Human-in-the-Loop Eider Duck Counting in Arctic Canada with an Open-Vocabulary Multi-Species Wildlife Detector

Jayden Hsiao[1], Aryan Kalia[1], Zhonghao Zhang[1], Hudson Sun[1], Muhammed Patel[1], David A. Clausi[1], Lincoln Linlin Xu[2], Becky Segal[3], Joel Heath[3]

[1]Vision and Image Processing Lab, University of Waterloo
[2]Department of Geomatics Engineering, University of Calgary
[3]Arctic Eider Society
Corresponding authors: {j3hsiao, m32patel}@uwaterloo.ca

Abstract

Accurate monitoring of eider duck populations in Arctic Canada is essential for understanding ecosystem health and supporting conservation efforts in a rapidly changing climate. Traditional manual counting from aerial imagery is time-consuming, labor-intensive, and prone to observer bias. In this work, we present a human-in-the-loop wildlife counting system that integrates an open-vocabulary multi-species object detector to streamline and enhance the accuracy of eider duck surveys. The system leverages a pre-trained open-vocabulary model, enabling the identification of both target and incidental species without retraining, and employs human validation to correct and refine automated detections. This collaborative workflow combines the scalability of machine learning with expert ecological knowledge, reducing annotation effort while maintaining high accuracy. Field validation using aerial imagery from Arctic Canada demonstrates that our approach can significantly accelerate population assessments, improve consistency across surveys, and facilitate adaptive monitoring in remote environments.

Code — https://github.com/echonax07/OpenWildlife

Introduction

The common eider duck holds significant cultural and ecological importance for Inuit communities in Hudson Bay. Eider down is the warmest natural insulator known and has been fundamental to Inuit survival during Arctic winters (Heath 2011). Modern environmental disruptions, such as changes in sea ice and altered ocean current patterns driven by hydroelectric infrastructure in eastern North America, are increasingly affecting both the birds and the communities that depend on them (Ridenour et al. 2019). Thus, the demographic structure of the eider population functions as a sensitive indicator of ecosystem change.

The eider is sexually dimorphic, with males exhibiting predominantly white plumage and females displaying brown coloration. This distinction enables male and female individuals to be identified in aerial imagery (Fig. 1). While long-term aerial survey campaigns by the Arctic Eider Society have produced a substantial archive of high-resolution imagery, individual eiders occupy approximately 10×10 pixels and frequently appear in densely packed flocks (Fig. 1). Furthermore, white male ducks can frequently be confused with ice chunks due to their color. These characteristics make manual identification and differentiation of male and female individuals extremely difficult and time-consuming.

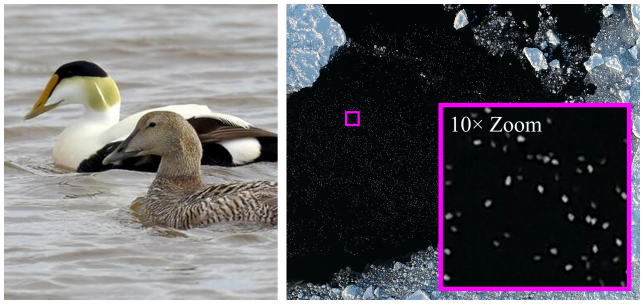

Figure 1: Visual characteristics of eider ducks in aerial surveys. Left: Reference images showing sexually dimorphic plumage between males (white) and females (brown) (Cornell Lab of Ornithology 2025). Right: Example aerial survey image (4368×2912) with $10\times$ zoom inset displaying the small size of eiders.

In this work, we address this gap by introducing a scalable approach for automated demographic analysis of eider ducks in aerial imagery. Specifically, our work's contributions are three-fold:

- **Approach**. We developed an AI-assisted annotation system combining the OpenWildlife foundation model with a human-in-the-loop workflow. Our solution features a regional correction mechanism, dynamic model refinement, and specialized visualization tools for high-density flocks. This innovative approach enables accurate sex classification of eiders while handling the challenges of small target sizes (10×10 pixels) and dense aggregations.
- **Evaluation**. We showed that the OpenWildlife model performed better than state-of-the-art object detectors at both zero-shot and finetuned detection of eider ducks due to domain-specific pretraining through aerial imagery combined with its open-vocabulary. We established a testing framework using 10 fully annotated aerial images (comprising 320 patches of size 512×512) contain-

ing approximately 35,000 individual eider annotations. Our human-in-the-loop system achieved a 77.6% recall rate with 22.2% counting error, while reducing annotation time by 87.5% compared to manual methods.

- **Deployment**. The system has been implemented as an extension to LabelStudio and deployed for use by the Arctic Eider Society. This deployment features practical innovations including image slicing for large surveys, confidence-based filtering, and real-time count estimation, all packaged in a Docker container for easy adoption by other wildlife monitoring organizations.

Related Work

Traditional methods for animal population monitoring using aerial imagery have relied heavily on manual annotation. While these approaches offer high accuracy, they are labor-intensive and not scalable for large datasets (Terletzky and Ramsey 2016). Early automated techniques, such as unsupervised classification and background subtraction, often suffered from high false positive rates, particularly when detecting dense or small-bodied species (Terletzky and Ramsey 2016).

The advent of deep learning, especially convolutional neural networks (CNNs), has significantly enhanced detection and counting accuracy. For instance, methods utilizing density maps have outperformed classical detectors in challenging scenarios, such as seabird colonies (Qian et al. 2023; Singh, Gangloff, and Pham 2023). Other strategies have employed transfer learning and domain adaptation to address the scarcity of labeled aerial data (Weinstein et al. 2022; Xu et al. 2024; Patel et al. 2023). However, despite their superior accuracy, density map-based methods remain labor-intensive to annotate, as they require marking all individual keypoints, making them difficult to scale. Furthermore, density map-based methods also struggle in providing precise locations for individual objects, especially in dense regions where objects overlap significantly, causing difficulty for human review and corrections (Gao et al. 2020).

Foundation models, pre-trained on large datasets, offer new possibilities for wildlife monitoring with limited data. MegaDetector (Beery, Morris, and Yang 2019) is one such model that aids in analyzing extensive camera trap data. Since foundation models are pre-trained on large amounts of data, they require minimal adaptation to new domains. Building upon this concept, we utilize OpenWildlife, a foundation model trained on extensive aerial imagery, as a backbone to develop an interactive Eider duck counting solution.

Methodology

Dataset

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 period of four months, 6 of these images were fully annotated by an Indigenous co-operative student using a default configuration of the open-source data labeling platform LabelStudio (Tkachenko et al. 2020-2025a), resulting in 8,339 keypoint annotations corresponding to individual ducks. These labels serve as the basis for fine-tuning our system.

OpenWildlife Model

Given the scarcity of labeled data, purely supervised training from scratch would severely overfit and fail to generalize to unseen viewpoints or colony configurations. Thus, we follow OpenWildlife (Patel et al. 2025) in building a large-scale pre-trained open-vocabulary detector based on MM-Grounding-DINO (Zhao et al. 2024), as described in Fig. 2.

Backbone Networks OpenWildlife follows the standard Grounding-DINO (Liu et al. 2024) architecture and uses two modality-specific backbones. For image inputs, we employ a Swin Transformer (Liu et al. 2021), which extracts multi-scale visual representations with shifted-window self-attention. In parallel, a BERT encoder (Devlin et al. 2019) is used to obtain contextualized embeddings of class descriptions. The two backbones operate independently and do not communicate until the feature fusion stage.

Feature Enhancer To jointly encode correspondences between the visual and textual modalities, we use a bi-directional attention-based Feature Enhancer. First, the encoded text and image features are passed through cross-attention blocks (text-to-image and image-to-text), which enable the model to fuse shared semantic information across modalities. The fused representations are further refined with self-attention layers—standard multi-head self-attention for the textual features, and deformable attention (Zhu et al. 2021) for the visual features.

Language-Guided Query Selection To account for the unusually large number of individuals per image in eider colonies, we increase the number of decoder queries from the 900 used in Grounding-DINO to 2,000. These queries are initialized via language-guided query selection, where the cosine similarity between each text feature and image feature is used to identify high-similarity regions as candidate object locations. The positional component of each selected region acts as a dynamic anchor box, while the content component is initialized as a zero vector. These queries then serve as informative starting points for the subsequent decoding process.

Cross-Modality Decoder The cross-modality decoder alternates between self-attention, image cross-attention, and text cross-attention layers. Image cross-attention enables refinement of object queries based on the visual feature maps, whereas text cross-attention injects class-specific semantic information into the queries. This facilitates alignment between bounding boxes and their corresponding textual descriptions. Each refined object query is then used to predict a bounding box and an associated class embedding.

Training Losses For bounding box regression, we employ a combination of L1 loss and Generalized IoU (GIoU) loss (Rezatofighi et al. 2019). For classification, we apply a focal-based contrastive loss (Lin et al. 2020) to align predicted object queries with the corresponding class tokens,

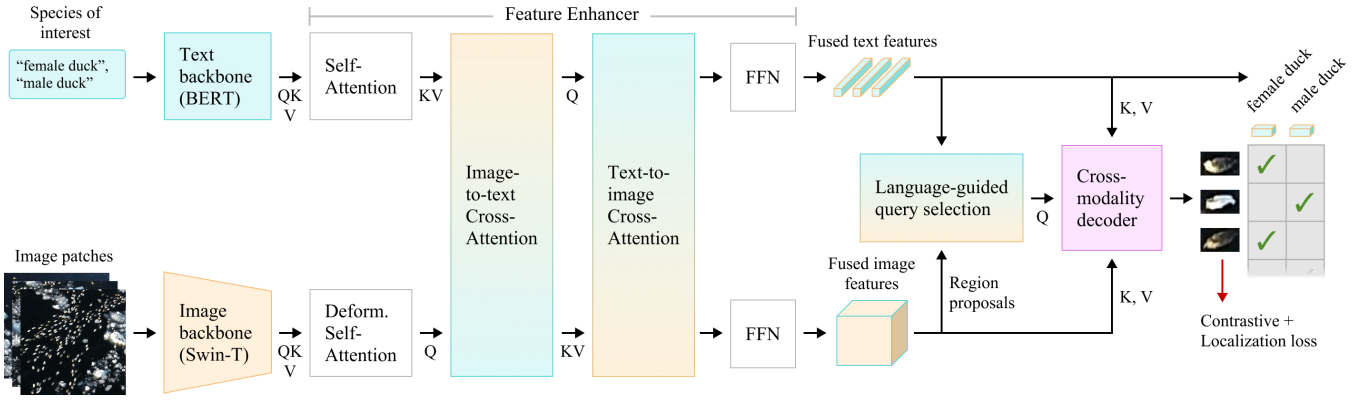

Figure 2: OpenWildlife model architecture adapted from MM-Grounding-DINO. The model processes image-text pairs through Swin Transformer and BERT backbones, with cross-modal fusion enabling open-vocabulary detection. Our approach builds upon pre-training across 15 aerial wildlife datasets before task-specific fine-tuning for eider duck detection.

which enables open-vocabulary detection. Bipartite matching loss (Carion et al. 2020) is additionally used during training to associate predicted queries with ground-truth annotations and avoid duplicate detections. Auxiliary losses are applied at each decoder layer following DETR (Carion et al. 2020).

The final loss for a training batch is defined as

$$\mathcal{L}_{total} = \lambda_{cls}\,\mathcal{L}_{cls} + \lambda_{L1}\,\mathcal{L}_{L1} + \lambda_{GIoU}\,\mathcal{L}_{GIoU}, \quad (1)$$

where $\mathcal{L}_{cls}$, $\mathcal{L}_{L1}$, and $\mathcal{L}_{GIoU}$ denote the classification, box L1, and GIoU components, respectively. Following DINO, we set the weights for these terms to $\lambda_{cls} = 1.0$, $\lambda_{L1} = 5.0$, and $\lambda_{GIoU} = 2.0$ in the final loss, and use (2.0, 5.0, 2.0) for the corresponding matching costs in the Hungarian assignment stage.

Pre-training Although Grounding-DINO benefits from large-scale object detection pre-training on general-purpose datasets (e.g., Objects365, V3Det), these datasets do not reflect the domain shift associated with aerial wildlife imagery. To address this, we follow OpenWildlife (Patel et al. 2025) and perform additional pre-training on 27,684 aerial wildlife images comprising 977,428 bounding boxes across 15 species. This domain-specific pre-training provides the model with coarse-grained representations of animal morphology and background context that are directly relevant to the downstream eider detection task. We refer to their paper for additional details about this pre-training dataset.

Experiments

Baseline Comparison Setup To validate the effectiveness of our approach, we conducted comparative evaluation against four state-of-the-art object detection architectures: DINO (Zhang et al. 2022), CO-DETR (Zong, Song, and Liu 2023), YOLOv8 (Sohan, Sai Ram, and Rami Reddy 2024), and MM-Grounding-DINO (Zhao et al. 2024). All these models were initially pretrained on the data described above. We then evaluate the models in both zero-shot and fine-tuned configurations on our eider duck dataset to assess generalization capabilities and adaptation potential.

All the models leveraged aerial image specific pre-training on 15 aerial wildlife datasets comprising 27,684 images with 977,428 annotations spanning diverse species including beluga whales, elephants, penguins, and various bird species (Patel et al. 2025). Training was conducted for 20 epochs using AdamW optimizer with learning rate 4×10^{-5} and 1000 warm-up steps. Images were preprocessed by slicing into 1024×1024 pixel patches to maintain computational efficiency while preserving fine-grained details essential for detecting small objects (10×10 pixels) typical in aerial surveys. Data augmentation included affine transformations, brightness/contrast adjustments, RGB/HSV shifts, JPEG compression, and blur effects to enhance model robustness across varying environmental conditions.

Evaluation Protocol The zero-shot evaluation assessed each model's ability to detect eider ducks without any task-specific training, providing insight into cross-domain generalization capabilities. Fine-tuned evaluation measured model performance after training on our limited set of six fully annotated, full-resolution eider duck images. All evaluations used mean Average Precision at IoU threshold 0.5 (mAP50) as the primary metric, calculated on our 10-image test set containing approximately 35,000 individual eider duck annotations.

Results

Figure 3 presents the performance comparison of OpenWildlife against four state-of-the-art detection architectures on the AES eider duck dataset. We evaluated both zero-shot and fine-tuned configurations to assess the models' generalization capabilities and adaptation potential for this novel species detection task.

Zero-shot and Fine-tuned Evaluation

In the zero-shot evaluation, OpenWildlife achieved 0.1 mAP50, performing better than all baseline methods. Traditional closed-set detectors CO-DETR and YOLOv8 completely failed to detect eider ducks (0.00 mAP50), demonstrating their inability to generalize to novel species not

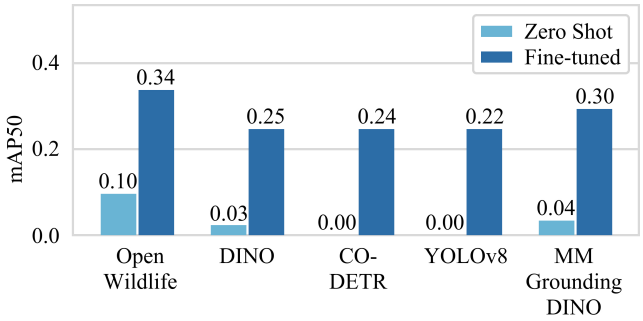

Figure 3: Comparative performance of object detectors on eider duck detection. OpenWildlife outperforms closed-vocabulary detectors (DINO, CO-DETR, YOLOv8) in zero-shot settings due to its open-vocabulary architecture and aerial domain pre-training. After fine-tuning, all models show improved performance, with OpenWildlife maintaining the highest mAP50 (0.34) on our 10-image test set containing 35,000 annotations.

present in their training data. DINO achieved minimal performance at 0.03 mAP50, while MM-Grounding-DINO reached 0.04 mAP50. Overall, these numbers are very poor, highlighting the distinct nature of eider duck imagery.

Since closed-set models cannot predict novel classes, we evaluated them using a relaxed localization-based metric, where predictions are counted as true positives if they are correctly localized, regardless of their class label. Even under this lenient setting, OpenWildlife outperforms the baselines, highlighting the importance of combining open-vocabulary architecture with aerial-domain pretraining for zero-shot detection of small, densely packed animals.

After fine-tuning, the performance gap between models narrowed. OpenWildlife maintained the highest performance at 0.34 mAP50, but closed-set detectors achieved similar results, as shown in Figure 3. This convergence occurs because fine-tuning provides visual supervision for the target species, reducing the relative advantage of text encoding capabilities. When models receive sufficient visual examples of eider ducks, visual features become sufficient to distinguish target classes, diminishing the benefit of open-vocabulary architecture in this specific closed-set scenario.

Architectural Advantage Analysis The results indicate that OpenWildlife's performance advantage arises from the combination of its open-vocabulary architecture and aerial-domain pretraining, which enables effective zero-shot generalization to novel species through transferable visual priors. The improvement from zero-shot to fine-tuned performance ($0.10 \rightarrow 0.34$ mAP50) demonstrates that limited human-in-the-loop annotations are sufficient to substantially improve detection accuracy. Across both evaluation settings, OpenWildlife consistently outperforms alternatives and can be deployed immediately due to its strong pretrained initialization, making it well suited for wildlife monitoring in resource-constrained settings where large-scale manual annotation is impractical.

Deployment

While tools such as LabelStudio (Tkachenko et al. 2020-2025b), Roboflow (Dwyer et al. 2025), and CVAT (CVAT.ai Corporation 2023) support pre-labeling using pre-trained models, our objective was to continuously refine the OpenWildlife model during annotation. To accomplish this, we extended LabelStudio so that annotators could predict, correct, and incrementally retrain the model. This design choice allows model quality to improve in parallel with annotation progress, rather than only after a complete labeling pass.

We selected LabelStudio over CVAT, another free and open-source alternative, due to its flexible XML-based configuration, the ability to self-host both the frontend (annotation interface) and machine learning backend (Tkachenko et al. 2020-2025b), as well as its proven success within AES for initial eider duck labeling.

Human-in-the-loop Workflow

Figure 4 shows the human-in-the-loop (HITL) pipeline we implemented within LabelStudio. Although the Community Edition of LabelStudio supports prediction and training through a machine learning backend (Tkachenko et al. 2020-2025b), it assumes that images are fully annotated prior to retraining. This assumption is infeasible in our setting, where individual aerial images frequently contain thousands of eider ducks. Fully correcting model predictions before retraining would be prohibitively time-consuming. To address this, we introduced a *regional* annotation workflow. When a task is opened for annotation, the current model generates an initial set of predictions. The annotator edits only a selected subregion and draws a training polygon to indicate verified ground truth. After pressing the *Train* button, the backend finetunes the model on this verified region, redeploys the updated weights, and enables the annotator to issue a new *Predict* on the remaining areas. This predict–annotate–train cycle can be repeated iteratively until all regions are accurately labeled, without requiring exhaustive full-image annotation before each update.

Quality-of-Life Features To make this incremental workflow practical for large, high-density images, we incorporated several quality-of-life enhancements directly into the annotation interface. A "bubble annotator" based on the Hierarchical-DBSCAN algorithm (Campello, Moulavi, and Sander 2013) groups spatially close detections, improving the responsiveness of zooming and panning. An image slicing tool can be used to split images into crops if the number of eiders within an image exceeds the number of detection queries accepted by the model. A notification system informs users when a training job completes, eliminating the need to check backend logs. To further improve annotator efficiency, we added a confidence threshold slider that allows users to filter low-confidence predictions and see real-time duck counts at the chosen threshold. A model settings panel enables users to revert to or name checkpoints, upload new model weights, and tweak training parameters. Finally, the system supports XLSX export, providing high-level project metrics in a familiar format.

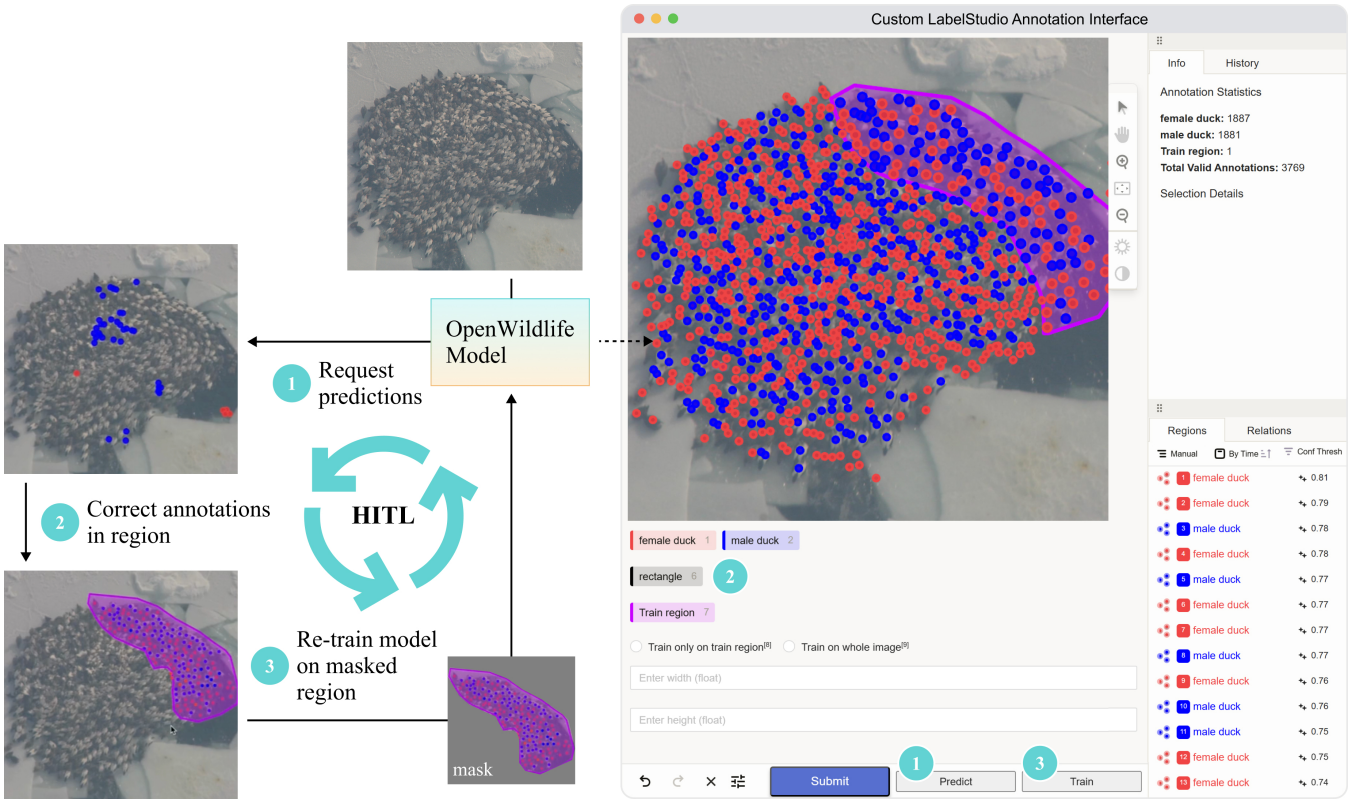

Figure 4: The human-in-the-loop pipeline built within LabelStudio. Given predictions from our OpenWildlife model, a human can correct annotations within a region and re-train the model on this masked area. This cycle can be repeated until the annotations are satisfactory. The right panel shows the interface in which this pipeline is executed, including bubbled numbers which reflect the pipeline's steps, annotation tools, and other quality-of-life features.

Technical Design The system consists of two operational workflows (prediction and training) and four main components: the LabelStudio frontend, the LabelStudio backend, the ML backend, and a Redis server. The frontend communicates with the backend via REST APIs, while the ML backend handles both prediction and training and maintains model checkpoints and state.

During prediction, the frontend calls the *force_predict* API. The ML backend loads the relevant model parameters, slices images, and returns the predictions after converting from MSCOCO (Lin et al. 2014) format to the LabelStudio JSON format. Ground-truth regions and their corresponding labels are preserved, while previous drafts and predictions of other image sections are replaced with the newly generated results. This typically completes within 30 seconds.

Training is handled similarly, except that it is executed asynchronously. The request queues a job in Redis, and the frontend periodically checks its status using the *job_status* API. Jobs usually complete in 2–4 minutes. Forking the training process improves GPU multitasking and simplifies memory clean-up by avoiding long-running blocking calls. All components are packaged within a Docker container to simplify deployment and avoid environment-specific configuration issues.

Experimental Setup

Dataset For quantitative evaluation, we hold out 10 aerial images as a test set. These images were selected to be representative of the full dataset, covering a range of geographic conditions (e.g. ducks swimming on sea ice vs. flying over open water), field-of-view sizes, behavioral contexts (e.g. swimming, launching, and in-flight), and population sizes (from 352 to 11,099 individuals per image). Although only ten images are used, they correspond to 320 patches of size 512×512, and required approximately 35,000 individual annotations to construct the ground-truth labels. Generating these labels required several days of dedicated effort and highlights the difficulty of preparing densely populated aerial imagery even at small scale.

Model Variants We evaluate three different model variants in increasing order of supervision:

- **Pre-trained**: model trained only on 15 aerial wildlife datasets following OpenWildlife pre-training.
- **Fine-tuned**: pre-trained model further fine-tuned using the 6 fully annotated eider colony images.
- **HITL**: fine-tuned model further refined by incorporating human feedback through the annotation interface.

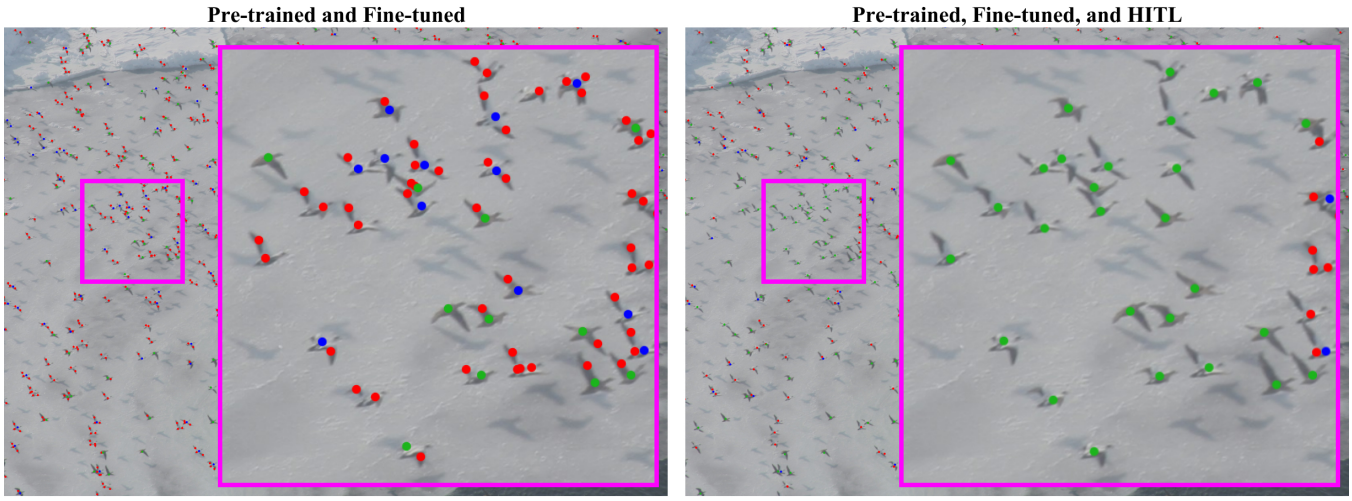

Figure 5: Qualitative improvement through human-in-the-loop refinement. Comparison shows detection results before (left) and after (right) HITL feedback, with true positives (green), false positives (red), and false negatives (blue). The HITL-refined model demonstrates improved precision in distinguishing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

User Study To evaluate the annotation workflow, we conducted a small-scale user study with five participants. Each participant first completed a short training session covering the use of the web-based LabelStudio interface and the task of identifying male and female eider ducks in aerial imagery. Participants then annotated a set of images over a four-month period. For each image, model-generated predictions were provided as a starting point, and participants reviewed and corrected these predictions as needed.

Evaluation Metrics Performance on the 10-image test set is evaluated using precision, recall, mean absolute error (MAE), and percentage error. For the user study, annotation efficiency is measured as the average time (in minutes) required to annotate one full image. All the models are evaluated at a score threshold of 0.3.

Results and Discussion

Accuracy and Efficiency Table 1 reports the performance of the three model variants. Accuracy improves consistently across training stages: fine-tuning yields large gains over pre-training alone, and the HITL stage produces further improvements, particularly in recall and MAE. In terms of annotation efficiency, the HITL workflow reduces the time required to annotate a single image by 87.5% compared to manual annotation (Table 2).

Transfer Learning To verify that HITL refinement does not overfit to the imaging conditions of a single site, we perform a leave-out experiment ("HITL Other"), in which HITL training is carried out on five non-overlapping images and the resulting model is evaluated on the same 10-image test set. As shown in Table 3, the HITL Other variant still improves over the fine-tuned baseline, demonstrating generalization across locations and conditions.

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 1: Performance comparison of three model variants on the 10-image held-out test set, evaluated at a detection confidence score threshold of 0.3. Checkmarks indicate which supervision stages were used during training.

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

Table 2: Annotation time comparison between fully manual labeling and the human-in-the-loop (HITL) workflow.

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

Table 3: Transfer learning results after HITL training on 5 non-overlapping images.

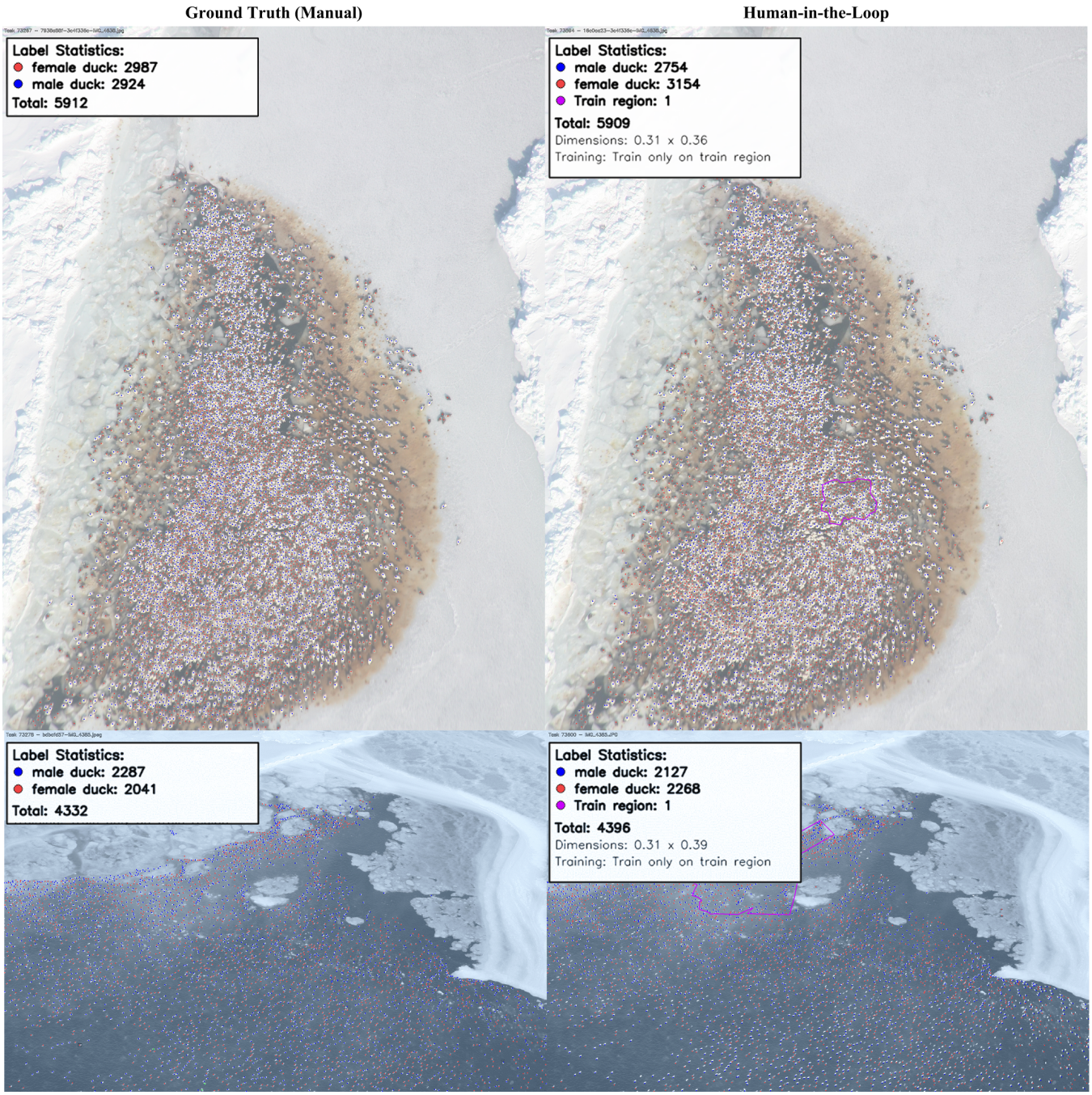

Figure 6: Comparison between a fully manual annotation workflow (left) and the human-in-the-loop process (right). Highlighted regions mark areas where annotators intervened during the HITL cycle, while the inset summarizes the final male and female counts for each approach. The side-by-side view illustrates how limited, localized corrections are sufficient to bring the HITL output into close alignment with fully manual labels.

Qualitative Analysis Figure 5 illustrates qualitative improvements obtained from HITL refinement. The model identifies a higher proportion of true positives and produces fewer false positives compared to the fine-tuned baseline. Participants also reported a substantially improved user experience, describing the model predictions as "accurate enough to require only minor adjustments" and noting the "orders of magnitude" time savings relative to fully manual annotation. Full-image comparisons of manually-annotated ground truth and human-in-the-loop results is available in Figure 6.

Discussion

Our work demonstrates the effectiveness of leveraging foundation models and human-in-the-loop (HITL) workflows for automating the demographic analysis of eider ducks in aerial imagery. Several key insights emerged from this study:

1. **Foundation Models in Low-Data Regimes**: The OpenWildlife model, pre-trained on diverse aerial wildlife datasets, proved invaluable for few-shot adaptation to eider duck detection. The vocabulary-guided pre-training enabled accurate predictions despite limited labeled data, significantly reducing annotation effort compared to training from scratch.
2. **Handling High-Density Imagery**: The sheer size of aerial images and the high density of eider ducks necessitated specialized functionality, such as image splitting, adjustable batch processing, and annotation clustering. These optimizations were critical for maintaining system responsiveness and usability during large-scale annotation tasks.
3. **Granular HITL Workflows**: The regional correction approach, where annotators refine predictions incrementally rather than correcting entire images before retraining, proved particularly beneficial for densely populated scenes. This iterative predict-annotate-train loop allowed for continuous model improvement without requiring exhaustive full-image annotations at each step.

Participants in our user study reported substantial efficiency gains, with annotation times reduced by nearly 90% compared to manual labeling. Qualitative analysis further confirmed that the HITL-refined model achieved higher precision and recall, particularly in challenging scenarios with overlapping or occluded ducks.

Future Work

While our system represents a significant step forward in automating wildlife population monitoring, several avenues for improvement remain:

- **Optimizing Canvas Rendering**: Large numbers of annotations can strain front-end rendering performance. Future work could explore more efficient visualization techniques, such as dynamic level-of-detail rendering or GPU-accelerated annotation clustering, to further enhance the user experience.
- **Open-Source Release**: To support broader adoption by conservation agencies, we plan to release both the annotation tool and the fine-tuned OpenWildlife model as open-source software. This will enable other organizations working with aerial wildlife imagery to benefit from our approach.
- **Expanding to Other Species**: The framework could be extended to support additional Arctic bird species or adapted for terrestrial wildlife monitoring, provided sufficient pre-training data is available.
- **Active Learning Integration**: Incorporating active learning strategies to prioritize uncertain or high-value regions for human review could further optimize the annotation process.

Conclusion

We presented a scalable, human-in-the-loop system for automated demographic analysis of eider ducks in aerial imagery. By combining OpenWildlife's foundation model with an iterative regional refinement workflow, we achieved substantial improvements in both accuracy and annotation efficiency compared to manual methods. Our approach not only addresses the immediate challenges of eider duck monitoring but also provides a flexible framework adaptable to other wildlife conservation efforts. As environmental changes continue to impact Arctic ecosystems, tools like this will play an increasingly vital role in supporting both ecological research and Indigenous-led conservation initiatives.

Ethical Statement

This research was conducted in close partnership with the Arctic Eider Society and Inuit communities, following principles of Indigenous data sovereignty and community-led research. The human-in-the-loop annotation system was designed to augment rather than replace Indigenous knowledge and monitoring practices. While our system improves monitoring efficiency, we recognize the importance of maintaining human oversight to ensure cultural context and traditional ecological knowledge are preserved in conservation decisions.

Acknowledgments

We thank the Inuit communities of Sanikiluaq and the Belcher Islands for their partnership, knowledge sharing, and guidance throughout this research. We also acknowledge the provision of computational resources through the Digital Research Alliance of Canada and the support of the University of Waterloo.

References

Beery, S.; Morris, D.; and Yang, S. 2019. Efficient Pipeline for Camera Trap Image Review. Associated arXiv preprint: arXiv:1907.06772.

Campello, R. J. G. B.; Moulavi, D.; and Sander, J. 2013. Density-Based Clustering Based on Hierarchical Density Estimates. In Pei, J.; Tseng, V. S.; Cao, L.; Motoda, H.; and Xu, G., eds., *Advances in Knowledge Discovery and Data Mining*, 160–172. Berlin, Heidelberg: Springer Berlin Heidelberg. ISBN 978-3-642-37456-2.

Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; and Zagoruyko, S. 2020. End-to-End Object Detection with Transformers. In *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I*, 213–229. Berlin, Heidelberg: Springer-Verlag. ISBN 978-3-030-58451-1.

Cornell Lab of Ornithology. 2025. Common Eider Identification.

CVAT.ai Corporation. 2023. Computer Vision Annotation Tool (CVAT).

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. 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), 4171–4186.

Dwyer, B.; Nelson, J.; Hansen, T.; et al. 2025. Roboflow (Version 1.0). Accessed: 2025-08-19.

Gao, G.; Gao, J.; Liu, Q.; Wang, Q.; and Wang, Y. 2020. Cnn-based density estimation and crowd counting: A survey. *arXiv preprint arXiv:2003.12783*.

Heath, J. 2011. People of a Feather.

Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; and Dollár, P. 2020. Focal Loss for Dense Object Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(2): 318–327.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, 740–755. Springer.

Liu, S.; Zeng, Z.; Ren, T.; Li, F.; Zhang, H.; Yang, J.; Jiang, Q.; Li, C.; Yang, J.; Su, H.; Zhu, J.; and Zhang, L. 2024. Grounding DINO: Marrying DINO with Grounded Pre-Training for Open-Set Object Detection. ArXiv:2303.05499 [cs].

Liu, Z.; Lin, Y.; Cao, Y.; Hu, H.; Wei, Y.; Zhang, Z.; Lin, S.; and Guo, B. 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, 10012–10022.

Patel, M.; Chen, X.; Xu, L.; Cantu, F. J. P.; Turnes, J. N.; Brubacher, N. C.; Clausi, D. A.; and Scott, K. A. 2023. The influence of input image scale on deep learning-based beluga whale detection from aerial remote sensing imagery. In *IGARSS 2023-2023 IEEE International Geoscience and Remote Sensing Symposium*, 5732–5734. IEEE.

Patel, M.; Turnes, J. N.; Hsiao, J.; Xu, L.; and Clausi, D. 2025. OpenWildlife: Open-Vocabulary Multi-Species Wildlife Detector for Geographically-Diverse Aerial Imagery. ArXiv:2506.19204 [cs].

Qian, Y.; Humphries, G. R.; Trathan, P. N.; Lowther, A.; and Donovan, C. R. 2023. Counting animals in aerial images with a density map estimation model. *Ecology and Evolution*, 13(4): e9903.

Rezatofighi, H.; Tsoi, N.; Gwak, J.; Sadeghian, A.; Reid, I.; and Savarese, S. 2019. Generalized Intersection Over Union: A Metric and a Loss for Bounding Box Regression. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 658–666.

Ridenour, N. A.; Hu, X.; Jafarikhasragh, S.; Landy, J. C.; Lukovich, J. V.; Stadnyk, T. A.; Sydor, K.; Myers, P. G.; and Barber, D. G. 2019. Sensitivity of freshwater dynamics to ocean model resolution and river discharge forcing in the Hudson Bay Complex. *Journal of Marine Systems*, 196: 48–64.

Singh, T.; Gangloff, H.; and Pham, M.-T. 2023. Object counting from aerial remote sensing images: application to wildlife and marine mammals. In *IGARSS 2023-2023 IEEE International Geoscience and Remote Sensing Symposium*, 6580–6583. IEEE.

Sohan, M.; Sai Ram, T.; and Rami Reddy, C. V. 2024. A review on yolov8 and its advancements. In *International Conference on Data Intelligence and Cognitive Informatics*, 529–545. Springer.

Terletzky, P. A.; and Ramsey, R. D. 2016. Comparison of three techniques to identify and count individual animals in aerial imagery. *Journal of Signal and Information Processing*, 7(3): 123–135.

Tkachenko, M.; Malyuk, M.; Holmanyuk, A.; and Liubimov, N. 2020-2025a. Label Studio: Data labeling software. Open source software available from https://github.com/HumanSignal/label-studio.

Tkachenko, M.; Malyuk, M.; Holmanyuk, A.; and Liubimov, N. 2020-2025b. Label Studio: Data labeling software. Open source software available from https://github.com/HumanSignal/label-studio.

Weinstein, B. G.; Garner, L.; Saccomanno, V. R.; Steinkraus, A.; Ortega, A.; Brush, K.; Yenni, G.; McKellar, A. E.; Converse, R.; Lippitt, C. D.; et al. 2022. A general deep learning model for bird detection in high-resolution airborne imagery. *Ecological Applications*, 32(8): e2694.

Xu, Z.; Wang, T.; Skidmore, A. K.; and Lamprey, R. 2024. A review of deep learning techniques for detecting animals in aerial and satellite images. *International Journal of Applied Earth Observation and Geoinformation*, 128: 103732.

Zhang, H.; Li, F.; Liu, S.; Zhang, L.; Su, H.; Zhu, J.; Ni, L. M.; and Shum, H.-Y. 2022. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. *arXiv preprint arXiv:2203.03605*.

Zhao, X.; Chen, Y.; Xu, S.; Li, X.; Wang, X.; Li, Y.; and Huang, H. 2024. An open and comprehensive pipeline for unified object grounding and detection. *arXiv preprint arXiv:2401.02361*.

Zhu, X.; Su, W.; Lu, L.; Li, B.; Wang, X.; and Dai, J. 2021. Deformable DETR: Deformable Transformers for End-to-End Object Detection. ArXiv:2010.04159 [cs].

Zong, Z.; Song, G.; and Liu, Y. 2023. Detrs with collaborative hybrid assignments training. In *Proceedings of the IEEE/CVF international conference on computer vision*, 6748–6758.