ denotes the intersection area between the predicted and ground-truth boxes, and $\cup$ denotes their union area. IoU ranges from 0 (no overlap) to 1 (perfect overlap). The Generalized IoU (GIoU) loss extends this by considering the smallest enclosing box C that contains both $\hat{b}$ and b:

$$\mathcal{L}_{\text{GIoU}} = 1 - \text{IoU}(\hat{b}, b) + \frac{|C \setminus (\hat{b} \cup b)|}{|C|}, \tag{2.10}$$

where $C \setminus (\hat{b} \cup b)$ represents the area inside C but outside the union of the two boxes [107]. This additional term penalizes non-overlapping boxes and provides meaningful gradients even when IoU is zero.

For text-conditioned or open-vocabulary detectors, *contrastive or relevance-based objectives* align visual and textual embeddings. Given image features f_v and a corresponding

text embedding f_t, a typical contrastive loss is:

$$\mathcal{L}_{\text{contrastive}} = -\log \frac{\exp(\text{sim}(f_v, f_t)/\tau)}{\sum_k \exp(\text{sim}(f_v, f_{t_k})/\tau)}, \tag{2.11}$$

where $\{f_{t_k}\}$ denotes the set of text embeddings in the current batch (including both the matched description and mismatched alternatives), $\text{sim}(\cdot, \cdot)$ is cosine similarity (see Equation (2.4)), and τ is a temperature scaling factor. Minimizing this loss increases similarity between matched image–text pairs while decreasing similarity to other text embeddings in the batch, thereby organizing both modalities in a shared embedding space.

Detection Metrics

Evaluation of object detection involves metrics that assess both classification and localization performance. *Precision* (P) measures the fraction of predicted objects that are correct, *recall* (R) measures the fraction of ground-truth objects detected, and the F1-score combines both:

$$P = \frac{\text{TP}}{\text{TP} + \text{FP}}, \quad R = \frac{\text{TP}}{\text{TP} + \text{FN}}, \quad F1 = 2\frac{P \cdot R}{P + R}, \tag{2.12}$$

where TP, FP, and FN denote true positives, false positives, and false negatives, respectively.

Mean Average Precision (mAP) summarizes precision-recall performance across multiple thresholds and classes. For a single class, average precision (AP) is the area under the precision-recall curve, and mAP is the mean across classes:

$$\text{mAP} = \frac{1}{C} \sum_{c=1}^{C} \text{AP}_c, \tag{2.13}$$

where mAP_{50} refers to mAP computed at an IoU threshold of 0.5.

For counting tasks, errors are measured using metrics such as mean absolute error (MAE):

$$\text{MAE} = \frac{1}{N} \sum_{i=1}^{N} |\hat{y}_i - y_i|, \tag{2.14}$$

root mean squared error (RMSE):

$$\text{RMSE} = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (\hat{y}_i - y_i)^2}, \tag{2.15}$$

and percentage error:

$$\%\text{error} = \frac{1}{N}\sum_{i=1}^{N}\frac{|\hat{y}_i - y_i|}{y_i} \cdot 100. \tag{2.16}$$

COCO Annotation Format

Although this work does not directly train or evaluate on the COCO dataset, we adopt the *COCO annotation format* as a standardized representation for object detection. The COCO format has become the de facto interchange standard for detection datasets and is widely supported by modern training pipelines and evaluation toolkits [72].

In COCO-style detection annotations, labels are stored in a structured JSON file containing three primary components:

- **Images**: metadata for each image, including a unique `id`, file name, width, and height.
- **Annotations**: object-level entries containing an `id`, the corresponding `image_id`, a `category_id`, and geometric information describing the object location.
- **Categories**: a list of class definitions, each with a unique `id` and name.

For object detection, bounding boxes are represented using an axis-aligned format:

$$b = (x_{\min}, y_{\min}, w, h), \tag{2.17}$$

where $(x_{\min}, y_{\min})$ denotes the top-left corner of the box in pixel coordinates, and w and h represent its width and height. All coordinates are defined in the image coordinate system, with the origin at the top-left corner. An optional `area` field records the box area in pixels, and the `iscrowd` flag indicates whether the annotation corresponds to a crowd region.

Standardizing annotations in COCO format enables uniform data loading, batching, augmentation, and evaluation across datasets. More importantly, it ensures compatibility with transformer-based, set-prediction detectors that assume per-instance bounding box supervision in a structured annotation schema.

2.2 AI for Biodiversity Monitoring

2.2.1 Biodiversity Monitoring

Biodiversity monitoring is essential for understanding ecosystem stability, resilience, and long-term sustainability. Biodiversity underpins ecosystem services such as pollination, carbon sequestration, nutrient cycling, water purification, and climate regulation, and declines in species richness or functional diversity are often early indicators of ecosystem stress. Effective monitoring considers multiple signals including species richness and evenness, genetic variation, phenological changes such as shifts in migration or breeding timing, behavioural dynamics, and habitat quality. Together, these indicators provide a holistic view of ecosystem health and inform conservation policy, land management, and environmental governance [73].

Historically, these signals were collected through labour-intensive methods such as ground transects, point counts, and observational surveys, which require trained personnel and are limited in spatial coverage, temporal resolution, and reproducibility [61, 30]. Technological advances in remote sensing and automated data acquisition have expanded monitoring capabilities. High-resolution aerial and satellite imagery, acoustic sensors, camera traps, and environmental DNA sampling enable the capture of species presence, behaviours, and habitat conditions over broad spatial and temporal scales. Deep learning and other AI techniques now allow the extraction of meaningful ecological metrics from these heterogeneous datasets, improving consistency, efficiency, and the ability to detect early ecological changes that might otherwise go unnoticed.

The foundation model paradigm described in Section 2.1.6 has begun to influence ecological research and biodiversity monitoring. Large-scale pre-training on image–text data enables models to associate visual patterns with taxonomic and ecological concepts, supporting tasks such as species recognition and open-vocabulary classification (see Section 2.1.1 and Section 2.1.4). Examples of image-text data include a camera trap image broadly labelled "white-tailed deer" or an aerial survey image annotated with instances of "beluga whale". Multimodal models trained on biological image corpora have demonstrated fine-grained taxonomic classification across diverse organism groups [119].

Transfer learning plays a central role in these applications. Pre-trained backbones can be adapted to domain-specific datasets, reducing the amount of labelled ecological data required for training. However, ecological data often exhibit substantial domain shift due to variations in sensor types, environmental conditions, geographic regions, and species distributions. When the pre-training data underrepresent certain habitats or taxa, performance may degrade, necessitating targeted fine-tuning or domain adaptation strategies.

Large language models can also support ecological workflows by generating image captions, synthesizing metadata, assisting with literature review, or extracting structured information from reports [34, 1]. When integrated into multimodal pipelines, they can provide contextual knowledge that complements visual models. At the same time, automatically generated annotations may contain inaccuracies or hallucinated details, and error propagation can occur if such outputs are used as training data without verification.

2.2.2 Population Counting from Aerial Imagery

Motivation and Context

Population counts provide a direct and quantifiable measure of biodiversity health, often reflecting species abundance, reproductive success, and ecosystem stability. Such counts can be obtained from direct field observations, harvest records, motion-sensitive camera traps, or signs of species presence, but these approaches are generally limited in spatial extent, observer reliability, and temporal coverage. Aerial imagery, captured via manned aircraft, unmanned aerial vehicles (UAVs), or increasingly, commercial satellites, provides a scalable alternative capable of monitoring extensive and often remote habitats with minimal disturbance to wildlife [118, 66, 25]. Crewed aircraft offer flexible coverage for large survey areas, UAVs provide high-resolution imagery at low altitudes and reduced operational risk, and satellite sensors, though historically constrained by spatial resolution, are now emerging as viable tools for on-demand, large-scale population monitoring.

The traditional protocol for large-scale aerial wildlife surveys has been the systematic reconnaissance flight (SRF), in which trained rear-seat observers count animals within predefined transect strips during flight [61, 30]. Because animals are visible only briefly as the aircraft passes overhead, this method is susceptible to perception bias and underestimation [15]. To mitigate these limitations, many surveys now rely on retrospective annotation of captured aerial imagery. In this workflow, high-resolution photographs are reviewed postflight and manually annotated to enumerate individuals [25, 20]. Retrospective counting improves accuracy relative to real-time observation by allowing zooming, contrast adjustment, and repeated inspection. However, manual annotation remains labor-intensive, slow, and cognitively demanding. Annotation fatigue, inconsistency across observers, and ambiguity in complex scenes introduce additional sources of error [65, 60, 11]. As survey volumes increase with higher-resolution sensors and broader coverage, fully manual workflows become increasingly impractical.

These limitations directly motivate the integration of AI-based automation. Despite its promise, extracting reliable counts from aerial imagery presents significant technical

challenges. Individuals may occupy only a few pixels, making detection sensitive to ground sampling distance and image resolution, while variations in altitude, sensor characteristics, and imaging conditions introduce additional scale variability. Wildlife can occur in dense aggregations, causing severe occlusion and overlap that complicate instance separation. Complex natural backgrounds, camouflage, seasonal changes, and heterogeneous terrain further reduce visual contrast. Multi-species scenes require fine-grained discrimination, and class imbalance or long-tailed distributions can bias model training. Domain shifts caused by geographic, climatic, or sensor variation further hinder generalization [98, 44, 24].

AI-Based Counting Strategies

Modern AI methods attempt to address these challenges through different representational strategies. Encoder–decoder architectures, including U-Net variants, learn hierarchical spatial features that enable pixel-level localization of animals in sparse or heterogeneous environments [110, 132, 35, 32]. Detection-based frameworks predict bounding boxes around individual instances, enabling explicit localization, species classification, and post-hoc verification of results. The availability of bounding boxes is particularly important in conservation contexts, where transparency, auditability, and expert validation are often required for policy decisions or population trend reporting.

In contrast, density-based counting architectures estimate a continuous density map over the image rather than predicting discrete object instances. Typically, a convolutional neural network is trained to regress from the input image to a spatial density surface, where integrating over the density map yields the total count [24, 101]. Ground truth supervision is commonly generated by convolving point annotations with a Gaussian kernel, allowing the network to learn spatially distributed representations of object presence. Density-based methods are particularly effective in scenes with tightly packed aggregations, where explicit bounding box detection becomes unreliable due to extreme overlap or limited resolution. However, because density maps do not inherently produce discrete instance-level predictions, they may limit interpretability and direct visual verification compared to detection-based approaches.

Transformer-based frameworks, such as WildlifeMapper, further enhance detection performance by leveraging attention mechanisms to model long-range dependencies and emphasize salient features against complex natural backgrounds [58]. Across these approaches, automated analysis has demonstrated improvements in speed, repeatability, and scalability compared to manual counting [123, 38, 96, 52]. Nevertheless, models often remain specialized to specific species, imaging platforms, or ecological contexts, and transferring

performance to new environments continues to require careful adaptation and, in many cases, human-in-the-loop refinement [106].

Overall, the evolution from real-time aerial observation to retrospective manual annotation, and now toward AI-assisted analysis, reflects a progression toward greater scalability and reproducibility in wildlife monitoring. While automation addresses many limitations of human counting, balancing accuracy, interpretability, and generalization remains a central challenge in deploying AI systems for operational biodiversity assessment.

2.3 Human-in-the-Loop Learning

Human-in-the-loop (HITL) machine learning integrates human expertise directly into model training and refinement, creating iterative feedback cycles between prediction, correction, and adaptation. Rather than separating annotation and model development into sequential phases, HITL systems allow annotators to validate and correct model outputs during labelling, enabling incremental improvement as new data are collected. This paradigm is particularly well suited to ecological monitoring, where expert knowledge is essential, annotation is costly, and data distributions shift across seasons, habitats, and imaging platforms.

HITL approaches address limitations of traditional fully supervised pipelines, in which large datasets must be exhaustively labelled before training begins. Interactive systems instead leverage model pre-labelling to reduce manual effort, using human corrections to refine parameters over time. This strategy aligns with active learning and interactive machine learning frameworks that seek to maximize model improvement per unit of human effort [115, 3, 57]. In biodiversity monitoring, where expert time is often the primary bottleneck, such efficiency gains are critical for scalability.

Iterative updating introduces theoretical and practical challenges. A central concern is the *stability–plasticity dilemma*, which describes the tension between adapting to new data and retaining previously acquired knowledge [89]. Naïve fine-tuning on newly labelled samples can lead to *catastrophic forgetting*, degrading performance on earlier species, regions, or survey conditions [88]. In long-term monitoring programs, where models must generalize across years and ecological contexts, controlled retraining and evaluation checkpoints are therefore essential.

Operational deployment further requires careful integration with annotation infrastructure. Wildlife monitoring projects often involve distributed expert teams, limited funding, and irregular data acquisition schedules. Annotation platforms such as LabelStudio

[122], CVAT [21], and Roboflow [28] support collaborative labelling and model-assisted pre-annotation, but typically assume static training cycles. Embedding incremental retraining directly within annotation workflows enables continuous improvement, reduces redundant corrections over time, and increases annotation throughput.

In ecological applications, the objective of human-in-the-loop learning extends beyond predictive accuracy to include interpretability and long-term robustness under domain shift. By integrating expert feedback into controlled iterative refinement, HITL systems provide a pathway toward adaptive, trustworthy, and scalable biodiversity monitoring pipelines that bridge machine learning research and operational conservation practice.

Chapter 3

Dataset

3.1 Overview

To evaluate open-world aerial wildlife detection, we construct a unified benchmark by consolidating diverse public datasets, generating aligned textual descriptions, and designing dataset splits that explicitly test generalization across species and environments.

3.2 Cross-Dataset Consolidation

We consolidated 24 publicly available aerial wildlife datasets covering multiple species, ecosystems, and imaging conditions, all consisting of RGB imagery. Figure 3.2 shows example images, and Figure 3.1 illustrates their global distribution. All annotations were converted to COCO format (Section 2.1.7) with bounding boxes; datasets originally providing only keypoints were adapted by generating fixed-size boxes centred on annotated locations, with sizes based on the species' average spatial extent.

Datasets were grouped according to their role in model development and evaluation. 14 datasets were used for open-set continuous pre-training (Table A.1), seven datasets containing annotation categories not observed during training were reserved for evaluating generalization to unseen categories (Table A.2), and three datasets containing previously observed labels but acquired under distinct geographic regions or environmental conditions were used to evaluate robustness to domain shift (Table A.3).

The resulting label- and environment-aware partitioning is summarized in Table 3.1. In total, the training split comprises 23,408 images with 810,492 annotations spanning 15

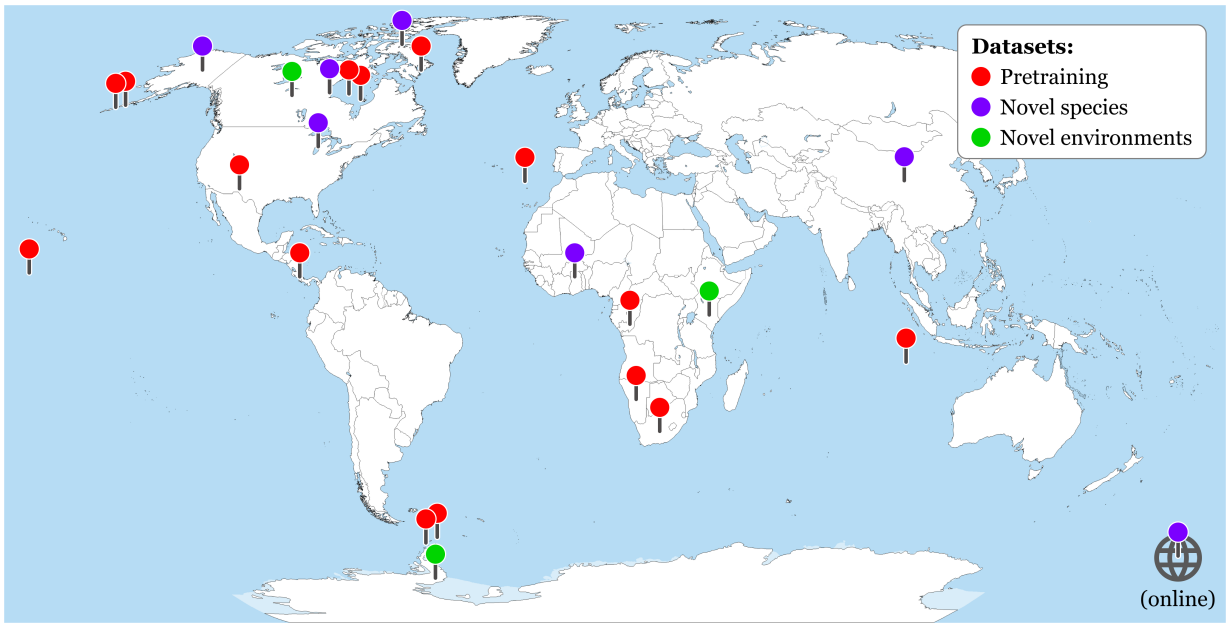


Figure 3.1: Global distribution of the aerial wildlife datasets used in this work. Pins indicate the approximate geographic regions where imagery was collected. Red denotes datasets used for open-set pre-training, purple indicates datasets containing novel species, and green marks datasets used to assess generalization to novel environments.

distinct annotation labels. These labels are inherited directly from the source datasets and are not uniformly defined at the species level; they include a mixture of species-level designations (e.g., beluga whale), higher-level taxonomic groupings (e.g., bird, penguin), and coarse ecological categories (e.g., other). Remapping these labels to a consistent taxonomic level was not feasible, as some labels (e.g., bird) aggregate multiple species that cannot be distinguished from the annotations alone. Consequently, "annotation label" or "detection category" is used throughout this work rather than "species," as the latter would imply a taxonomic precision that the data does not consistently support.

The novel category test split contains 16 annotation labels absent from the training set at the label level, enabling evaluation of generalization to unseen detection categories without additional supervision. Due to the coarse-grained nature of some labels, semantic or visual overlap may still exist at higher taxonomic levels; this split therefore evaluates generalization to new annotation categories within related visual domains, rather than to entirely disjoint biological classes.

The novel environment split contains imagery associated with 7 annotation labels, of which 3 (beluga whale, elephant, penguin) are present in the training set and 4 (bowhead whale, giraffe, narwhal, zebra) are not. This split is not designed to guarantee label-level novelty; rather, it is constructed to isolate the effect of geographic and environmental domain shift by holding out imagery captured under distinct conditions from those seen during training.

The two evaluation splits are constructed independently and are not required to be mutually exclusive in terms of semantic content: while some labels or closely related categories may appear across both test sets, no test data is used during training. This design enables controlled analysis of two complementary generalization challenges—(i) transferring to previously unseen annotation categories and (ii) maintaining performance under environmental variation—while explicitly acknowledging that these factors cannot be perfectly disentangled given the heterogeneity of real-world ecological datasets.

	Original annotation labels	# of annotated images	# of annotations
Train	Alcelaphinae, Beluga whale, Bird, Brant, Buffalo, Canada goose, Elephant, Emperor goose, Kob, Other, Penguin, Sea lion, Turtle, Warthog, Waterbuck	23,408	810,492
Test (novel species)	Bearded seal, Camel, Caspian tern, Cattle, Great cormorant, Great white pelican, Gray headed gull, Gull, Kiang, Narwhal, Polar bear, Ringed seal, Royal tern, Sheep, Slender-billed gull, Yak, Zebra	18,606	283,151
Test (novel environments)	Beluga whale, Bowhead whale, Elephant, Giraffe, Narwhal, Penguin, Zebra	708	5,307

Table 3.1: Dataset partitioning for open-set continuous pre-training and evaluation. Labels are inherited directly from source datasets and reflect mixed taxonomic granularity (see Appendix A). The test (novel species) set contains 16 annotation labels absent from the training set. The test (novel environments) set contains 7 labels, of which 3 overlap with training labels; this split isolates environmental and geographic domain shift rather than label-level novelty.

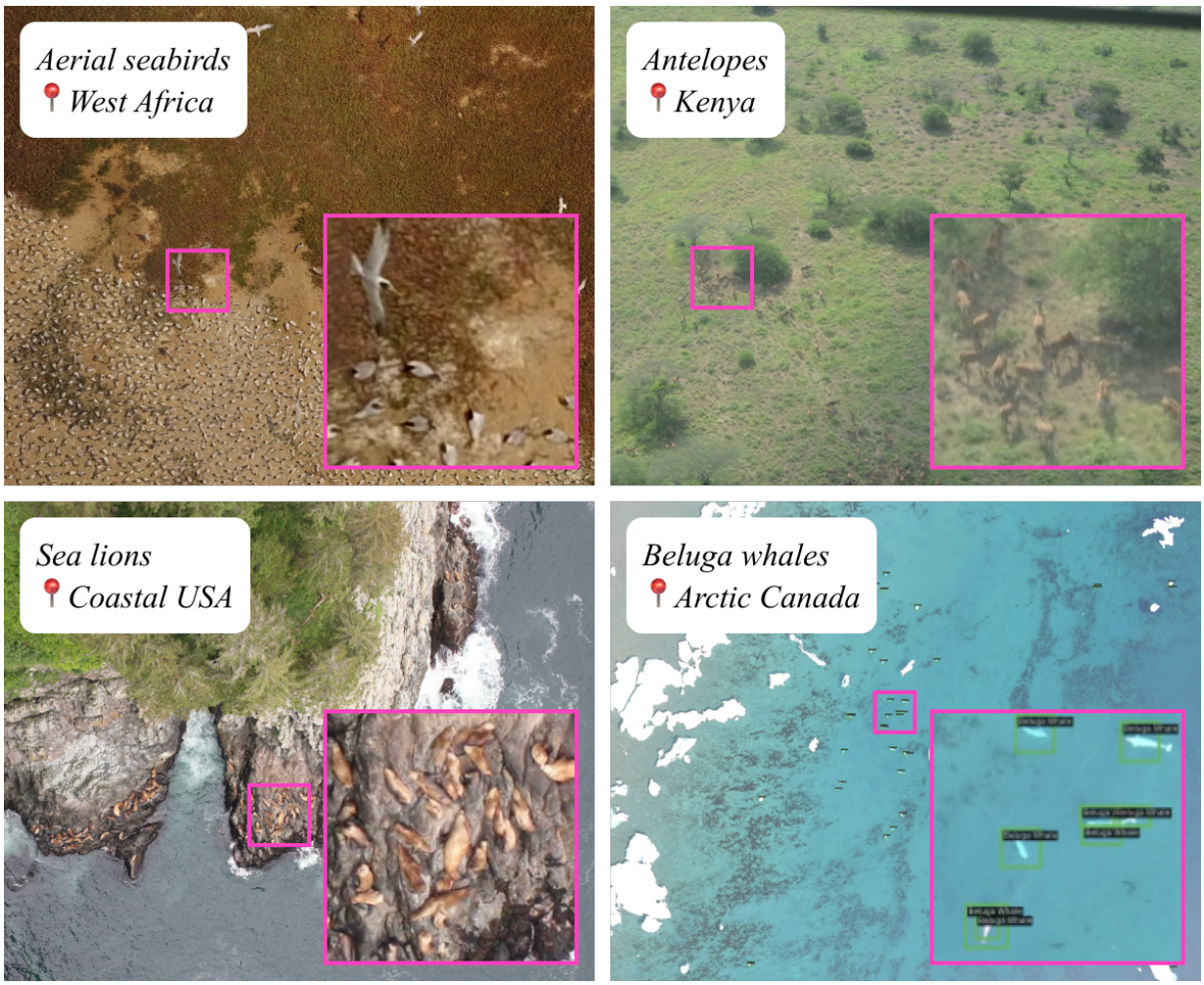


Figure 3.2: Example imagery from the consolidated dataset [53, 29, 48], showcasing the variety of environments and species that OpenWildlife was trained on.

Chapter 4

OpenWildlife Model

4.1 Overview

This chapter presents the OpenWildlife (OW) model, a framework for open-vocabulary wildlife detection in aerial imagery. We detail the model's architecture, highlighting its adaptation of MM-Grounding-DINO [141] to the ecological domain, including caption generation, cross-modal feature extraction and fusion, language-guided query selection, and iterative cross-modality decoding. The chapter also describes the experimental protocol, including training procedures, dataset preparation, and captioning strategies, followed by a comprehensive evaluation of OW's performance. Comparisons with closed-set state-of-the-art detectors and literature benchmarks demonstrate the benefits of open-vocabulary text guidance and aerial-domain pre-training. Finally, we analyze detection results on novel species and in novel environments, and discuss limitations observed in certain failure cases.

4.2 Architecture

Figure 4.1 illustrates the architecture of the OpenWildlife model, which builds upon MM-Grounding-DINO [141] to enable open-vocabulary wildlife detection in aerial imagery. Within the taxonomy introduced in Section 2.1.7, OpenWildlife follows a transformer-based, anchor-free, set-based detection paradigm derived from DETR-style architectures. Object queries interact with multi-scale image features through deformable attention, pro-

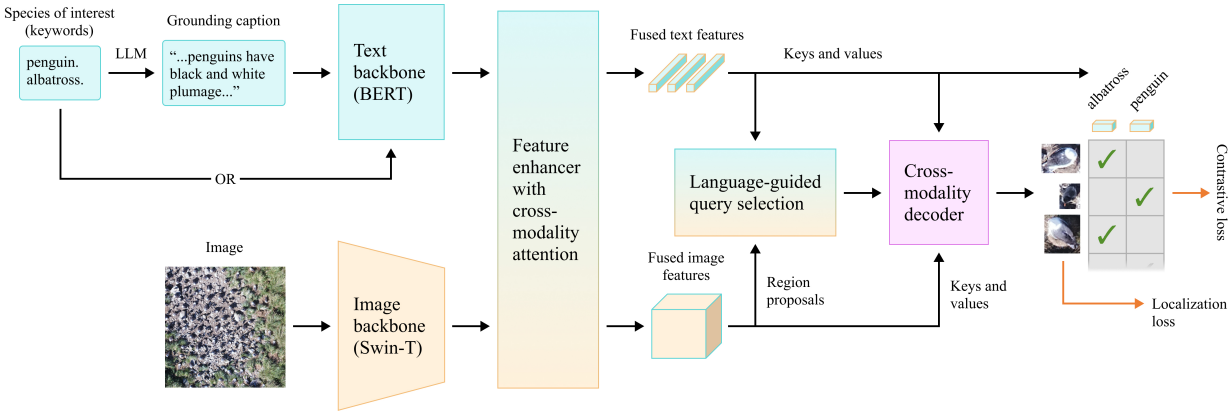


Figure 4.1: In the OpenWildlife architecture, given an aerial image and either a list of species of interest or an LLM-generated grounding caption, the model extracts image and text features. These features are fused through cross-modality enhancement, enabling language-guided queries to highlight relevant regions in both modalities.

ducing a fixed set of detections without requiring dense prediction or post-processing steps such as non-maximum suppression.

The design explicitly targets challenges unique to aerial ecological surveys, including extreme object sparsity or density, small object size, and the absence of semantically aligned image–caption datasets. The model comprises five main components: caption generation, feature extraction and fusion, language-guided query selection, cross-modality decoding, and training losses.

4.2.1 Caption Generation

A fundamental limitation in applying grounding-based detectors to aerial wildlife imagery is the lack of semantically rich, domain-aligned image–caption datasets. While MM-Grounding-DINO [141] is pre-trained on large-scale datasets such as Objects365 [117] and V3Det [125], these sources do not capture the visual scale, viewpoints, or ecological semantics present in aerial surveys.

To bridge this gap, OpenWildlife optionally leverages large language models (LLMs) to generate grounding captions tailored to aerial wildlife imagery. Each caption explicitly enumerates the species present, followed by concise visual descriptors constrained to features observable from high-altitude platforms (e.g., colour, body shape, group structure).

These captions act as semantic priors that guide localization and classification through the text encoder. Further details on caption generation are discussed in Section 4.3.2.

4.2.2 Feature Extraction and Fusion

Given an image and caption text pair (see Sections 2.1.1 and 2.1.4), OpenWildlife extracts modality-specific representations using independent backbones. Visual features are obtained using a Swin Transformer (Swin-T) [80], chosen for its ability to efficiently capture multi-scale image features, which is critical for aerial wildlife detection. Unlike the original Vision Transformer (ViT) [27], which processes images at a fixed patch size and relies on global self-attention, Swin-T employs hierarchical feature maps with shifted window attention. This design enables the model to maintain both fine-grained spatial resolution for small, sparse objects (e.g., individual birds or seals) and broader contextual information for large or clustered animal groups, while keeping computational cost tractable for high-resolution aerial imagery.

Textual inputs are encoded using a BERT-based language encoder [26], which generates contextualized embeddings of species names and descriptive phrases. BERT was selected over simpler embedding methods (e.g., word2vec [90] or GloVe [100]) because its bidirectional transformer architecture captures semantic relationships in context, allowing the model to interpret descriptive captions that reference visual attributes, relative size, and grouping of animals. This contextual understanding is essential for aligning language with visual features in aerial images, particularly when captions contain nuanced descriptors.

The two modalities are fused using a bi-directional, attention-based Feature Enhancer. Cross-attention layers allow text features to attend to spatial image regions and vice versa, enabling alignment between linguistic semantics and visual patterns. These fused representations are further refined using self-attention: standard multi-head self-attention for text, and deformable attention [143] for visual features. This design is critical for aerial imagery, where targets exhibit large scale variation and are frequently embedded in heterogeneous backgrounds.

4.2.3 Language-Guided Query Selection

Unlike conventional object detectors that rely on fixed or randomly initialized queries, OpenWildlife employs language-guided query selection to efficiently localize candidate objects. Specifically, cosine similarity is computed between visual feature vectors and species-

specific text embeddings, producing a relevance map that highlights regions most consistent with the provided textual description.

To accommodate high-density scenarios common in wildlife colonies, we increase the number of decoder queries from the 900 used in standard Grounding-DINO to 2,000. High-similarity regions are selected as candidate object locations, where the positional component of each query acts as a dynamic anchor box and the content component is initialized to zero. This initialization strategy biases the decoder toward semantically relevant regions while preserving flexibility for subsequent refinement.

4.2.4 Cross-Modality Decoder

The selected queries are refined using a cross-modality decoder similar to MM-Grounding-DINO [141]. Each decoder layer consists of self-attention over object queries, followed by image cross-attention and text cross-attention, and a feedforward network (FFN). The inclusion of an explicit text cross-attention block at every decoder layer distinguishes this architecture from standard DINO and ensures that species-level semantic information directly influences bounding box refinement.

Through iterative decoding, object queries converge to precise spatial locations while maintaining alignment with their corresponding textual descriptions. Each refined query produces both a bounding box prediction and an associated class embedding in the shared vision–language space.

4.2.5 Training Losses

Training is formulated as a set-based prediction problem following DETR [14]. Bounding box regression is optimized using a combination of L1 loss and Generalized IoU (GIoU) loss [107], while classification is performed using a focal-based contrastive loss [71]. Auxiliary decoder-layer losses and a bipartite matching loss associate predicted queries with ground-truth annotations and prevent duplicate detections. The total loss is defined as

$$\mathcal{L}_{total} = \lambda_{cls}\,\mathcal{L}_{cls} + \lambda_{L1}\,\mathcal{L}_{L1} + \lambda_{GIoU}\,\mathcal{L}_{GIoU}, \tag{4.1}$$

where the weights λ_{cls}, λ_{L1}, and λ_{GIoU} are tunable.

4.3 Experimental Protocol

4.3.1 Implementation and Training

Our OpenWildlife (OW) model was initialized with publicly available MM-GDINO-T weights [141]. Due to the high resolution of aerial imagery, training images were pre-sliced into 1024×1024 pixel patches. At inference time, larger images are processed using a sliding-window strategy over overlapping patches. Although set-based detectors do not require non-maximum suppression (NMS) to remove duplicate predictions within a single image, NMS is applied to consolidate overlapping detections produced across adjacent inference windows (Section 2.1.7). All evaluation metrics are computed on the reconstructed full-scale images.

Continuous pre-training was conducted on 14 aerial datasets (Section 3.2) using standard train, validation, and test splits. A single model was trained for 20 epochs with a global batch size of 16 across 8 NVIDIA A100-40GB GPUs on 2 nodes, requiring approximately 20 hours. Key training hyperparameters are summarized in Table 4.1.

Hyperparameter	**Value**
Optimizer	AdamW
Learning Rate	4e-05
Warm-up Steps	1000
Max Epochs	20
Batch Size	16 (global)
Data Augmentations	Affine, Brightness, Contrast, RGBShift, HSVShift, JPEG compression, Blur

Table 4.1: Key training hyperparameters for OW model pre-training and fine-tuning.

The weights λ_{cls}, λ_{L1}, and λ_{GIoU} used in Section 4.2.5, together with the Hungarian matching costs, are summarized in Table 4.2.

Stage	$(\lambda_{cls}, \lambda_{L1}, \lambda_{GIoU})$
Training	$(1.0, 5.0, 2.0)$
Hungarian Matching	$(2.0, 5.0, 2.0)$

Table 4.2: Training loss weights and Hungarian matching cost settings for the OW model. Each triplet corresponds to the classification, L1, and GIoU components, respectively.

4.3.2 Caption Generation

We trained two versions of the OpenWildlife (OW) model that differ in how textual supervision is constructed: *OW Sentence*, which uses full-sentence captions, and *OW Keyword*, which uses only species keyword lists.

For OW Sentence, we generated one caption per image using OpenAI's GPT-4o-mini model via the Batch API [97]. Each caption is aligned with the ground-truth bounding box labels. The caption begins with a list of species corresponding exactly to the provided labels, followed by a short visual description derived from the fine-grained species annotations in the source dataset. Descriptions emphasize coarse, aircraft-visible attributes such as colour, size, and body shape, while avoiding mention of species names beyond those listed in the labels.

The prompt used for caption generation was:

> *You are an annotator generating captions of an aerial image. The caption should start with the species present, following exactly the LABELS, followed by a very short description of their appearance based on the TRUE SPECIES. The TRUE SPECIES should not be mentioned, only their descriptions. If LABELS is "NO INSTANCES," only say "No animals present." Only describe details visible from very far away.*

Keyword captions consist solely of the species labels used as text prompts, while sentence captions augment these labels with coarse visual descriptions derived from fine-grained annotations. Representative examples are shown in Table 4.3.

4.3.3 Comparisons with Closed-Set SOTA and Literature

We report two types of comparisons:

Closed-Set SOTA Comparison We also compare OW against state-of-the-art (SOTA) closed-set detectors (DINO [137], CO-DETR [144], and YOLOv8 [135]) pre-trained on the same 14 aerial datasets. This comparison was performed under zero-shot (Section 2.1.5) and fine-tuning conditions. Since closed-set models cannot predict unseen classes, we adopt a localization-based evaluation: a prediction counts as a true positive if the object is correctly localized regardless of class label. These results quantify the benefit of open-vocabulary text guidance and aerial-domain pre-training in zero-shot scenarios.

Dataset	Caption Type	
	Keyword	**Sentence**
AED [29]	Elephant	Species that may be present: elephant. The species may be large, with a grayish-brown skin that appears rough and wrinkled. They have large ears shaped like fans and long trunks.
DFOW14 [85]	Beluga whale	Species that may be present: beluga whales. Belugas are characterized by their predominantly white coloration, which can appear almost ivory in bright lighting. They have rounded heads and a lack of dorsal fin, giving them a distinctive streamlined shape.
NOAA [48]	Sea lion	Species that may be present: sea lion. The species may have a robust body with light brown fur and prominent foreflippers, forming a compact silhouette against coastal backgrounds.

Table 4.3: Sample keyword and sentence captions for representative datasets. Sentence captions extend keyword labels with short, aircraft-visible appearance descriptions derived from fine-grained species annotations.

Literature Comparison For each dataset, we include metrics from prior publications. These were *not rerun* unless original train, validation, and test splits were unavailable (marked ‡). Bold entries indicate the better metric between literature and OW Keyword results. This comparison highlights OW's performance relative to previously reported approaches under the same data conditions.

4.4 Results

4.4.1 Closed-Set SOTA Comparison

Figure 4.2 compares OW against closed-set detectors—DINO [137], CO-DETR [144], and YOLOv8 [135]—as well as MM-Grounding-DINO [141], which retains open-vocabulary capability but lacks aerial-domain pre-training. In the zero-shot setting, OW outperforms all baselines, demonstrating the combined importance of semantic text guidance and aerial pre-training. MM-Grounding-DINO performs substantially worse, isolating the impact of domain-specific pre-training. Under fine-tuning, closed-set models improve and approach

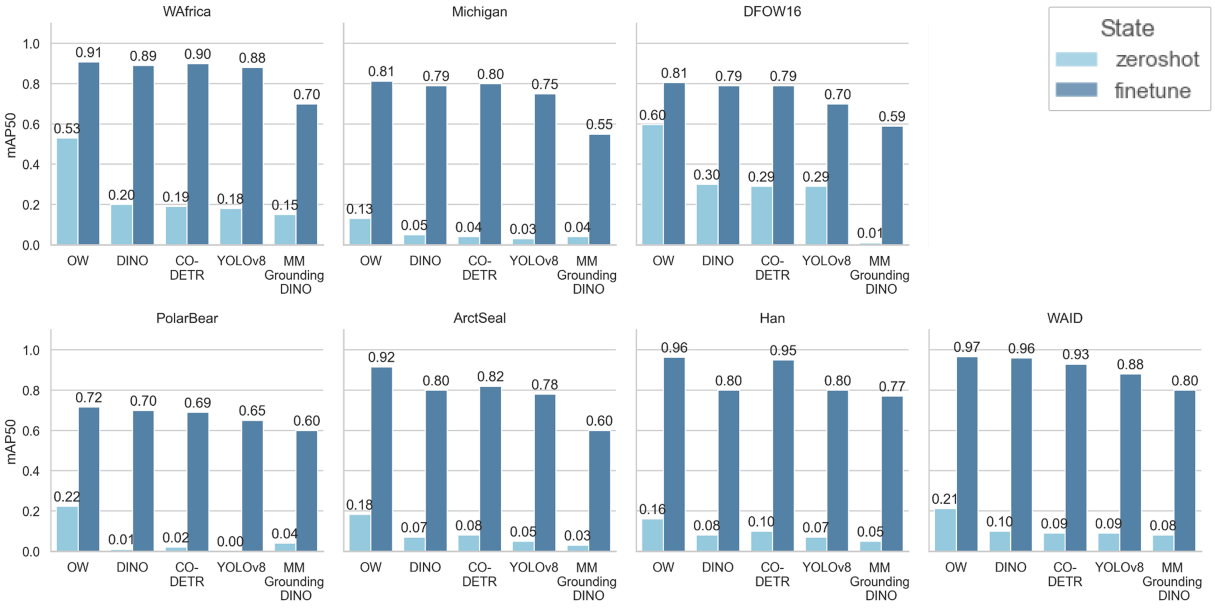


Figure 4.2: Comparison of OW with closed-set detectors (DINO [137], CO-DETR [144], YOLOv8 [135]) and MM-Grounding-DINO [141] on novel-species datasets under zero-shot and fine-tuning settings, illustrating the effects of open-vocabulary detection and aerial-domain pre-training.

OW performance, reducing the relative advantage of text guidance while confirming the complementary roles of open-vocabulary design and aerial-domain pre-training.

4.4.2 Literature Comparison

As reported in Table 4.4, OW surpasses most prior methods, with the exceptions of the Qian and Virunga-Garamba datasets. The Qian dataset, which contains densely packed penguin colonies, is particularly suited to crowd-counting regression methods [101] that better handle occlusions [74]. For Virunga-Garamba, OW is slightly outperformed by [58], likely due to their use of the larger SAM ViT-L backbone [54] with 308M parameters, compared to OW's more compact 172M-parameter model.

Dataset	**Original Label(s)**	**Literature Results**	**OW Keyword**	
			Lit. Metric	mAP50
AED [94]	Elephant	0.89 mAP † [94]	**0.90** mAP †	0.856
DFOW14 [85]	Beluga whale	-	-	0.856
DFOW15 [86]	Beluga whale	-	-	0.82
DFOW17 [128]	Beluga whale	-	-	0.857
Gray [33]	Turtle	0.27 F1 ‡	**0.38** F1	0.26
IndOcean [130]	Bird	0.72 F1 [130]	**0.80** F1	0.833
Izembek [131]	Brant, Canada goose, Emperor goose	-	-	0.343
NABird [130]	Bird	0.72 F1 [130]	**0.93** F1	0.959
NewMex [130]	Bird	0.76 F1 [130]	**0.826** F1	0.728
NOAA [48]	Sea lion	10.86 RMSE [48]	**8.13** RMSE	0.662
Palmyra [130]	Bird	0.72 F1 [130]	**0.78** F1	0.789
Pfeifer [130]	Penguin	0.66 F1 [130]	**0.86** F1	0.842
Qian [101]	Penguin	**66.27** RMSE ‡	85.32 RMSE	0.671
Virunga-Gar. [22]	Alcelaphinae, Buffalo, Kob, Warthog, Waterbuck, Elephant	**0.85** mAP50 [58], 0.835 F1 [24]	0.72 mAP50, **0.85** F1	0.719

Table 4.4: Results of open-set continuous fine-tuning. **Bold** numbers denote the best metric between literature results and OW Keyword results. Results marked with † indicate a 200-pixel Chebyshev distance allowance to be considered a true positive. Literature results marked with ‡ were reproduced by us due to the original train, validation, and test splits not being available. Descriptions of the datasets are available in Table A.1.

Dataset	Original Label(s)	OW Sentence	OW Keyword
AFSC [2]	Seal	**0.257**	0.183
Chabot [16]	Polar bear	**0.369**	0.224
DFOW16 [84]	Narwhal	0.576	**0.597**
Han [35]	Yak, Sheep	0.106	**0.161**
Michigan [130]	Gull	**0.179**	0.131
WAfrica [53]	Royal tern, Caspian tern, Slender-billed gull, Gray headed gull, Great cormorant, Great white pelican	0.465	**0.533**
WAID [92]	Sheep, Cattle, Seal, Camelus, Kiang, Zebra	**0.229**	0.212

Table 4.5: Comparison of OW Sentence and OW Keyword models on novel species datasets, with all results reported as mAP50. **Bold** numbers indicate the best mAP50 between the two models. Descriptions of the datasets are available in Table A.2.

4.4.3 Detection of Novel Labels

We evaluate detection of novel labels on seven datasets, with performance reported in two separate tables. Table 4.5 compares the OW Sentence and OW Keyword models directly, reporting mAP50 scores. OW demonstrates moderate novel-label detection performance, achieving up to 0.597 mAP50. In this comparison, OW Sentence outperforms OW Keyword when the sentence captions are sufficiently descriptive, improving bounding box precision and robustness to background features.

Table 4.6 presents finetuned OW Keyword results alongside literature-reported metrics. In most cases, OW matches or surpasses prior methods, confirming that even minimal finetuning (10 epochs) allows the model to effectively generalize to unseen labels.

Descriptive captions that highlight shape, size, and other contextual elements enable OW to better align visual features with label semantics, reducing confusion between visually similar objects. However, overly long or inaccurate LLM-generated captions can hinder performance, as reflected in some failure cases. Qualitative examples demonstrating OW's multi-label detection capability in both terrestrial and marine scenarios are shown in Figure 4.3.

Dataset	**Original Label(s)**	**Lit. Results**	**Lit. Metric**	**Keyword (mAP50)**
AFSC [2]	Seal	**0.87** F1 ‡	**0.87** F1	0.916
Chabot [16]	Polar bear	-	-	0.716
DFOW16 [84]	Narwhal	-	-	0.805
Han [35]	Yak, Sheep	0.89 mAP50 [35]	**0.96** mAP50	0.964
Michigan [130]	Gull	0.72 F1 [130]	**0.82** F1	0.813
WAfrica [53]	Royal tern, Caspian tern, Slender-billed gull, Gray headed gull, Great cormorant, Great white pelican	0.79 F1 [53]	**0.91** F1	0.908
WAID [92]	Sheep, Cattle, Seal, Camelus, Kiang, Zebra	**0.98** mAP50 [92]	0.97 mAP50	0.967

Table 4.6: Comparison of finetuned OW Keyword results with literature results on the same datasets. **Bold** numbers indicate the best metric. Literature results marked with ‡ were reproduced by us due to unavailable original splits. Descriptions of the datasets are available in Table A.2.

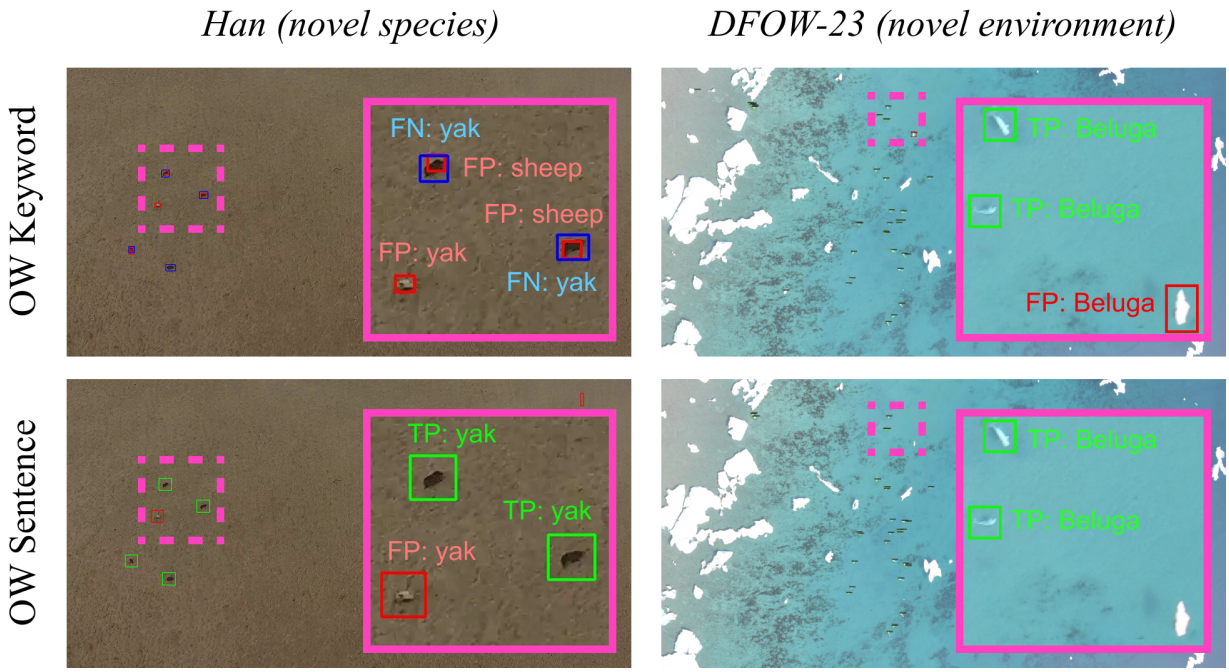


Figure 4.3: Qualitative detection results of the OW Keyword and OW Sentence models on novel species [35] and in novel environments. TP is marked as green, FP as red, FN as blue. Grounding with descriptive sentences can reduce class confusion and improve robustness to background features.

Dataset	Original Label(s)	OW Sentence	OW Keyword
DFOW23	Narwhal, Beluga whale, Bowhead whale	0.024	**0.362**
Kenya [29]	Zebra, Elephant, Giraffe	**0.384**	0.326
Liu [78]	Penguin	**0.733**	0.619

Table 4.7: Comparison of OW Sentence and OW Keyword models on novel environment datasets, reported as mAP50. **Bold** numbers indicate the better mAP50 between the two models. Descriptions of the datasets are available in Table A.3.

4.4.4 Detection in Novel Environments

We evaluate detection in novel environments using a test set of three datasets comprising labels that were present in the training dataset but captured in previously unseen environments. This setup allows us to assess domain generalization: the model must detect familiar species or label groups under new visual contexts, rather than encountering entirely new species.

Table 4.7 reports the direct comparison between OW Sentence and OW Keyword models, with mAP50 scores for each dataset. OW demonstrates strong domain transfer capabilities, achieving up to 0.733 mAP50. In these comparisons, OW Sentence outperforms OW Keyword on some datasets, while OW Keyword excels on others, highlighting the complementary benefits of descriptive sentence captions versus concise keyword guidance.

Table 4.8 presents finetuned OW Keyword results alongside literature-reported metrics. Across all datasets, OW matches or exceeds previously reported methods after minimal fine-tuning (10 epochs), confirming its ability to generalize previously learned labels to novel environmental contexts.

4.4.5 Limitations

While OW demonstrates strong open-vocabulary and aerial-domain detection capabilities, certain failure cases highlight its limitations. Misleading or incomplete LLM-generated captions can negatively impact detection performance. For example, OW failed to detect a gull [130] because the caption described it as "white" based on adult coloration. This description does not account for age-related variation (e.g. juveniles could be brown). Detection was further complicated because the bird is faint against the background.

Another failure case involved misclassification of snow mounds as polar bears [16], likely

Dataset	**Original Label(s)**	**Lit. Results**	**Lit. Metric**	**OW Keyword**
DFOW23	Narwhal, Beluga whale, Bowhead whale	-	-	0.435
Kenya [29]	Zebra, Elephant, Giraffe	0.77 mAP30 [29]	**0.869** mAP30	0.854
Liu [78]	Penguin	-	-	0.755

Table 4.8: Comparison of finetuned OW Keyword results with literature results on novel environment datasets. **Bold** numbers indicate the best metric. Descriptions of the datasets are available in Table A.3.

due to the model's lack of exposure to that type of background during training. These examples, visualized in Figure 4.4, illustrate that both caption quality and scene context can limit OW's performance, particularly in complex terrestrial and marine environments.

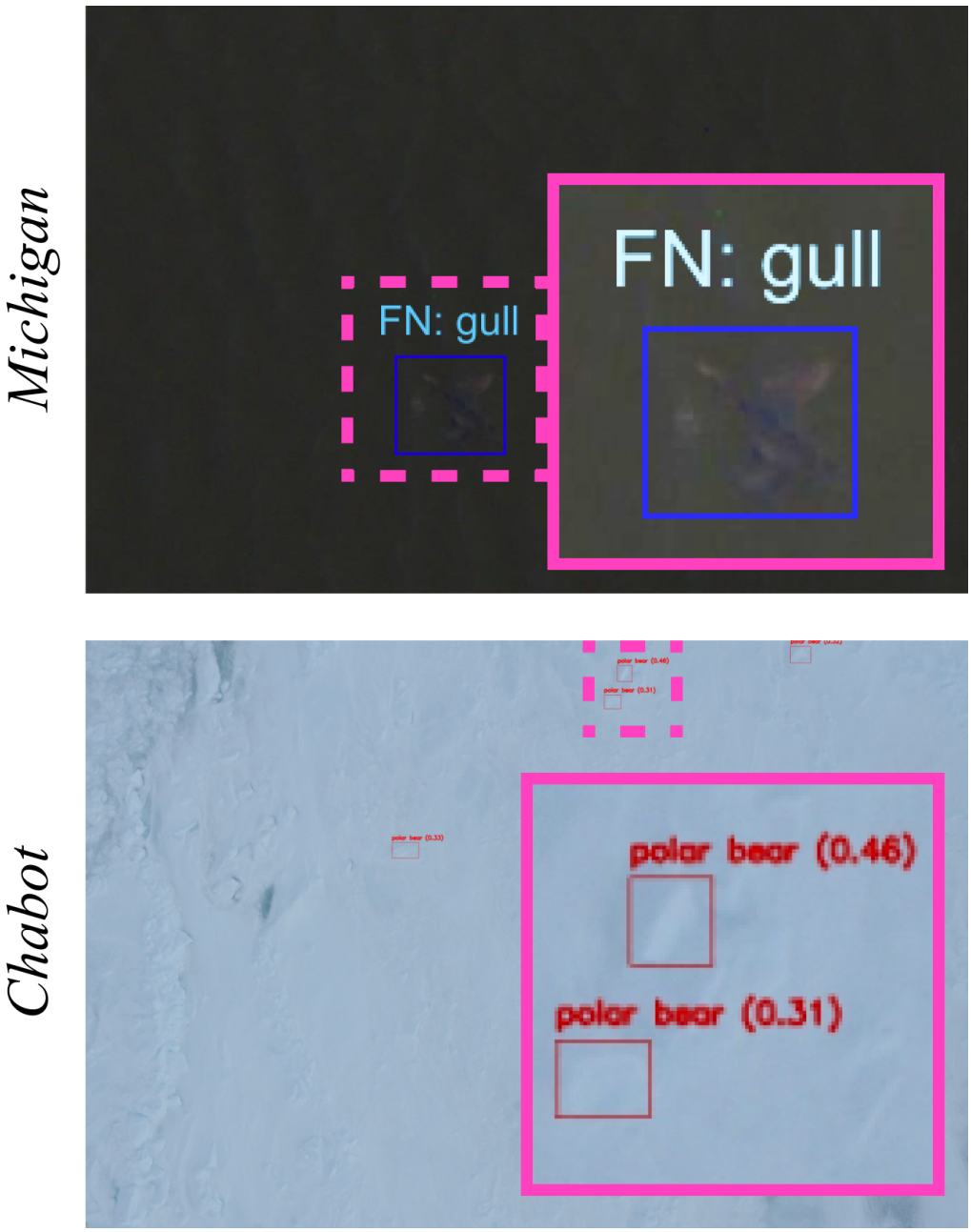


Figure 4.4: Limitations of novel species detection. Left: OW Sentence missed a gull [130] due to age-related coloration and low contrast. Right: OW misclassified snow mounds as polar bears [16]; insets are brightened for clarity.

Chapter 5

Deployment

This chapter describes the deployment of our human-in-the-loop annotation system in a real-world wildlife monitoring context. While previous sections detailed model design and training, here we evaluate how the OpenWildlife Keyword described in Section 4.2 performs when embedded into an operational workflow, with particular attention to annotation efficiency, usability, and incremental model improvement.

The system is developed in collaboration with the Arctic Eider Society to support long-term monitoring of the common eider (*Somateria mollissima*) in Hudson Bay. The common eider holds significant cultural and ecological importance for Inuit communities, as eider down is among the warmest natural insulators and has historically been essential for Arctic survival [37]. Beyond its cultural value, the demographic structure of eider colonies functions as a sensitive ecological indicator. Environmental changes, including shifting sea ice dynamics and altered ocean current patterns associated with hydroelectric development in eastern North America, increasingly affect both the birds and the communities that depend on them [108]. Reliable, large-scale demographic monitoring is therefore of both scientific and societal importance.

Eiders are sexually dimorphic, with males exhibiting predominantly white plumage and females displaying brown coloration. This distinction enables sex-specific identification in aerial imagery and allows demographic ratios to be estimated directly from survey data. However, the practical challenges are substantial. In high-resolution aerial images, individual eiders occupy roughly 10×10 pixels and frequently appear in densely packed flocks. White males can also be visually confounded with ice fragments due to similar coloration. As a result, manual identification and differentiation of individuals is extremely time-consuming and prone to fatigue-related errors. These characteristics make the task

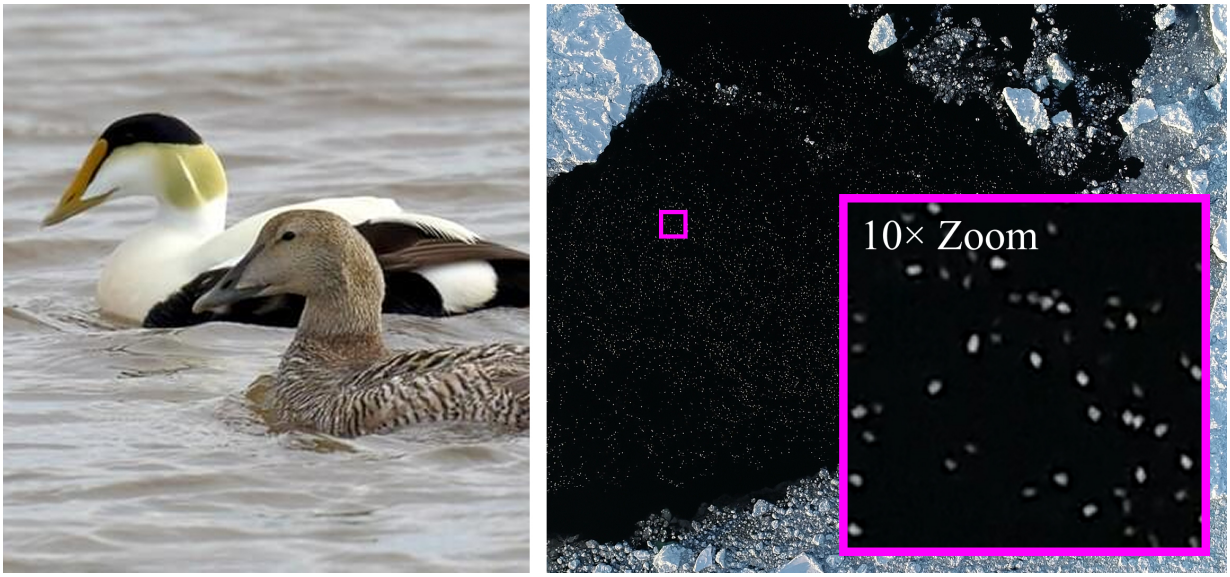


Figure 5.1: Visual characteristics of eider ducks in aerial surveys. Left: Reference images showing sexually dimorphic plumage between males (white) and females (brown) [19]. Right: Example aerial survey image (4368×2912) with 10× zoom inset displaying the small size of eiders.

well-suited to a semi-automated annotation system that supports human expertise rather than replacing it.

5.1 Deployment-specific Dataset

The system targets a recurring aerial survey program conducted by the Arctic Eider Society. The dataset consists of 750 high-resolution (4368 × 2912 pixels) aerial images of eider duck colonies collected by manned aircraft or drones from oblique and nadir viewpoints over the Belcher Islands in 2002, 2008, and 2021. Over a four-month period, six images were fully annotated using a default configuration of the open-source labelling platform Label Studio [122], producing 8,339 keypoint annotations corresponding to individual ducks. The annotation effort was carried out in collaboration with the Arctic Eider Society through an Indigenous co-operative program. These labels serve as the basis for fine-tuning our system.

5.2 Workflow Implementation

Figure 5.2 illustrates the HITL pipeline. Unlike the standard LabelStudio workflow, which requires full-image annotations before retraining, our setting involves images containing thousands of ducks, making full-image corrections impractical. We implemented a *regional* workflow where annotators interact with smaller image regions:

1. The current model generates initial predictions for the full image.
2. The annotator edits a selected subregion, drawing polygons to indicate verified ground truth.
3. Pressing the *Train* button fine-tunes the model solely on this verified region, following similar existing strategies for learning with partial annotations (Section 2.1.5). During fine-tuning, a mask is applied such that all pixels outside the annotated polygons are ignored (set to zero), ensuring the model only updates on verified areas.
4. The updated model is redeployed, allowing predictions on remaining areas.

This iterative predict–annotate–train cycle is repeated until the entire image is annotated, without requiring exhaustive corrections at each step.

To facilitate this workflow, the interface incorporates several optimizations. Dense clusters of eiders are grouped visually using a bubble-based layout to improve navigation responsiveness (Figure B.1). Model versioning, checkpoint management, and overall configuration are handled through an enhanced settings page (Figure B.2), while detailed training and testing parameters can be edited through a dedicated dialog within this interface (Figure B.3). The underlying software architecture, including the frontend, backend, ML backend, and asynchronous training workflows, is shown in Figure B.4. These features collectively ensure that annotators can efficiently correct predictions, update models, and continue working without interruption.

5.3 Experiments

Evaluation focuses on two aspects: the reduction in annotation time achieved by iterative model refinement, and the improvement in detection and counting accuracy as human feedback accumulates. Ten aerial images, representative of varying ice cover, viewing

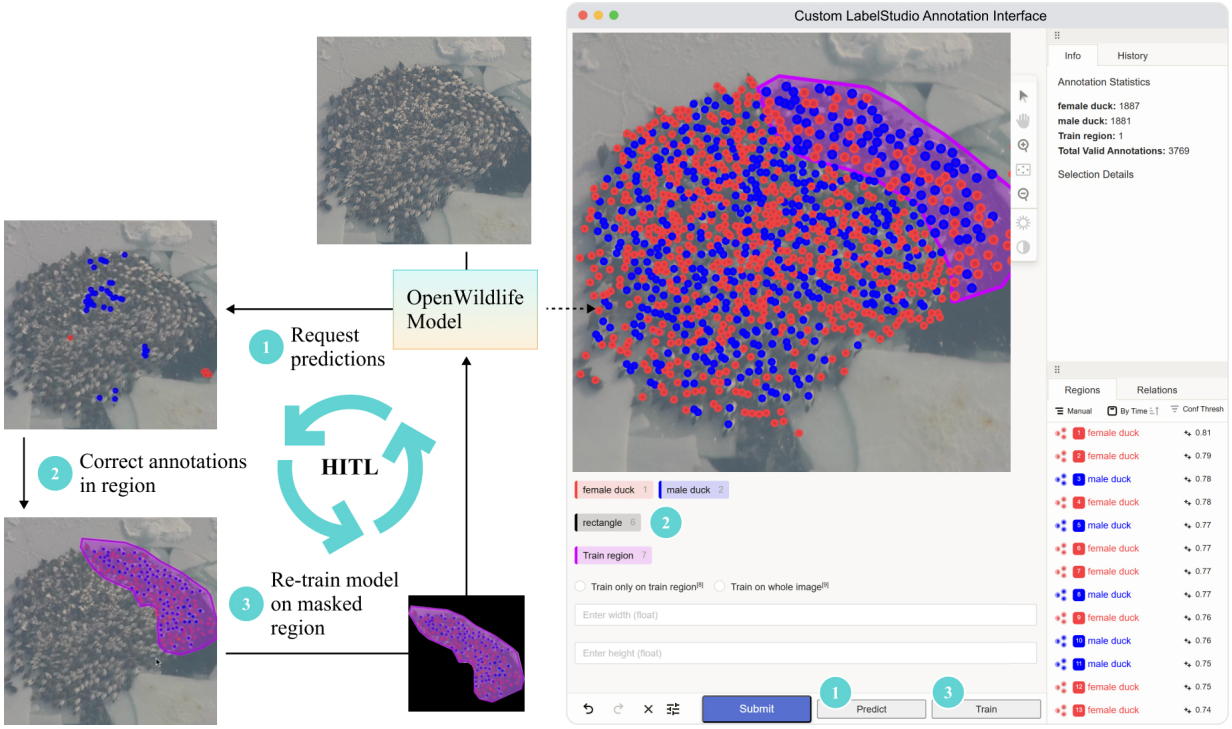


Figure 5.2: The human-in-the-loop pipeline built within LabelStudio. Given predictions from our OpenWildlife model, a human can correct annotations within a region and re-train the model on this masked area. This cycle can be repeated until the annotations are satisfactory. The right panel shows the interface in which this pipeline is executed, including bubbled numbers which reflect the pipeline's steps, annotation tools, and other quality-of-life features.

geometry, flock density, and behavioural context, were held out as a test set. Despite their small number, these images contain roughly 35,000 eider ducks across 320 patches of size 512×512 pixels, illustrating the difficulty of dense aerial annotation.

Three model variants were evaluated: the pre-trained OpenWildlife Keyword model, the same model fine-tuned on six fully annotated images, and the fine-tuned model further refined through human-in-the-loop corrections.

Performance metrics include precision, recall, mean absolute error (MAE), and percentage counting error, all evaluated at a detection confidence threshold of 0.3. Annotation efficiency was measured as the average time to label a full-resolution image.

5.4 Results

5.4.1 Accuracy and Efficiency

Table 5.1 summarizes the performance of the three model variants. Fine-tuning substantially improves accuracy over the pre-trained model, while human-in-the-loop refinement yields further gains, particularly in recall and MAE. The improvement in counting accuracy demonstrates that localized corrections can propagate meaningful global improvements across dense colonies.

Precision and recall are reported as zero for the pre-trained model because it fails to produce detections that satisfy the evaluation criteria on this dataset. The pre-trained OpenWildlife model was not trained on eider imagery, and the extremely small object size (approximately 10×10 pixels) and dense colony structure differ substantially from the distributions encountered during pre-training. As a result, none of the model's predictions exceed the detection confidence threshold of 0.3 while also matching ground-truth annotations under the evaluation protocol. Consequently, the model registers zero true positives, yielding precision and recall values of zero.

Annotation efficiency is dramatically improved under the human-in-the-loop workflow, with average labelling time per image reduced by 87.5% relative to fully manual annotation (Table 5.2).

5.4.2 Generalization Across Images

To assess potential overfitting, we performed a leave-out experiment in which HITL training was applied to five non-overlapping images and evaluated on the same 10-image test set.

Pre-trained	✓	✓	✓
Fine-tuned	✗	✓	✓
HITL	✗	✗	✓
Precision ↑	0.0	**67.11**	65.97
Recall ↑	0.0	75.64	**77.56**
MAE ↓	3384.7	548	**264.3**
% Error ↓	98.08	29.20	**22.23**

Table 5.1: Performance comparison of deployed model variants on the 10-image held-out test set.

Workflow	Time Per Image (minutes) ↓
Manual	117.08 ± 111.29
HITL	**14.70 ± 9.73**

Table 5.2: Annotation time comparison between manual labelling and the deployed HITL workflow.

Results (Table 5.3) indicate that HITL refinement improves counting and detection across locations and conditions, confirming that localized corrections produce broadly applicable gains.

Fine-tuned	✓	✓
HITL Other	✗	✓
Precision ↑	**67.1**	63.5
Recall ↑	75.6	**81.1**
MAE ↓	548	**531.7**
% Error ↓	29.2	**28.9**

Table 5.3: Generalization performance after HITL refinement on non-overlapping images.

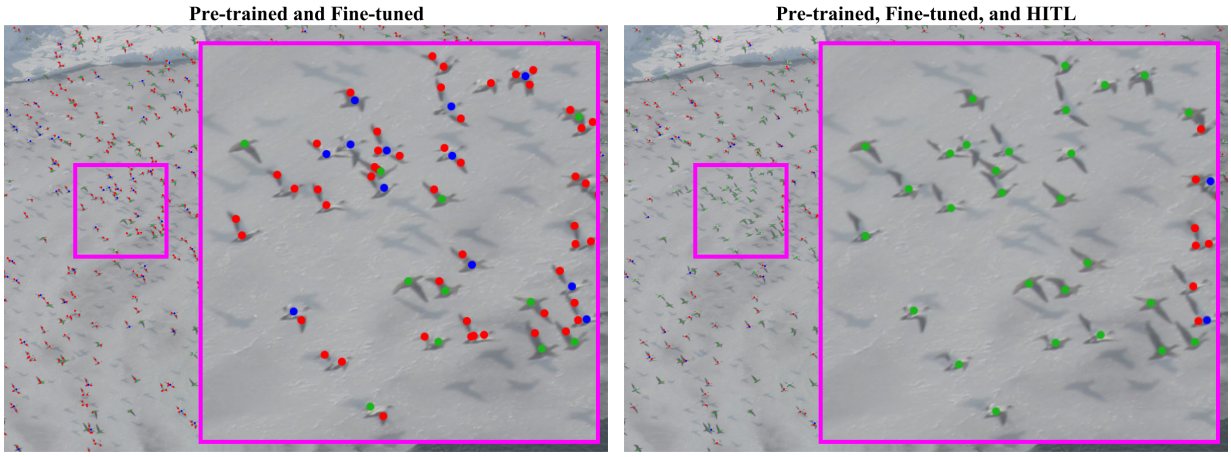


Figure 5.3: Qualitative improvement through human-in-the-loop refinement. Comparison shows detection results before (left) and after (right) HITL feedback, with true positives (green), false positives (red), and false negatives (blue). The HITL-refined model demonstrates improved precision in distinguishing eiders from background clutter and better recall in dense flock regions, as visible in the 5× zoom insets.

5.4.3 Qualitative Analysis

Figure 5.3 illustrates qualitative improvements achieved through HITL refinement. Compared with the fine-tuned baseline, the model produces fewer false positives in cluttered regions and recovers more true positives in dense flocks, as also exemplified in Figure 5.1. Figure 5.4 demonstrates that a small number of localized corrections is sufficient to bring model output close to fully manual labels.

5.5 Summary

Embedding the OpenWildlife Keyword model within a regional human-in-the-loop workflow substantially reduces annotation effort while steadily improving detection and counting accuracy. The deployment confirms that dense aerial surveys of eider ducks can be analyzed efficiently without exhaustive labelling.

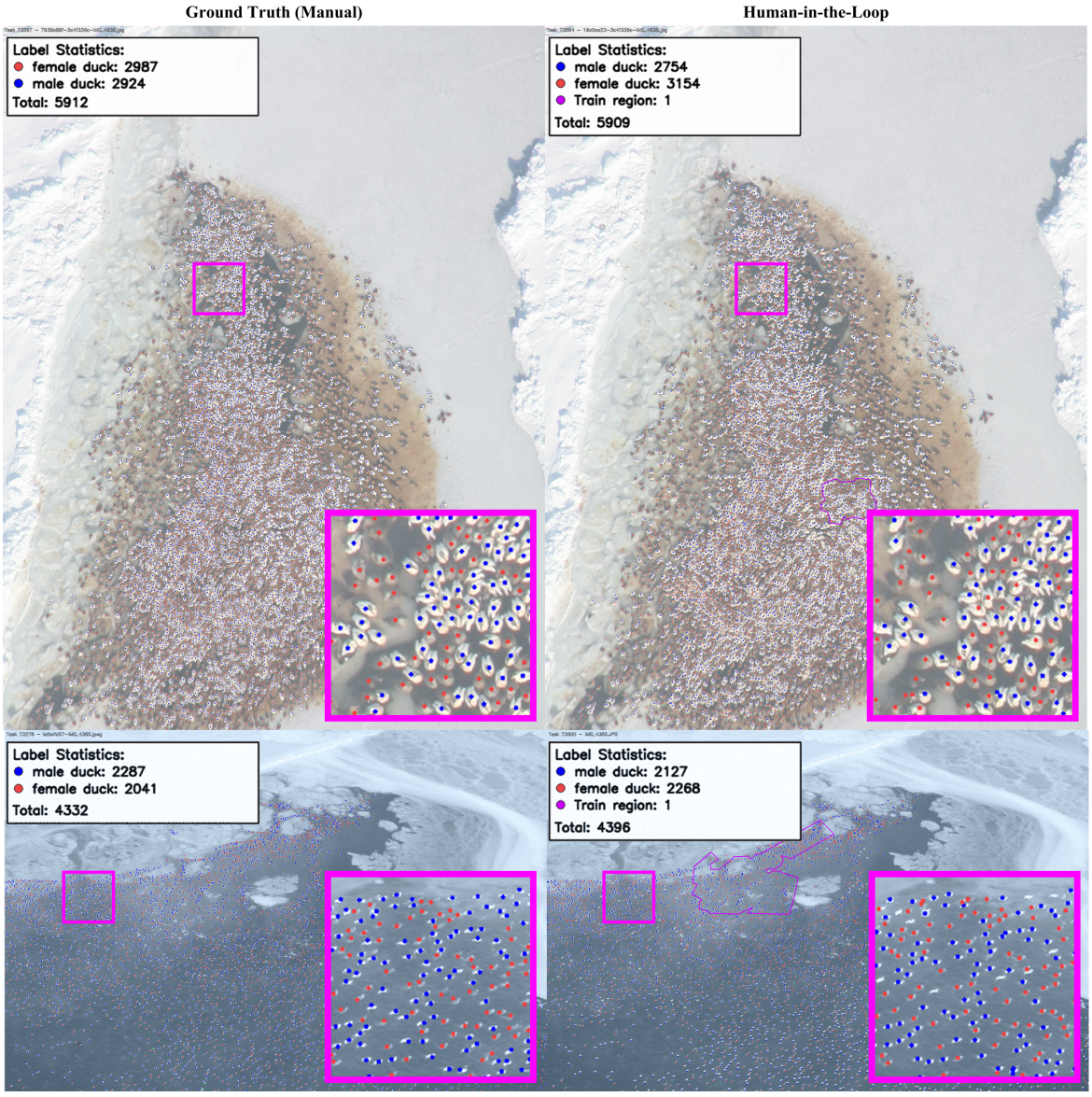


Figure 5.4: Comparison of a fully manual annotation workflow (left) and the human-in-the-loop (HITL) process (right). Pink regions highlight areas where annotators intervened during the HITL cycle, while cyan insets show 4× zoomed views of the corresponding regions. Final counts of male and female ducks are shown in the upper left of each image. The side-by-side comparison illustrates how a small number of localized corrections can bring the HITL output into close agreement with fully manual annotations.

Chapter 6

Conclusion

6.1 Summary

This thesis investigated the use of open-vocabulary deep learning models for multi-species wildlife detection in aerial imagery, addressing the challenges of species generalization, domain shift, and high-density annotation. Chapter 3 detailed the construction of a unified benchmark by consolidating publicly available aerial wildlife datasets. Textual descriptions were aligned with species instances to enable natural language–guided detection, and dataset splits were designed to explicitly test cross-species and cross-environment generalization. This benchmark establishes a foundation for evaluating open-vocabulary detection models and provides a standard for comparing performance across diverse ecological contexts.

Chapter 4 introduced **OpenWildlife (OW)**, a foundation-model–based wildlife detector that supports flexible species specification through natural language prompts. The architecture leverages pre-trained visual-language embeddings and fine-tuning strategies to enable few-shot adaptation to novel species, reducing reliance on extensive task-specific annotation. Human-in-the-loop workflows were incorporated to refine predictions iteratively, allowing expert corrections to directly improve subsequent model iterations. Quantitative evaluation showed that OW achieves strong detection performance across known and previously unseen species, while qualitative analysis highlighted its ability to maintain accuracy in complex, high-density scenes.

Chapter 5 demonstrated the practical application of OW in collaboration with the Arctic Eider Society. The system reduced annotation times by nearly 90% compared to fully

manual labelling while achieving a 22.23% error. Iterative regional correction allowed the model to improve continuously, particularly in densely populated areas where conventional full-image annotation would be prohibitive. At the same time, challenges remain: domain shifts and visually ambiguous clusters occasionally lead to misclassifications, and reliance on text-based descriptions can introduce errors when captions are inaccurate or incomplete. Visualized failure cases illustrate these limitations and provide guidance for future improvement.

Overall, this work demonstrates that combining open-vocabulary foundation models with human-in-the-loop workflows can enable scalable and accurate wildlife monitoring. By bridging the gap between automated detection and expert oversight, OW provides a flexible framework capable of supporting both scientific research and conservation management in data-scarce environments.

6.2 Future Work

6.2.1 Ecological Impact Assessment

While this thesis focused on improving detection accuracy and annotation efficiency, the ultimate objective of OW is to support actionable ecological inference. A critical next step is to explicitly quantify how residual detection and classification errors propagate to population-level metrics such as abundance, density, and species distribution. Ecological inference frameworks have long recognised that imperfect detection introduces bias into downstream estimates, motivating the development of hierarchical models that explicitly separate observation error from ecological processes [59, 112]. However, these frameworks typically assume missed detections rather than systematic misclassification, an important distinction in the context of open-vocabulary detection.

In aerial wildlife surveys, detection errors arise from occlusion, scale variation, and visual similarity between species, all of which can induce both false negatives and class-dependent misclassification [17, 39]. While missed detections can often be modelled through detection probabilities, misclassification introduces structured error that may violate assumptions of standard models such as N-mixture formulations [111]. In particular, when errors are correlated with species identity, density, or environmental conditions, they can lead to biased estimates that are not easily corrected without explicitly modelling the classification process.

To address this, future work will conduct a comparative analysis of three annotation regimes: (i) fully manual labelling, (ii) fully automated or fine-tuned model predictions,

and (iii) HITL-assisted classification. For each regime, outputs will be used to estimate ecological quantities such as colony size, spatial density, and occupancy. These estimates will be benchmarked against high-confidence reference annotations to quantify bias, variance, and error propagation. Importantly, this analysis will distinguish between detection failure (missed individuals) and misclassification (incorrect species labels), enabling assessment of how each error type affects ecological conclusions.

Understanding error propagation is particularly important in regimes involving rare species or densely aggregated colonies, where small absolute miscounts can produce disproportionately large relative errors. Prior work has shown that detection bias in aerial surveys can distort abundance estimates and spatial patterns, particularly when detection probability varies across space or with group size [104]. Similarly, recent studies integrating deep learning with UAV imagery highlight that automated pipelines can achieve high apparent accuracy while still introducing systematic biases in population estimates if errors are not explicitly accounted for [82].

By systematically evaluating these effects, it becomes possible to define thresholds of acceptable detection performance for specific ecological tasks. For example, coarse-grained monitoring of population trends may tolerate higher error rates than applications requiring precise estimation of endangered species abundance. This analysis also provides a principled basis for refining the HITL workflow, identifying conditions under which additional human intervention yields the greatest improvement in ecological reliability.

Ultimately, integrating detection outputs with ecological modelling frameworks bridges the gap between technical performance metrics (e.g., mAP) and domain-relevant outcomes. At the same time, this integration must be approached cautiously, as standard models may require extension to explicitly account for classification uncertainty. Establishing this connection is essential for substantiating claims that OW can support conservation programs, as it enables direct evaluation of whether efficiency gains from HITL annotation are achieved without compromising the validity of ecological inference.

6.2.2 Multi-Modal Sensor Integration

Enhancing wildlife detection through additional sensor modalities offers a promising avenue for improving OW performance under challenging ecological conditions. Thermal infrared (TIR) imagery, multispectral sensors, and LiDAR provide complementary information to conventional RGB imagery, enabling detection where animals are visually cryptic or obscured by dense vegetation.

Thermal imaging captures the heat signatures of warm-blooded animals, making it possible to distinguish individuals from complex backgrounds even when RGB contrast is low or lighting is poor. Studies have demonstrated the feasibility of integrating UAV-acquired thermal and RGB data for real-time wildlife detection, showing that combined modalities improve precision and recall compared with RGB alone, despite limitations such as thermal resolution and sensor altitude [64]. TIR-specific deep learning networks, such as ALSS-YOLO, further enhance detection of small or overlapping targets by optimizing feature extraction and attention mechanisms for thermal imagery [36].

Multispectral imagery adds additional spectral bands that provide information on vegetation, soil, and temperature, which can improve background modelling and reduce false positives. Cross-modality attentive fusion strategies have been shown to effectively combine RGB and thermal or multispectral data, enhancing feature representations and improving robustness in low-visibility or complex environments [102, 140]. Early field studies using multispectral and thermal UAV imagery confirmed that multiple species can be detected simultaneously, including bison, deer, wolves, and elk, with minimal misclassification, demonstrating practical viability for ecological monitoring [18].

Integrating these modalities into OW would require both hardware and software adaptations. UAV platforms must capture registered RGB, thermal, and multispectral imagery at sufficient resolution to allow pixel- or feature-level fusion. On the algorithmic side, detection networks could employ early fusion at the input level, feature-level fusion within the network, or decision-level fusion of modality-specific predictions. The resulting system would be more resilient to environmental variability, such as canopy cover, low light, or seasonal changes in habitat.

Overall, multi-modal integration has the potential to significantly enhance OW's operational utility. By leveraging thermal and multispectral signals, the system could maintain high detection performance in conditions where RGB alone is insufficient, enabling robust, all-weather, and large-area wildlife monitoring campaigns.

6.2.3 Enhanced Caption Generation and Alignment

Advancing wildlife detection through richer textual supervision offers a promising direction for improving OW performance, particularly under conditions of visual ambiguity or high species diversity. While current captioning approaches summarize coarse appearance features, future work could leverage region-aware Visual Large Language Models (VLLMs) to generate detailed, semantically precise descriptions directly from detected bounding boxes. This would allow captions to capture subtle attributes such as posture, relative size, or

behaviour-specific traits, potentially improving open-vocabulary generalization to novel species or challenging environments [70].

Region-grounded captioning has been shown to enhance detection performance by linking textual descriptions to specific visual features, enabling models to distinguish similar species or individual instances more accurately [134, 138]. Cross-modal attention and feature-fusion strategies allow the model to integrate visual context with linguistic embeddings, improving the relevance and discriminative power of generated captions [69, 126, 127, 139]. Recent advances in unified vision-language frameworks, such as GLIPv2 and UniTAB, further demonstrate the feasibility of combining localization and language supervision within a single architecture, facilitating more robust alignment between text and region proposals [134, 138].

In practical terms, integrating region-aware captioning into OW would involve generating fine-grained textual descriptions for each predicted bounding box, which could then guide both open-vocabulary detection and post-hoc analysis. Attention mechanisms could prioritize features most informative for distinguishing visually similar species or behaviours, while multimodal fusion could ensure that captions remain coherent even in cluttered aerial imagery [69, 126]. By capturing richer semantic detail, this approach has the potential to reduce misclassification in complex or densely populated scenes and to support identification down to subspecies or behaviour-specific distinctions.

Overall, enhanced caption generation and alignment would complement OW's current human-in-the-loop pipeline by providing more informative textual supervision, enabling models to leverage both visual and linguistic cues effectively. This extension could improve detection accuracy, accelerate adaptation to novel species, and support scalable, high-fidelity wildlife monitoring in diverse ecological settings [70].

6.2.4 Active Learning Integration

A natural extension of the human-in-the-loop (HITL) component in this work is to incorporate formal active learning strategies that prioritize annotation effort on the most informative samples. In conventional supervised learning, each unlabeled image is treated as equally valuable; in active learning, the model actively selects which images should be labelled next based on quantitative criteria such as uncertainty, expected model improvement, or diversity in visual features. This can reduce annotation cost for a given performance target [115].

Active learning operates in a cyclical loop: (1) train the detector on the current labelled set, (2) score unlabelled images according to an acquisition function that estimates their

informativeness, (3) request human annotations for the top-ranked images, and (4) update the model with these new labels. Approaches that combine classification confidence with localization cues, as well as uncertainty- or diversity-based metrics, have been proposed for object detection tasks [49, 68, 12].

In aerial wildlife imagery, selective annotation can drastically reduce human effort. Experiments with UAV datasets have shown that less than 1% of labelled images can be sufficient to retrieve the majority of animal instances when active learning drives sample selection, compared with random sampling [50]. Similarly, active learning pipelines applied to ecological camera trap images have demonstrated reduced annotation requirements while maintaining or improving predictive performance across multiple species [10].

More general work on deep active learning for object detection highlights strategies that can be adapted to ecological datasets. Methods that consider location stability, localization tightness, and class-balanced uncertainty sampling can further improve efficiency in object detection tasks [49, 68, 12]. These studies indicate that actively selecting samples based on their potential to improve the model can reduce annotation effort and guide continuous performance improvement.

Operationalizing active learning within the deployed HITL pipeline entails several practical considerations. Acquisition functions must be computed efficiently at scale, especially if OW is tasked with scoring thousands of unlabelled images for a new survey region. The choice of criterion (uncertainty, diversity, or hybrid) should be validated in the ecological context of interest; what reduces uncertainty most effectively in wildlife imagery may differ from generic benchmarks due to foreground-background imbalance, species camouflage, or sensor noise. The system should support an annotation interface that can seamlessly present selected samples to human experts and capture corrected labels in a way that integrates back into OW's training loop.

Integrating active learning would not only reduce annotation workload further, but also systematically identify and mitigate model weaknesses, especially in areas of domain shift or rare species. By guiding human effort to where the model is least confident or most likely to improve, this extension would strengthen OW's HITL framework and support sustainable deployment in long-term monitoring campaigns.

6.2.5 Keypoint-Native Detection for Wildlife Monitoring

A key direction for future work is to extend DINOKey to operate within larger wildlife monitoring pipelines that natively support point annotations [43]. Many ecological datasets, including those used in aerial surveys and OpenWildlife-style benchmarks, are annotated with

points due to lower annotation cost and reduced inter-annotator variability [129, 101, 23]. Converting these annotations to bounding boxes introduces assumptions about object extent that are often inconsistent, particularly for small or densely packed animals [4].

By aligning model outputs directly with point supervision, DINOKey avoids these limitations and provides a natural interface for counting and localization tasks. Extending this framework to multi-species or open-vocabulary detection would eliminate the need for heuristic box generation, but introduces challenges in matching and evaluation. In particular, the absence of scale information complicates query initialization and assignment, suggesting the need for class-aware or adaptive matching strategies beyond the current Hausdorff-based formulation [87].

Evaluation protocols also require reconsideration. Fixed distance thresholds, such as the Chebyshev criterion used in this work, may not generalize across species with varying physical sizes, motivating adaptive or uncertainty-aware metrics.

Finally, integrating keypoint-native detection into human-in-the-loop workflows offers practical advantages. Point annotations are faster to correct, enabling more efficient iterative refinement, while the set-based formulation inherited from DETR and DINO preserves global reasoning without post-processing [14, 137].

References

[1] Evgenios Agathokleous, Costas J. Saitanis, Chao Fang, and Zhen Yu. Use of ChatGPT: What does it mean for biology and environmental science? *Science of The Total Environment*, 888:164154, 8 2023. 18

[2] Alaska Fisheries Science Center. A dataset for machine learning algorithm development from 2010-06-15 to 2010-08-15. https://www.fisheries.noaa.gov/inport/item/63322, 2026. Accessed: 2026-04-02. 36, 37, 79

[3] Saleema Amershi, Maya Cakmak, W. Bradley Knox, and Todd Kulesza. Power to the People: The Role of Humans in Interactive Machine Learning. *AI Magazine*, 35(4):105–120, 2014. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1609/aimag.v35i4.2513. 2, 20

[4] Marie R. G. Attard, Richard A. Phillips, Ellen Bowler, Penny J. Clarke, Hannah Cubaynes, David W. Johnston, and Peter T. Fretwell. Review of Satellite Remote Sensing and Unoccupied Aircraft Systems for Counting Wildlife on Land. *Remote Sensing*, 16(4):627, January 2024. 56

[5] Jayme Garcia Arnal Barbedo, Luciano Vieira Koenigkan, Thiago Teixeira Santos, and Patrícia Menezes Santos. A Study on the Detection of Cattle in UAV Images Using Deep Learning. *Sensors*, 19(24):5436, January 2019. 2

[6] Sara Beery, Grant Van Horn, and Pietro Perona. Recognition in Terra Incognita. In *Computer Vision – ECCV 2018: 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XVI*, pages 472–489, Berlin, Heidelberg, September 2018. Springer-Verlag. 2

[7] Emanuel Ben-Baruch, Tal Ridnik, Itamar Friedman, Avi Ben-Cohen, Nadav Zamir, Asaf Noy, and Lihi Zelnik-Manor. Multi-label Classification with Partial Annotations using Class-aware Selective Loss. In *2022 IEEE/CVF Conference on Computer*

Vision and Pattern Recognition (CVPR), pages 4754–4762, June 2022. ISSN: 2575-7075. 11

[8] Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation Learning: A Review and New Perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1798–1828, August 2013. 2, 7

[9] Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Ben Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Rob Reich, Hongyu Ren, Frieda Rong, Yusuf Roohani, Camilo Ruiz, Jack Ryan, Christopher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishnan Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. On the Opportunities and Risks of Foundation Models. 8 2021. 2, 11

[10] Ludwig Bothmann, Lisa Wimmer, Omid Charrakh, Tobias Weber, Hendrik Edelhoff, Wibke Peters, Hien Nguyen, Caryl Benjamin, and Annette Menzel. Automated wildlife image classification: An active learning tool for ecological applications. *Ecological Informatics*, 77:102231, November 2023. 55

[11] Ellen Bowler, Peter T. Fretwell, Geoffrey French, and Michal Mackiewicz. Using Deep Learning to Count Albatrosses from Space: Assessing Results in Light of Ground

Truth Uncertainty. *Remote Sensing 2020, Vol. 12, Page 2026*, 12(12):2026, 6 2020. 1, 18

[12] Clemens-Alexander Brust, Christoph Käding, and Joachim Denzler. Active Learning for Deep Object Detection. pages 181–190, March 2026. 55

[13] Stuart H. M. Butchart, Matt Walpole, Ben Collen, Arco van Strien, Jörn P. W. Scharlemann, Rosamunde E. A. Almond, Jonathan E. M. Baillie, Bastian Bomhard, Claire Brown, John Bruno, Kent E. Carpenter, Geneviève M. Carr, Janice Chanson, Anna M. Chenery, Jorge Csirke, Nick C. Davidson, Frank Dentener, Matt Foster, Alessandro Galli, James N. Galloway, Piero Genovesi, Richard D. Gregory, Marc Hockings, Valerie Kapos, Jean-Francois Lamarque, Fiona Leverington, Jonathan Loh, Melodie A. McGeoch, Louise McRae, Anahit Minasyan, Monica Hernández Morcillo, Thomasina E. E. Oldfield, Daniel Pauly, Suhel Quader, Carmen Revenga, John R. Sauer, Benjamin Skolnik, Dian Spear, Damon Stanwell-Smith, Simon N. Stuart, Andy Symes, Megan Tierney, Tristan D. Tyrrell, Jean-Christophe Vié, and Reg Watson. Global biodiversity: indicators of recent declines. *Science*, 328(5982):1164–1168, May 2010. 1

[14] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-End Object Detection with Transformers, 5 2020. 2, 12, 13, 30, 56

[15] Graeme Caughley. Bias in Aerial Survey. *The Journal of Wildlife Management*, 38(4):921, 10 1974. 18

[16] Dominique Chabot, Seth Stapleton, and Charles M. Francis. Measuring the spectral signature of polar bears from a drone to improve their detection from space. *Biological Conservation*, 237:125–132, 9 2019. 36, 37, 39, 41, 79

[17] Katherine S Christie, Sophie L Gilbert, Casey L Brown, Michael Hatfield, and Leanne Hanson. Unmanned aircraft systems in wildlife research: current and future applications of a transformative technology. *Frontiers in Ecology and the Environment*, 14(5):241–251, 2016. _eprint: https://esajournals.onlinelibrary.wiley.com/doi/pdf/10.1002/fee.1281. 51

[18] L.-P. Chrétien, J. Théau, and P. Ménard. WILDLIFE MULTISPECIES REMOTE SENSING USING VISIBLE AND THERMAL INFRARED IMAGERY ACQUIRED FROM AN UNMANNED AERIAL VEHICLE (UAV). *The International Archives*

of the Photogrammetry, Remote Sensing and Spatial Information Sciences, XL-1-W4:241–248, August 2015. Conference Name: ISPRS UAV-g2015 (Volume XL-1/W4) - 30 August–2 September 2015, Toronto, Canada. 53

[19] Cornell Lab of Ornithology. Common Eider Identification, 2025. 43

[20] Hannah C. Cubaynes and Peter T. Fretwell. Whales from space dataset, an annotated satellite image dataset of whales for training machine learning models. *Scientific Data 2022 9:1*, 9(1):1–8, 5 2022. 18

[21] CVAT.ai Corporation. Computer Vision Annotation Tool (CVAT), November 2023. 21

[22] Alexandre Delplanque, Samuel Foucher, Philippe Lejeune, Julie Linchant, and Jérôme Théau. Multispecies detection and identification of African mammals in aerial imagery using convolutional neural networks. *Remote Sensing in Ecology and Conservation*, 8(2):166–179, 4 2022. 35, 78

[23] Alexandre Delplanque, Samuel Foucher, Jérôme Théau, Elsa Bussiere, Cédric Vermeulen, and Philippe Lejeune. From crowd to herd counting: How to precisely detect and count african mammals using aerial imagery and deep learning? *ISPRS Journal of Photogrammetry and Remote Sensing*, 197:167–180, 2023. 56

[24] Alexandre Delplanque, Samuel Foucher, Jérôme Théau, Elsa Bussière, Cédric Vermeulen, and Philippe Lejeune. From crowd to herd counting: How to precisely detect and count African mammals using aerial imagery and deep learning? *ISPRS Journal of Photogrammetry and Remote Sensing*, 197:167–180, 3 2023. 19, 35

[25] Alexandre Delplanque, Jérôme Théau, Samuel Foucher, Ghazaleh Serati, Simon Durand, and Philippe Lejeune. Wildlife detection, counting and survey using satellite imagery: are we there yet? *GIScience & Remote Sensing*, 61(1), 12 2024. 1, 18

[26] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186, 2019. 29

[27] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words:

Transformers for Image Recognition at Scale. *ICLR 2021 - 9th International Conference on Learning Representations*, 10 2020. 9, 29

[28] B. Dwyer, J. Nelson, T. Hansen, et al. Roboflow (version 1.0), 2025. Accessed: 2025-08-19. 21

[29] Jasper A.J. Eikelboom, Johan Wind, Eline van de Ven, Lekishon M. Kenana, Bradley Schroder, Henrik J. de Knegt, Frank van Langevelde, and Herbert H.T. Prins. Improving the precision and accuracy of animal population estimates with aerial image object detection. *Methods in Ecology and Evolution*, 10(11):1875–1887, 11 2019. 26, 33, 39, 40, 81

[30] S. K. Eltringham. Review of Counting Animals. *Journal of Applied Ecology*, 14(3):996–997, 1977. 17, 18

[31] Chuanxing Geng, Sheng-Jun Huang, and Songcan Chen. Recent Advances in Open Set Recognition: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(10):3614–3631, October 2021. 2, 9

[32] B. C. Gonçalves, B. Spitzbart, and H. J. Lynch. SealNet: A fully-automated pack-ice seal detection pipeline for sub-meter satellite imagery. *Remote Sensing of Environment*, 239:111617, 3 2020. 19

[33] Patrick C. Gray, Abram B. Fleishman, David J. Klein, Matthew W. McKown, Vanessa S. Bézy, Kenneth J. Lohmann, and David W. Johnston. A convolutional neural network for detecting sea turtles in drone imagery. *Methods in Ecology and Evolution*, 10(3):345–355, 3 2019. 35, 76

[34] Sajjad Reyhani Haghighi, Mikaeel Pasandideh Saqalaksari, and Scott N. Johnson. Artificial Intelligence in Ecology: A Commentary on a Chatbot's Perspective. *The Bulletin of the Ecological Society of America*, 104(4):e2097, 10 2023. 18

[35] Liang Han, Pin Tao, and Ralph R. Martin. Livestock detection in aerial images using a fully convolutional network. *Computational Visual Media*, 5(2):221–228, 6 2019. 19, 36, 37, 38, 79

[36] Ang He, Xiaobo Li, Ximei Wu, Chengyue Su, Jing Chen, Sheng Xu, and Xiaobin Guo. ALSS-YOLO: An Adaptive Lightweight Channel Split and Shuffling Network for TIR Wildlife Detection in UAV Imagery. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 17:17308–17326, 2024. 53

[37] Joel Heath. People of a Feather, May 2011. 42

[38] Jarrod C. Hodgson, Shane M. Baylis, Rowan Mott, Ashley Herrod, and Rohan H. Clarke. Precision wildlife monitoring using unmanned aerial vehicles. *Scientific Reports 2016 6:1*, 6(1):1–7, 3 2016. 19

[39] Jarrod C. Hodgson, Rowan Mott, Shane M. Baylis, Trung T. Pham, Simon Wotherspoon, Adam D. Kilpatrick, Ramesh Raja Segaran, Ian Reid, Aleks Terauds, and Lian Pin Koh. Drones count wildlife more accurately and precisely than humans. *Methods in Ecology and Evolution*, 9(5):1160–1167, 2018. _eprint: https://besjournals.onlinelibrary.wiley.com/doi/pdf/10.1111/2041-210X.12974. 51

[40] Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5):359–366, January 1989. 7

[41] Jayden Hsiao, Aryan Kalia, Zhonghao Zhang, Hudson Sun, Muhammed Patel, David A. Clausi, Lincoln Linlin Xu, Becky Segal, and Joel Heath. Human-in-the-Loop Eider Duck Counting in Arctic Canada with an Open-Vocabulary Multi-Species Wildlife Detector. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(47):39978–39986, March 2026. iii

[42] Jayden Hsiao, Aryan Kalia, Zhonghao Zhang, Hudson Sun, Muhammed Patel, David A. Clausi, Linlin Xu, Becky Segal, and Joel Heath. Human-in-the-loop eider duck counting in arctic canada with an open-vocabulary multi-species wildlife detector. In *Proceedings of the AAAI Conference on Artificial Intelligence: Innovative Applications of Artificial Intelligence (IAAI)*, 2026. Accepted to IAAI 2026 (to appear). iii

[43] Jayden Hsiao, Muhammed Patel, David A. Clausi, and Linlin Xu. Dinokey: Transformer-based keypoint detection for wildlife monitoring in aerial imagery. In *Proceedings of the International Society for Photogrammetry and Remote Sensing (ISPRS)*, 2026. Accepted to ISPRS 2026 (to appear). iii, 55

[44] H. Jachmann. Comparison of aerial counts with ground counts for large African herbivores. *Journal of Applied Ecology*, 39(5):841–852, 10 2002. 19

[45] Ylber Januzaj and Artan Luma. Cosine Similarity – A Computing Approach to Match Similarity Between Higher Education Programs and Job Market Demands Based on Maximum Number of Common Words. *International Journal of Emerging Technologies in Learning (iJET)*, 17(12):258–268, June 2022. 10

[46] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A.A. Kohl, Andrew J. Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W. Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. Highly accurate protein structure prediction with AlphaFold. *Nature 2021 596:7873*, 596(7873):583–589, 7 2021. 11

[47] Suvarna Kadam and Vinay Vaidya. Review and Analysis of Zero, One and Few Shot Learning Approaches. In Ajith Abraham, Aswani Kumar Cherukuri, Patricia Melin, and Niketa Gandhi, editors, *Intelligent Systems Design and Applications*, pages 100–112, Cham, 2020. Springer International Publishing. 11

[48] Kaggle. NOAA Fisheries Steller Sea Lion Population Count — Kaggle, 6 2017. 26, 33, 35, 77

[49] Chieh-Chi Kao, Teng-Yok Lee, Pradeep Sen, and Ming-Yu Liu. Localization-Aware Active Learning for Object Detection. In C.V. Jawahar, Hongdong Li, Greg Mori, and Konrad Schindler, editors, *Computer Vision – ACCV 2018*, pages 506–522, Cham, 2019. Springer International Publishing. 55

[50] Benjamin Kellenberger, Diego Marcos, Sylvain Lobry, and Devis Tuia. Half a Percent of Labels is Enough: Efficient Animal Detection in UAV Imagery Using Deep CNNs and Active Learning. *IEEE Transactions on Geoscience and Remote Sensing*, 57(12):9524–9533, December 2019. 55

[51] Benjamin Kellenberger, Diego Marcos, and Devis Tuia. Detecting mammals in UAV images: Best practices to address a substantially imbalanced dataset with deep learning. *Remote Sensing of Environment*, 216:139–153, October 2018. 2

[52] Benjamin Kellenberger, Diego Marcos, and Devis Tuia. Detecting mammals in UAV images: Best practices to address a substantially imbalanced dataset with deep learning. *Remote Sensing of Environment*, 216:139–153, 10 2018. 19

[53] Benjamin Kellenberger, Thor Veen, Eelke Folmer, and Devis Tuia. 21 000 birds in 4.5 h: efficient large-scale seabird detection with machine learning. *Remote Sensing in Ecology and Conservation*, 7(3):445–460, 9 2021. 26, 36, 37, 80

[54] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4015–4026, 2023. 34

[55] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. ImageNet classification with deep convolutional neural networks. *Commun. ACM*, 60(6):84–90, May 2017. 8

[56] Jan Kukacka, Vladimir Golkov, and Daniel Cremers. Regularization for Deep Learning: A Taxonomy. October 2017. 7

[57] Todd Kulesza, Margaret Burnett, Weng-Keen Wong, and Simone Stumpf. Principles of Explanatory Debugging to Personalize Interactive Machine Learning. In *Proceedings of the 20th International Conference on Intelligent User Interfaces*, IUI '15, pages 126–137, New York, NY, USA, March 2015. Association for Computing Machinery. 2, 20

[58] Satish Kumar, Bowen Zhang, Chandrakanth Gudavalli, Connor Levenson, Lacey Hughey, Jared A Stabach, Irene Amoke, Gordon Ojwang, Joseph Mukeka, Stephen Mwiu, Joseph Ogutu, Howard Frederick, and B.S. Manjunath. WildlifeMapper : Aerial Image Analysis for Multi-Species Detection and Identification. *IEEE Xplore*, pages 12594–12604, 2024. 19, 34, 35

[59] Marc Kéry and Michael Schaub. *Bayesian Population Analysis using WinBUGS: A hierarchical perspective*. Academic Press, November 2011. 51

[60] Richard Lamprey, David Ochanda, Rob Brett, Charles Tumwesigye, and Iain Douglas-Hamilton. Cameras replace human observers in multi-species aerial counts in Murchison Falls, Uganda. *Remote Sensing in Ecology and Conservation*, 6(4):529–545, 12 2020. 1, 18

[61] Richard Lamprey, Frank Pope, Shadrack Ngene, Michael Norton-Griffiths, Howard Frederick, Benson Okita-Ouma, and Iain Douglas-Hamilton. Comparing an automated high-definition oblique camera system to rear-seat-observers in a wildlife survey in Tsavo, Kenya: Taking multi-species aerial counts to the next level. *Biological Conservation*, 241:108243, 1 2020. 17, 18

[62] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, November 1998. 8

[63] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, May 2015. 1, 7, 8

[64] Seunghyeon Lee, Youngkeun Song, and Sung-Ho Kil. Feasibility Analyses of Real-Time Detection of Wildlife Using UAV-Derived Thermal and RGB Images. *Remote Sensing*, 13:2169, June 2021. ADS Bibcode: 2021RemS...13.2169L. 53

[65] Daniel L. Leedy. Aerial Photographs, Their Interpretation and Suggested Uses in Wildlife Management. *The Journal of Wildlife Management*, 12(2):191, 4 1948. 1, 18

[66] Euridice Leyequien, Jochem Verrelst, Martijn Slot, Gabriela Schaepman-Strub, Ignas M.A. Heitkönig, and Andrew Skidmore. Capturing the fugitive: Applying remote sensing to terrestrial animal distribution and diversity. *International Journal of Applied Earth Observation and Geoinformation*, 9(1):1–20, 2 2007. 1, 18

[67] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pages 12888–12900. PMLR, 2022. 7

[68] Ying Li, Binbin Fan, Weiping Zhang, Weiping Ding, and Jianwei Yin. Deep active learning for object detection. *Information Sciences*, 579:418–433, November 2021. 55

[69] Yunpeng Li, Chengjin Tao, Meng Liu, Xiangrong Zhang, Guanchun Wang, Tianyang Zhang, Dong Zhao, and Dabao Wang. Feature refinement and rethinking attention for remote sensing image captioning. *Scientific Reports*, 15(1):8742, March 2025. 54

[70] Zongxia Li, Xiyang Wu, Hongyang Du, Fuxiao Liu, Huy Nghiem, and Guangyao Shi. A Survey of State of the Art Large Vision Language Models: Alignment, Benchmark, Evaluations and Challenges, April 2025. arXiv:2501.02189 [cs]. 54

[71] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal Loss for Dense Object Detection, February 2020. 14, 30

[72] 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 *European conference on computer vision*, pages 740–755. Springer, 2014. 16

[73] Julie Linchant, Jonathan Lisein, Jean Semeki, Philippe Lejeune, and Cédric Vermeulen. Are unmanned aircraft systems (UASs) the future of wildlife monitoring? A review of accomplishments and challenges. 45(4):239–252. 1, 17

[74] Jiang Liu, Chenqiang Gao, Deyu Meng, and Alexander G Hauptmann. Decidenet: Counting varying density crowds through attention guided detection and density estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5197–5206, 2018. 34

[75] Li Liu, Wanli Ouyang, Xiaogang Wang, Paul Fieguth, Jie Chen, Xinwang Liu, and Matti Pietikäinen. Deep Learning for Generic Object Detection: A Survey. *International Journal of Computer Vision*, 128(2):261–318, February 2020. 1, 12

[76] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding DINO: Marrying DINO with Grounded Pre-Training for Open-Set Object Detection, July 2024. arXiv:2303.05499 [cs]. 13

[77] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C. Berg. SSD: Single Shot MultiBox Detector. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision – ECCV 2016*, pages 21–37, Cham, 2016. Springer International Publishing. 12

[78] Yang Liu, Vikrant Shah, Alexander Borowicz, Michael Wethington, Noah Strycker, Steve Forrest, Heather Lynch, and Hanumant Singh. Towards Efficient Machine Learning Methods for Penguin Counting in Unmanned Aerial System Imagery. *2020 IEEE/OES Autonomous Underwater Vehicles Symposium, AUV 2020*, 9 2020. 39, 40, 81

[79] Yen-Cheng Liu, Chih-Yao Ma, and Zsolt Kira. Unbiased Teacher v2: Semi-Supervised Object Detection for Anchor-Free and Anchor-Based Detectors. pages 9819–9828, 2022. 11

[80] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 9, 29

[81] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. ViLBERT: pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Proceedings*

of the 33rd International Conference on Neural Information Processing Systems, number 2, pages 13–23. Curran Associates Inc., Red Hook, NY, USA, December 2019. 7, 9

[82] Mitchell B. Lyons, Kate J. Brandis, Nicholas J. Murray, John H. Wilshire, Justin A. McCann, Richard T. Kingsford, and Corey T. Callaghan. Monitoring large and complex wildlife aggregations with drones. *Methods in Ecology and Evolution*, 10(7):1024–1035, 2019. _eprint: https://besjournals.onlinelibrary.wiley.com/doi/pdf/10.1111/2041-210X.13194. 52

[83] Anne E. Magurran. Measuring biological diversity. *Current Biology*, 31(19):R1174–R1177, October 2021. 1

[84] Marianne Marcoux, Leah M Montsion, J Blair Dunn, Steven H Ferguson, and Cory J D Matthews. Canadian Science Advisory Secretariat (CSAS) Estimate of the abundance of the Eclipse Sound narwhal (Monodon Monoceros) summer stock from the 2016 photographic aerial survey. Technical report, 2019. 36, 37, 79

[85] Marianne Marcoux, Brent G Young, Natalie C Asselin, Cortney A Watt, J Blair Dunn, Steven H Ferguson, and Oceans Canada. Estimate of Cumberland Sound beluga (Delphinapterus leucas) population size from the 2014 visual and photographic aerial survey. Technical report. 33, 35, 76

[86] Cory J D Matthews, Cortney A Watt, Natalie C Asselin, J Blair Dunn, Brent G Young, Leah M Montsion, Kristin H Westdal, Patricia A Hall, Jack R Orr, Steven H Ferguson, and Marianne Marcoux. Canadian Science Advisory Secretariat (CSAS) Estimated abundance of the Western Hudson Bay beluga stock from the 2015 visual and photographic aerial survey. Technical report, 2017. 35, 76

[87] Giacomo May, Emanuele Dalsasso, Benjamin Kellenberger, and Devis Tuia. POLO – Point-based, multi-class animal detection, April 2025. arXiv:2410.11741 [cs]. 56

[88] Michael McCloskey and Neal J. Cohen. Catastrophic Interference in Connectionist Networks: The Sequential Learning Problem. 24:109–165, 1989. Book Title: Psychology of Learning and Motivation ISBN: 9780125433242. 20

[89] Martial Mermillod, Aurélia Bugaiska, and Patrick Bonin. The stability-plasticity dilemma: investigating the continuum from catastrophic forgetting to age-limited learning effects. *Frontiers in Psychology*, 4, August 2013. 20

[90] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient Estimation of Word Representations in Vector Space, September 2013. arXiv:1301.3781 [cs]. 29

[91] Millennium Ecosystem Assessment. *Ecosystems and Human Well-being: Synthesis*. Island Press, Washington, DC, 2005. 1

[92] Chao Mou, Tengfei Liu, Chengcheng Zhu, and Xiaohui Cui. WAID: A Large-Scale Dataset for Wildlife Detection with Drones. *Applied Sciences 2023, Vol. 13, Page 10397*, 13(18):10397, 9 2023. 36, 37, 80

[93] Robert Munro. Crowdsourcing and the crisis-affected community: Lessons learned and looking forward from Mission 4636. *Inf. Retr.*, 16(2):210–266, April 2013. 2

[94] Johannes Naude and Deon Joubert. The Aerial Elephant Dataset: A New Public Benchmark for Aerial Object Detection., 2019. 35, 76

[95] A. Neubeck and L. Van Gool. Efficient Non-Maximum Suppression. In *18th International Conference on Pattern Recognition (ICPR'06)*, volume 3, pages 850–855, August 2006. ISSN: 1051-4651. 12

[96] Mohammad Sadegh Norouzzadeh, Anh Nguyen, Margaret Kosmala, Alexandra Swanson, Meredith S. Palmer, Craig Packer, and Jeff Clune. Automatically identifying, counting, and describing wild animals in camera-trap images with deep learning. *Proceedings of the National Academy of Sciences of the United States of America*, 115(25):E5716–E5725, 6 2018. 19

[97] OpenAI. GPT-4o mini: advancing cost-efficient intelligence — OpenAI, 7 2024. 32

[98] Muhammed Patel, Xinwei Chen, Linlin Xu, Fernando J. Pena Cantu, Javier Noa Turnes, Neil C. Brubacher, David A. Clausi, and K. Andrea Scott. The Influence of Input Image Scale on Deep Learning-Based Beluga Whale Detection from Aerial Remote Sensing Imagery. *International Geoscience and Remote Sensing Symposium (IGARSS)*, 2023-July:5732–5734, 2023. 19

[99] Muhammed Patel, Javier Noa Turnes, Jayden Hsiao, Linlin Xu, and David Clausi. OpenWildlife: Open-Vocabulary Multi-Species Wildlife Detector for Geographically-Diverse Aerial Imagery, June 2025. arXiv:2506.19204 [cs]. iii

[100] Jeffrey Pennington, Richard Socher, and Christopher Manning. GloVe: Global Vectors for Word Representation. In Alessandro Moschitti, Bo Pang, and Walter Daelemans, editors, *Proceedings of the 2014 Conference on Empirical Methods in Natural*

Language Processing (EMNLP), pages 1532–1543, Doha, Qatar, October 2014. Association for Computational Linguistics. 29

[101] Yifei Qian, Grant R.W. Humphries, Philip N. Trathan, Andrew Lowther, and Carl R. Donovan. Counting animals in aerial images with a density map estimation model. *Ecology and Evolution*, 13(4):e9903, 4 2023. 19, 34, 35, 56, 78

[102] Fang Qingyun and Wang Zhaokui. Cross-modality attentive feature fusion for object detection in multispectral remote sensing imagery. *Pattern Recognition*, 130:108786, October 2022. 53

[103] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 2, 7

[104] Jason I. Ransom. Detection probability in aerial surveys of feral horses. *The Journal of Wildlife Management*, 76(2):299–307, 2012. _eprint: https://wildlife.onlinelibrary.wiley.com/doi/pdf/10.1002/jwmg.204. 52

[105] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6):1137–1149, 6 2017. 12

[106] Sam A. Reynolds, Sara Beery, Neil Burgess, Mark Burgman, Stuart H. M. Butchart, Steven J. Cooke, David Coomes, Finn Danielsen, Enrico Di Minin, América Paz Durán, Francis Gassert, Amy Hinsley, Sadiq Jaffer, Julia P. G. Jones, Binbin V. Li, Oisin Mac Aodha, Anil Madhavapeddy, Stephanie A. L. O'Donnell, William M. Oxbury, Lloyd Peck, Nathalie Pettorelli, Jon Paul Rodríguez, Emily Shuckburgh, Bernardo Strassburg, Hiromi Yamashita, Zhongqi Miao, and William J. Sutherland. The potential for AI to revolutionize conservation: A horizon scan. *Trends in Ecology & Evolution*, 40(2):191–207, February 2025. 20

[107] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized Intersection over Union: A Metric and A Loss for Bounding Box Regression. 14, 30

[108] Natasha A. Ridenour, Xianmin Hu, Shabnam Jafarikhasragh, Jack C. Landy, Jennifer V. Lukovich, Tricia A. Stadnyk, Kevin Sydor, Paul G. Myers, and David G.

Barber. Sensitivity of freshwater dynamics to ocean model resolution and river discharge forcing in the Hudson Bay Complex. *Journal of Marine Systems*, 196:48–64, August 2019. 42

[109] Alberto Rivas, Pablo Chamoso, Alfonso González-Briones, and Juan Manuel Corchado. Detection of Cattle Using Drones and Convolutional Neural Networks. *Sensors*, 18(7):2048, July 2018. 2

[110] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 9351:234–241, 2015. 19

[111] J. Andrew Royle. N-Mixture Models for Estimating Population Size from Spatially Replicated Counts. *Biometrics*, 60(1):108–115, 2004. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.0006-341X.2004.00142.x. 51

[112] J. Andrew Royle and Robert M. Dorazio. *Hierarchical Modeling and Inference in Ecology: The Analysis of Data from Populations, Metapopulations and Communities*. Academic Press, September 2008. 51

[113] David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. Learning representations by back-propagating errors. *Nature*, 323(6088):533–536, October 1986. 6

[114] Walter J. Scheirer, Anderson de Rezende Rocha, Archana Sapkota, and Terrance E. Boult. Toward Open Set Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(7):1757–1772, July 2013. 2, 9

[115] Burr Settles. Active Learning Literature Survey. Technical Report, University of Wisconsin-Madison Department of Computer Sciences, 2009. Accepted: 2012-03-15T17:23:56Z. 2, 20, 54

[116] A. C. Seymour, J. Dale, M. Hammill, P. N. Halpin, and D. W. Johnston. Automated detection and enumeration of marine wildlife using unmanned aircraft systems (UAS) and thermal imagery. *Scientific Reports*, 7:45127, March 2017. 2

[117] Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. Objects365: A Large-Scale, High-Quality Dataset for Object Detection. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 8429–8438. IEEE, 10 2019. 28

[118] Andrew K. Skidmore, Nicholas C. Coops, Elnaz Neinavaz, Abebe Ali, Michael E. Schaepman, Marc Paganini, W. Daniel Kissling, Petteri Vihervaara, Roshanak Darvishzadeh, Hannes Feilhauer, Miguel Fernandez, Néstor Fernández, Noel Gorelick, Ilse Geijzendorffer, Uta Heiden, Marco Heurich, Donald Hobern, Stefanie Holzwarth, Frank E. Muller-Karger, Ruben Van De Kerchove, Angela Lausch, Pedro J. Leitão, Marcelle C. Lock, Caspar A. Mücher, Brian O'Connor, Duccio Rocchini, Woody Turner, Jan Kees Vis, Tiejun Wang, Martin Wegmann, and Vladimir Wingate. Priority list of biodiversity metrics to observe from space. *Nature Ecology & Evolution 2021 5:7*, 5(7):896–906, 5 2021. 1, 18

[119] Samuel Stevens, Jiaman Wu, Matthew J Thompson, Elizabeth G Campolongo, Chan Hee Song, David Edward Carlyn, Li Dong, Wasila M Dahdul, Charles Stewart, Tanya Berger-Wolf, Wei-Lun Chao, and Yu Su. BioCLIP: A Vision Foundation Model for the Tree of Life. 2023. 17

[120] Shashi Thota, Subrahmanyasarma Chitta, Vinay Kumar Reddy Vangoor, Chetan Sasidhar Ravi, and Venkata Sri Manoj Bonam. Few-Shot Learning in Computer Vision: Practical Applications and Techniques. *Human-Computer Interaction Perspectives*, 3(1):29–59, June 2023. 10

[121] Zhi Tian, Chunhua Shen, Hao Chen, and Tong He. FCOS: Fully Convolutional One-Stage Object Detection. pages 9627–9636, 2019. 12

[122] Maxim Tkachenko, Mikhail Malyuk, Andrey Holmanyuk, and Nikolai Liubimov. Label Studio: Data labeling software, 2020-2025. Open source software available from https://github.com/HumanSignal/label-studio. 21, 43

[123] Colin J. Torney, David J. Lloyd-Jones, Mark Chevallier, David C. Moyer, Honori T. Maliti, Machoke Mwita, Edward M. Kohi, and Grant C. Hopcraft. A comparison of deep learning and citizen science techniques for counting wildlife in aerial survey images. *Methods in Ecology and Evolution*, 10(6):779–787, 6 2019. 19

[124] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is All you Need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. 9

[125] Jiaqi Wang, Pan Zhang, Tao Chu, Yuhang Cao, Yujie Zhou, Tong Wu, Bin Wang, Conghui He, and Dahua Lin. V3Det: Vast Vocabulary Visual Detection Dataset. 4 2023. 28

[126] Qi Wang, Wei Huang, Xueting Zhang, and Xuelong Li. GLCM: Global–Local Captioning Model for Remote Sensing Image Captioning. *IEEE Transactions on Cybernetics*, 53(11):6910–6922, November 2023. 54

[127] Shuang Wang, Xiutiao Ye, Yu Gu, Jihui Wang, Yun Meng, Jingxian Tian, Biao Hou, and Licheng Jiao. Multi-label semantic feature fusion for remote sensing image captioning. *ISPRS Journal of Photogrammetry and Remote Sensing*, 184:1–18, February 2022. 54

[128] Cortney A.. Watt. Abundance and total allowable landed catch estimates from the 2017 aerial survey of the Cumberland Sound beluga (Delphinapterus leucas) population. page 33, 2021. 35, 76

[129] Ben G Weinstein, Lindsey Garner, Vienna R Saccomanno, Ashley Steinkraus, Andrew Ortega, Kristen Brush, Glenda Yenni, Ann E McKellar, Rowan Converse, Christopher D Lippitt, et al. A general deep learning model for bird detection in high-resolution airborne imagery. *Ecological Applications*, 32(8):e2694, 2022. 56

[130] Ben G. Weinstein, Lindsey Garner, Vienna R. Saccomanno, Ashley Steinkraus, Andrew Ortega, Kristen Brush, Glenda Yenni, Ann E. McKellar, Rowan Converse, Christopher D. Lippitt, Alex Wegmann, Nick D. Holmes, Alice J. Edney, Tom Hart, Mark J. Jessopp, Rohan H Clarke, Dominik Marchowski, Henry Senyondo, Ryan Dotson, Ethan P. White, Peter Frederick, and S.K. Morgan Ernest. A general deep learning model for bird detection in high resolution airborne imagery. *bioRxiv*, page 2021.08.05.455311, 3 2022. 35, 36, 37, 39, 41, 77, 79

[131] Emily L. Weiser, Paul L. Flint, Dennis K. Marks, Brad S. Shults, Heather M. Wilson, Sarah J. Thompson, and Julian B. Fischer. Optimizing surveys of fall-staging geese using aerial imagery and automated counting. *Wildlife Society Bulletin*, 47(1), 3 2023. 35, 77

[132] Zijing Wu, Ce Zhang, Xiaowei Gu, Isla Duporge, Lacey F. Hughey, Jared A. Stabach, Andrew K. Skidmore, J. Grant C. Hopcraft, Stephen J. Lee, Peter M. Atkinson, Douglas J. McCauley, Richard Lamprey, Shadrack Ngene, and Tiejun Wang. Deep learning enables satellite-based monitoring of large populations of terrestrial mammals across heterogeneous landscape. *Nature Communications 2023 14:1*, 14(1):1–15, 5 2023. 19

[133] Yongqin Xian, Christoph H. Lampert, Bernt Schiele, and Zeynep Akata. Zero-Shot Learning—A Comprehensive Evaluation of the Good, the Bad and the Ugly.

IEEE Transactions on Pattern Analysis and Machine Intelligence, 41(9):2251–2265, September 2019. 11

[134] Zhengyuan Yang, Zhe Gan, Jianfeng Wang, Xiaowei Hu, Faisal Ahmed, Zicheng Liu, Yumao Lu, and Lijuan Wang. UniTAB: Unifying Text and Box Outputs for Grounded Vision-Language Modeling. In Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, editors, *Computer Vision – ECCV 2022*, pages 521–539, Cham, 2022. Springer Nature Switzerland. 54

[135] Muhammad Yaseen. What is YOLOv8: An In-Depth Exploration of the Internal Features of the Next-Generation Object Detector, August 2024. arXiv:2408.15857 [cs]. 32, 33, 34

[136] Alireza Zareian, Kevin Dela Rosa, Derek Hao Hu, and Shih-Fu Chang. Open-Vocabulary Object Detection Using Captions. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14388–14397, June 2021. ISSN: 2575-7075. 2, 10

[137] 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. 32, 33, 34, 56

[138] Haotian Zhang, Pengchuan Zhang, Xiaowei Hu, Yen-Chun Chen, Liunian Harold Li, Xiyang Dai, Lijuan Wang, Lu Yuan, Jenq-Neng Hwang, and Jianfeng Gao. GLIPv2: unifying localization and VL understanding. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, pages 36067–36080, Red Hook, NY, USA, November 2022. Curran Associates Inc. 54

[139] Kai Zhao and Wei Xiong. Exploring region features in remote sensing image captioning. *International Journal of Applied Earth Observation and Geoinformation*, 127:103672, March 2024. 54

[140] Pujie Zhao, Xia Ye, and Ziang Du. Object Detection in Multispectral Remote Sensing Images Based on Cross-Modal Cross-Attention. *Sensors*, 24(13):4098, January 2024. 53

[141] Xiangyu Zhao, Yicheng Chen, Shilin Xu, Xiangtai Li, Xinjiang Wang, Yining Li, and Haian Huang. An Open and Comprehensive Pipeline for Unified Object Grounding and Detection. *arXiv preprint arXiv:2401.02361*, 1 2024. 27, 28, 30, 31, 33, 34

[142] Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl. Objects as Points. *ArXiv*, April 2019. 12

[143] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable DETR: Deformable Transformers for End-to-End Object Detection, March 2021. arXiv:2010.04159 [cs]. 13, 29

[144] Zhuofan Zong, Guanglu Song, and Yu Liu. Detrs with collaborative hybrid assignments training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6748–6758, 2023. 32, 33, 34

Appendices

Appendix A

Dataset Metadata

Table A.1: The 14 datasets used for open-set continuous fine-tuning. "Label(s)" refers to the class annotations provided with the bounding boxes, while "Reported Taxon" denotes any additional taxonomic information reported in the original source publications, when available. A dash (—) indicates that no additional taxonomic information beyond the labels was reported.

Name	**Label(s)**	**Reported Taxon**	**Region**	**# Train (Img, Ann)**	**# Test (Img, Ann)**
AED [94]	Elephant	African bush elephant	South Africa, Botswana, Zambia	(1,635, 12,611)	(439, 2,970)
DFOW14 [85]	Beluga whale	—	Cumberland Sound, Canada	(476, 1,935)	(244, 507)
DFOW15 [86]	Beluga whale	—	West Hudson Bay, Canada	(278, 3,717)	(186, 8,000)
DFOW17 [128]	Beluga whale	—	Cumberland Sound Bay, Nunavut, Canada	(666, 5,181)	(156, 608)
Gray [33]	Turtle	Olive ridley sea turtle	Ostional, Costa Rica	(847, 1,673)	(212, 488)

Table A.1 continued from previous page

Name	**Label(s)**	**Reported Taxon**	**Region**	**# Train (Img, Ann)**	**# Test (Img, Ann)**
IndOcean [130]	Bird	Lesser frigatebird, Red-footed booby	Pulu Keeling Island and Christmas Island, Oceania	(1,898, 9,845)	(89, 326)
Izembek [131]	Brant, Canada goose, Emperor goose	Black brant, Gray-bellied brant, Cackling goose, Emperor goose, Duck	Izembek Lagoon, Alaska, USA	(7,000, 428,000)	(1,892, 92,512)
NABird [130]	Bird	Black-legged kittiwake, Common murre, Thick-billed murre	North Atlantic	(6,269, 124,000)	(124, 7,137)
NewMex [130]	Bird	Sandhill crane, Canada goose, Mallard, Northern pintail	New Mexico, USA	(185, 4,333)	(31, 282)
NOAA [48]	Sea lion	Steller sea lion	Western Aleutian Islands, USA	(579, 50,000)	(179, 17,121)
Palmyra [130]	Bird	Red-footed booby	Palmyra Atoll, South Pacific	(298, 1,315)	(154, 454)
Pfeifer [130]	Penguin	Chinstrap penguin, Antarctic shag	Nelson and King George Islands, Antarctica	(843, 43,000)	(48, 2,687)

Table A.1 continued from previous page

Name	**Label(s)**	**Reported Taxon**	**Region**	**# Train (Img, Ann)**	**# Test (Img, Ann)**
Qian [101]	Penguin	Brush-tailed penguin	Antarctic Peninsula and South Shetland Islands	(605, 104,000)	(133, 30,408)
Virunga-Gar. [22]	Alcelaphinae, Buffalo, Kob, Warthog, Waterbuck, Elephant, Other	Buffalo, Giraffe, Hartebeest, Hippopotamus, Kob, Topi, Warthog, Waterbuck	Democratic Republic of Congo	(928, 6,962)	(258, 2,299)
Total			**(23,408, 810,492)**	**(4,276, 166,936)**	

Table A.2: The seven datasets used for testing on novel detection categories. "Label(s)" refers to the class annotations provided with the bounding boxes, while "Reported Taxon" denotes any additional taxonomic information reported in the original source publications, when available. A dash (—) indicates that no additional taxonomic information beyond the labels was reported.

Name	**Label(s)**	**Reported Taxon**	**Region**	**# Train (Img, Ann)**	**# Test (Img, Ann)**
AFSC [2]	Bearded seal, Ringed seal	—	Alaska, USA	(3,290, 11,000)	(823, 2,941)
Chabot [16]	Polar bear	—	Churchill, Manitoba, Canada	(5, 7)	(14, 21)
DFOW16 [84]	Narwhal	—	Eclipse Sound, Canada	(233, 47,459)	(441, 948)
Han [35]	Yak, Sheep	—	Asia	(61, 3,685)	(17, 665)
Michigan [130]	Gull	Mallard, American white pelican, Gull	Lake Michigan, USA	(4,850, 40,000)	(647, 6,198)

Table A.2 continued from previous page

Name	**Label(s)**	**Reported Taxon**	**Region**	**# Train (Img, Ann)**	**# Test (Img, Ann)**
WAfrica [53]	Royal tern, Caspian tern, Slender-billed gull, Gray headed gull, Great cormorant, Great white pelican	—	West Africa	(167, 18,000)	(43, 2,637)
WAID [92]	Sheep, Cattle, Seal, Camel, Kiang, Zebra	—	Various Locations, Online	(10,000, 163,000)	(1,437, 23,820)
Total			**(18,606, 283,151)**	**(3,422, 37,230)**	

Table A.3: The three datasets used for evaluation in novel environments. "Label(s)" refers to the class annotations provided with the bounding boxes, while "Reported Taxon" denotes any additional taxonomic information reported in the original source publications, when available. A dash (—) indicates that no additional taxonomic information beyond the labels was reported. Note that DFOW23 was obtained through an internal collaboration with Fisheries and Oceans Canada (DFO) and is not publicly available, hence no citation is provided.

Name	**Label(s)**	**Reported Taxon**	**Region**	**# Train (Img, Ann)**	**# Test (Img, Ann)**
DFOW23	Narwhal, Beluga whale, Bowhead whale	—	North America	(143, 288)	(343, 699)
Kenya [29]	Zebra, Elephant, Giraffe	—	Kenya	(393, 3,000)	(112, 850)
Liu [78]	Penguin	Chinstrap penguin	Antarctic Peninsula, Antarctica (Elephant, Low, Snow, and Anvers Islands)	(172, 2,019)	(63, 1,504)
Total			**(708, 5,307)**	**(518, 3,053)**	

Appendix B

Human-in-the-loop Interface Details

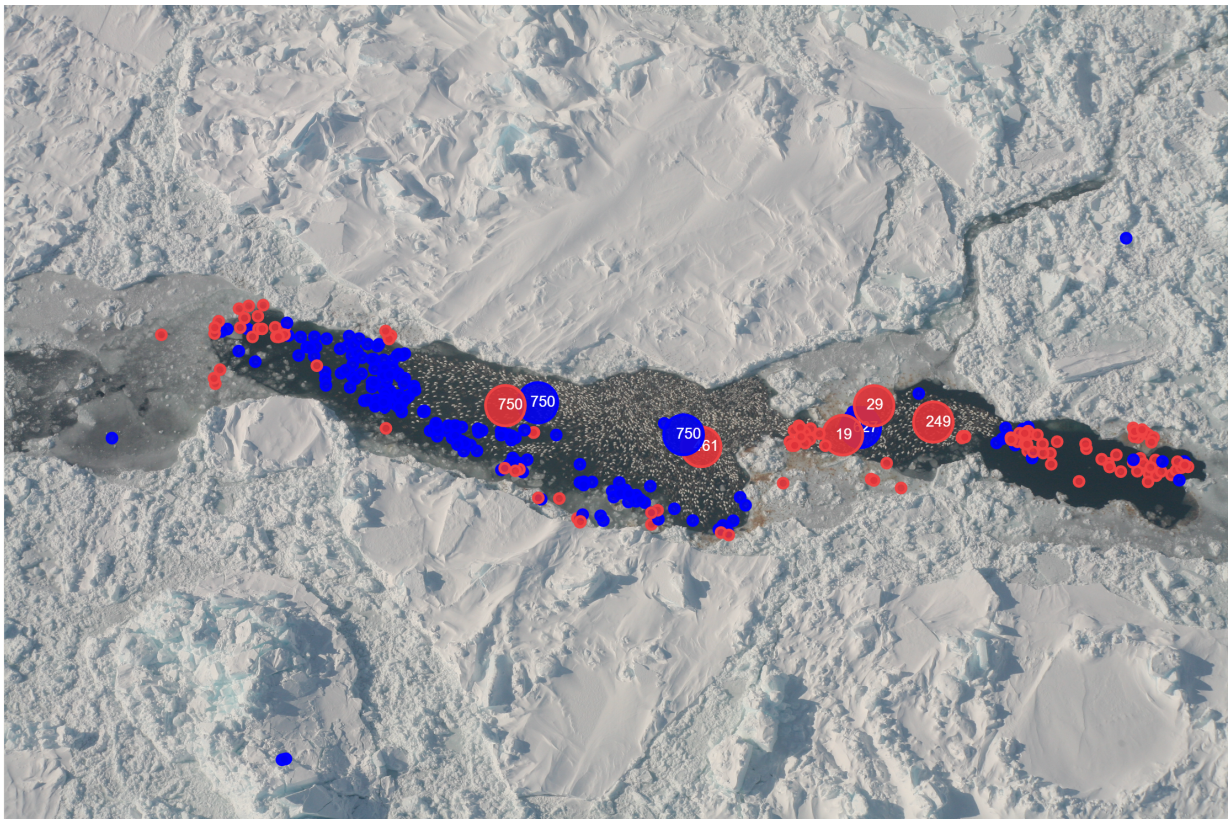


Figure B.1: Example of the bubble clustering algorithm used to improve UI performance.

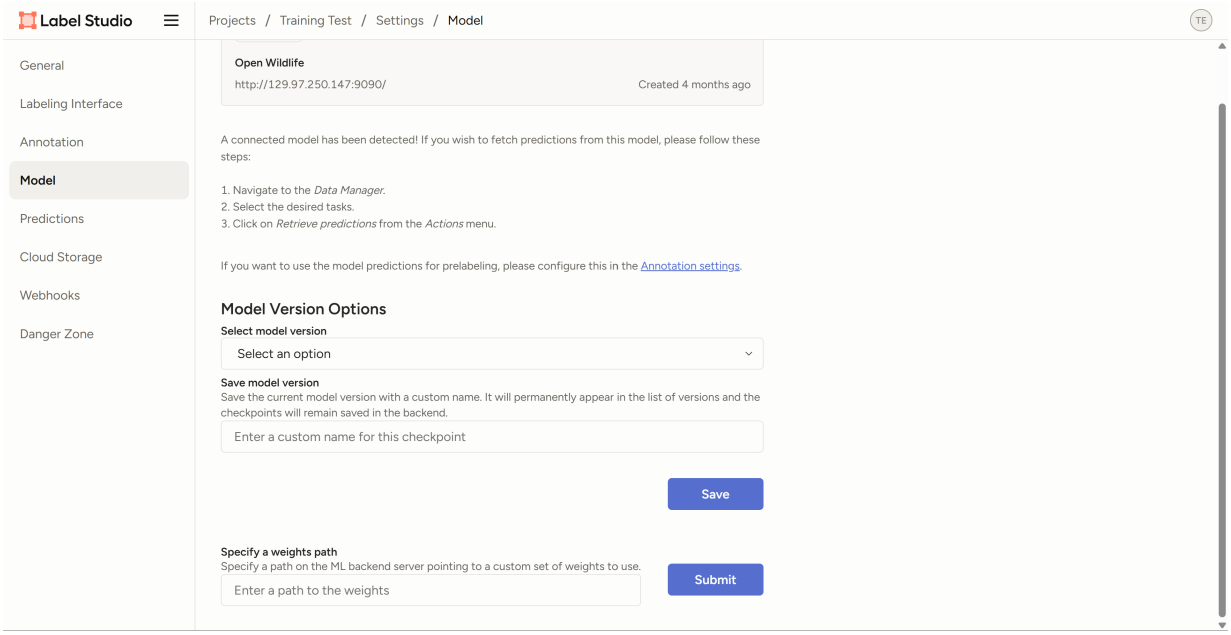


Figure B.2: Enhanced model settings page, containing options for model versioning.

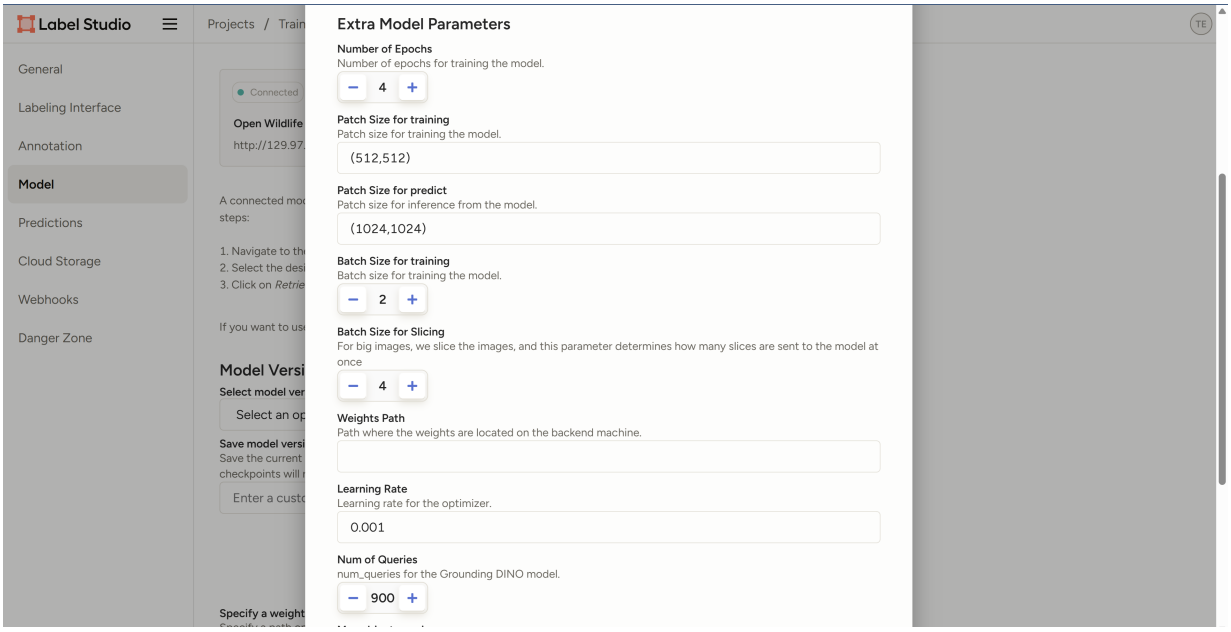


Figure B.3: Settings dialog to edit model training and testing parameters.

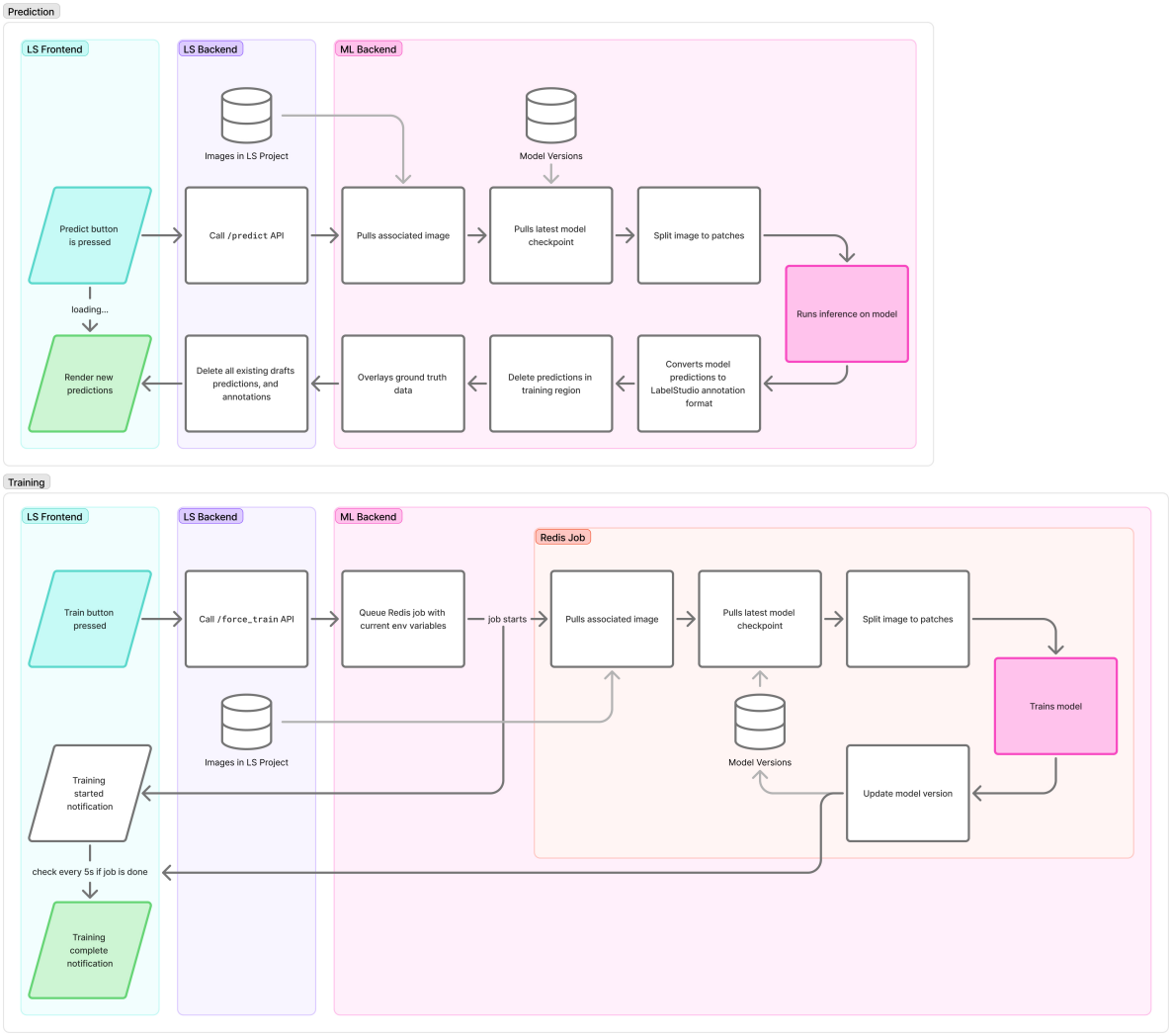


Figure B.4: Software design diagram of prediction and training workflows across front-end, back-end, and machine learning back-end systems.