Semi-Supervised Object Segmentation via Active Learning for Efficient Ecological Monitoring

Dane Brown[1 [0000−0001−7395−7370]] and Karen Bradshaw[1 [0000−0003−3979−5675]]

Rhodes University, Grahamstown, Makhanda, 6140, South Africa
d.brown@ru.ac.za
https://www.ru.ac.za/computerscience/people/academicstaff/profdanebrown/

Abstract. Deep learning models for instance segmentation have achieved remarkable success, yet their deployment in specialised domains like ecological monitoring is constrained by the prohibitive cost of acquiring high-quality polygonal annotations. This annotation dependency creates a fundamental bottleneck, limiting both scalability and adaptability of these models in real-world conservation scenarios. This paper introduces a data-centric workflow that addresses this challenge through an iterative, multi-stage active learning strategy enhanced with foundation models. The methodology integrates CLIP diversity sampling as an acquisition function with a semi-automated annotation pipeline that combines YOLOv8x6 detection proposals, human-in-the-loop verification, and SAM2-prompted segmentation refinement. A progressive training strategy using YOLOv11s-seg with quality-controlled pseudo-labelling iteratively expands the training dataset while maintaining annotation quality standards. Validated on African Penguin monitoring using Open Images V7 data and independent SANParks field data, experimental results demonstrate that CLIP diversity sampling achieves mAP_{50} of 0.82 with 400 training samples (without SAM2 refinement), compared to 0.69 for random sampling; with SAM2-refined annotations, performance reaches 0.88 mAP_{50} using the same 400 samples. Cross-domain generalisation on independent SANParks field data achieves 0.81–0.84 mAP_{50}. The framework reduces annotation requirements while providing a practical solution for deploying instance segmentation in data-scarce domains.

Keywords: Active Learning · Instance Segmentation · Contrastive Language–Image Pre-training (CLIP) · Ecological Monitoring · Segment Anything 2 (SAM2) · You Only Look Once (YOLO)

1 Introduction

The development of deep neural networks for instance segmentation has provided unprecedented capabilities for fine-grained image analysis, transforming fields from medical imaging to autonomous systems [5, 13, 19]. However, the supervised learning paradigm used by these models has a critical dependency that significantly limits their practical deployment: the need for vast quantities of meticulously annotated training data [8, 20]. This requirement represents the primary obstacle preventing widespread adoption of advanced computer vision solutions in specialised domains, where the cost and complexity of data preparation often exceed those of model development by orders of magnitude [1, 25].

In ecological monitoring applications, this challenge becomes particularly acute [16, 24]. The task of manually delineating pixel-perfect segmentation polygons demands not only significant time investment but also specialised domain knowledge to accurately identify species boundaries, handle occlusions, and account for environmental variations [29]. Conservation biologists and trained field technicians represent a scarce and expensive resource, making large-scale annotation campaigns financially and logistically prohibitive. This annotation bottleneck constitutes the central obstacle preventing widespread adoption of AI-driven monitoring solutions in critical conservation contexts.

The African Penguin (*Spheniscus demersus*) serves as a compelling case study for this challenge. Listed as critically endangered with fewer than 20,000 breeding pairs remaining compared with historical estimates of over a million breeding pairs [21], this species requires intensive monitoring across multiple colonies spanning diverse geographical and environmental conditions. While automated monitoring using computer vision represents a clear conservation objective [16, 33], the volume and diversity of imagery from monitoring sites render traditional annotate-everything approaches intractable. Furthermore, models trained on data from one

colony often fail to generalise to others due to domain shift – variations in lighting conditions, backgrounds, camera perspectives, and environmental factors that cause significant performance degradation [11, 26].

This research addresses the annotation dependency by shifting from model-centric to data-centric AI before later shifting back to tuning the model on the "right data" [35]. A workflow is designed that maximises data utility whilst minimising human labelling effort through semi-automation. This involves leveraging recent advances in foundation models – specifically Contrastive Language–Image Pre-training (CLIP) for semantic understanding [27] and Segment Anything 2 (SAM2) for zero-shot segmentation [28] – to create a semi-automated annotation pipeline that preserves quality while dramatically reducing manual effort. Within the broader agenda of computational ecology and Imageomics, which seeks to make biological traits computable from field imagery and enable in situ pipelines, there is a clear need for robust, deployable data-centric workflows that integrate intelligent selection, semi-automated labelling, and efficient model training [9]. This paper targets three gaps exposed in prior ecological vision work: first, the prohibitive cost of pixel-accurate instance segmentation annotations; second, the fragility of early-stage model-dependent acquisition functions for active learning; and third, the limited cross-domain generalisation under realistic field shifts. The resulting technical contributions were yielded by directly addressing these gaps.

Technical contributions:

1. A systematic integration of state-of-the-art foundation models into a cohesive active learning framework specifically designed for instance segmentation in data-scarce domains.
2. A practical, reproducible methodology that can be readily adapted to other conservation monitoring applications where traditional supervised learning approaches prove prohibitively expensive.
3. A novel combination of CLIP diversity sampling, detector-prompted annotation, and progressive pseudo-labelling that reduces manual annotation dependency.
4. A comprehensive evaluation demonstrating significant annotation cost reduction while maintaining model performance across diverse environmental conditions.

2 Literature Review

This section examines three interconnected research domains: the application of instance segmentation in ecological monitoring contexts, active learning strategies for reducing annotation requirements, and the transformative impact of foundation models on data-centric AI workflows.

2.1 Instance Segmentation in Conservation Applications

Computer vision applications in wildlife monitoring have evolved considerably over the past decade, progressing from classical detection methods to sophisticated deep learning approaches [24].

Contemporary research has gravitated toward recent Ultralytics YOLO variants such as v8 and v11, [14] due to their exceptional balance of speed, accuracy, and implementation versatility. These architectures have been successfully applied across diverse taxonomic groups, from marine environments to terrestrial wildlife surveys [6, 30]. However, the majority of this research assumes the existence of substantial, fully-annotated datasets and focuses primarily on bounding box detection rather than the more demanding task of instance segmentation.

Instance segmentation, while providing superior morphological detail and sophisticated handling of occlusion scenarios, remains significantly underexplored in conservation contexts due to its substantially higher annotation requirements. This represents a critical gap in the literature, as conservation applications often require precise object boundaries for behavioural analysis, population assessment, and automated monitoring systems where simple bounding boxes prove insufficient [22]. Recent object detection and tracking studies illustrate wildlife detection in camera-trap imagery [9, 36]. Animal-focused instance segmentation in diverse settings is still understudied. To the best of our knowledge, no studies explore instance segmentation on penguin images collected from camera traps and other non-satellite media.

2.2 Active Learning Paradigms for Annotation Efficiency

Active learning provides a principled framework for addressing annotation scarcity by iteratively selecting the most informative samples for human labelling from a pool of unlabelled data. This premise enables models

to guide their own training data acquisition, theoretically achieving superior performance with significantly fewer labelled examples compared to random sampling strategies.

The selection mechanism, termed the acquisition function, represents the core algorithmic component distinguishing different active learning approaches. Traditional uncertainty-based methods, including entropy sampling [17] and margin-based selection [31], have dominated the field but suffer from several practical limitations. These methods require access to well-calibrated model confidence scores, which are often unreliable in early training phases or when using pre-trained models on out-of-domain data – precisely the scenario encountered in ecological monitoring applications.

Recent research has demonstrated the superior performance of diversity-based approaches in ecological contexts. Norouzzadeh et al. [23] showed that geometric diversity sampling using k-center clustering [32] outperformed uncertainty-based methods for wildlife camera trap classification. This finding proves particularly relevant to practical constraints: diversity sampling requires no initial model assumptions and focuses on building representative datasets from project inception – a more robust strategy for bootstrapping performance in data-scarce environments.

Table 1: Comparison of active learning acquisition functions for ecological segmentation tasks. Feasibility assessment considers practical deployment constraints including model requirements, computational overhead, and compatibility with foundation model pipelines.

Acquisition Function	**Methodology and Characteristics**	**Deployment Feasibility**
Random Sampling	Uniform random selection serving as baseline for comparative evaluation	**High**: No dependencies
Uncertainty Sampling [17]	Selects samples with highest model uncertainty using entropy or variation ratios	**Low**: Requires calibrated confidence estimates
CLIP Diversity [27]	Leverages CLIP embeddings to identify visually unique samples ensuring broad feature coverage	**High**: No model dependencies, robust across domains
Query-by-Committee [10]	Employs ensemble disagreement as selection criterion	**Low**: Computationally expensive
k-Center Sampling [32]	Selects geometrically diverse points covering embedding space efficiently	**Medium**: More complex than uniqueness sorting

2.3 Foundation Models in Data-Centric Workflows

The emergence of large-scale foundation models has fundamentally transformed data-centric AI methodologies, providing powerful new tools for intelligent data curation and annotation acceleration. CLIP's multimodal architecture enables the extraction of rich, semantically meaningful image representations that capture both low-level visual features and high-level conceptual relationships [27]. This capability proves invaluable for implementing sophisticated diversity sampling strategies that extend beyond simple pixel-level similarity measures to capture semantic content variations.

Concurrently, the Segment Anything Model (SAM) and its successor SAM2 have revolutionised annotation workflows through unprecedented zero-shot segmentation capabilities [15, 28]. SAM2's ability to transform simple spatial prompts – such as bounding boxes or point coordinates – into detailed instance masks represents a paradigm shift in annotation efficiency. The model's promptable interface enables seamless integration with existing detection pipelines, creating hybrid workflows that leverage both discriminative detection models and generative segmentation capabilities.

The synergistic combination of these foundation models addresses complementary aspects of the annotation challenge: CLIP enables intelligent data selection by identifying diverse, representative samples, while SAM2 accelerates the annotation process itself by automating the most labour-intensive component – precise boundary delineation. This integration represents a key innovation in the proposed approach, moving beyond individual model capabilities to create comprehensive data-centric workflows that are both theoretically grounded and practically deployable.

3 Proposed System

The framework includes a data-centric methodology designed to efficiently construct high-quality instance segmentation datasets from large pools of unlabelled images. The FiftyOne data management platform serves as the orchestrating infrastructure, providing robust data versioning, visualisation, and workflow management capabilities. The methodology centres on an iterative four-stage active learning cycle that progressively refines both dataset quality and model performance through systematic data selection, semi-automated annotation, incremental training, and quality-controlled pseudo-labelling.

3.1 Methodology Overview

Before detailing the technical implementation, a high-level summary of the four main components within the active learning loop is presented:

1. **Intelligent Data Selection:** CLIP diversity sampling identifies the most informative, representative images from the unlabelled pool by leveraging semantic embeddings to ensure broad coverage of environmental and contextual variations while minimising redundancy.
2. **Semi-Automated Annotation:** A hybrid pipeline combines YOLOv8x6 detection proposals with human-in-the-loop verification, followed by SAM2-prompted segmentation to generate high-quality instance masks efficiently.
3. **Progressive Model Training:** YOLOv11s-seg is trained incrementally using both verified human annotations and high-confidence pseudo-labels, with quality control mechanisms preventing error accumulation.
4. **Systematic Optimisation:** Upon convergence, Bayesian hyperparameter optimisation fine-tunes the model for deployment-ready performance.

This iterative process continues until convergence criteria are met (performance improvement $< 1\%$ mAP_{50} between iterations or pseudo-labelling confidence stabilisation), ensuring both annotation efficiency and model quality.

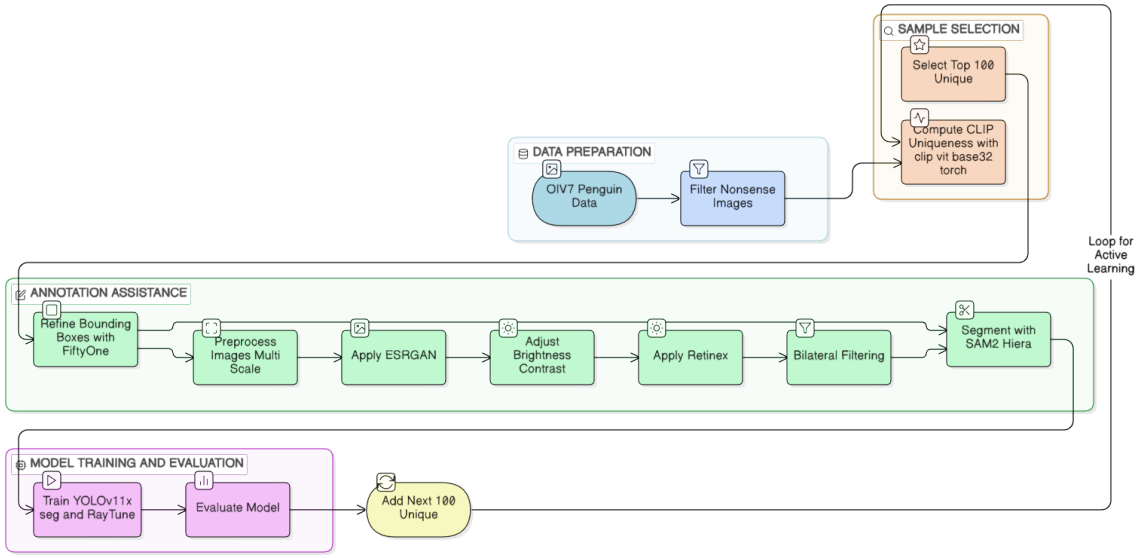

Fig. 1: Comprehensive overview of the foundation model-driven active learning workflow. The methodology progresses through four interconnected stages: 1) CLIP diversity sampling for intelligent data selection, 2) hybrid annotation combining YOLOv8x6 proposals with SAM2 refinement, 3) progressive model training with quality-controlled pseudo-labelling, and 4) systematic hyperparameter optimisation. The iterative nature enables continuous dataset and model improvement while minimising annotation effort.

3.2 Stage 1: CLIP-Driven Diversity Data Selection

The active learning cycle initiates with intelligent subset selection designed to maximise dataset diversity whilst minimising redundancy. This stage implements a principled approach to identifying the most informative samples from the unlabelled data pool, leveraging CLIP's semantic understanding capabilities.

The diversity sampling process operates through three sequential steps with computational complexity $O(n^2)$ for n images in the unlabelled pool. First, CLIP ViT-B/32 (`clip-vit-base32`) computes high-dimensional image embeddings for every image in the unlabelled pool via FiftyOne Brain. Embeddings are unit-normalised, and distances are computed with L_2, capturing both low-level visual features and high-level semantic content for robust similarity comparisons beyond pixel-level analysis [7]. Second, the FiftyOne Brain component calculates uniqueness scores by measuring each sample's mean distance to its k-nearest neighbours in the embedding space (typically $k = 5$). This metric effectively identifies visual outliers that represent underexplored regions of the feature space, ensuring broad coverage of potential data distributions. Finally, the top-N most unique samples (typically $N = 100$ for practical annotation batches) are selected to form the valuable set for the current annotation cycle.

The convergence criterion for the active learning loop is established when the performance improvement between consecutive iterations falls below 1% mAP_{50} or when the pseudo-labelling confidence distribution stabilises with less than 5% variation in high-confidence predictions across three consecutive iterations.

Algorithm 1 CLIP Diversity Sampling

Require: Unlabelled image pool $\mathcal{U}$, batch size N
Ensure: Selected diverse subset $\mathcal{S} \subset \mathcal{U}$
Initialise embedding collection $\mathbf{E} \leftarrow \{\}$
for each image $x_i \in \mathcal{U}$ **do**
  Compute CLIP embedding $\mathbf{e}_i \leftarrow \text{CLIP}(x_i)$
  $\mathbf{E} \leftarrow \mathbf{E} \cup \{\mathbf{e}_i\}$
end for
Initialise uniqueness scores $\mathbf{U} \leftarrow \{\}$
for each embedding $\mathbf{e}_i \in \mathbf{E}$ **do**
  Find k-nearest neighbours: $\text{kNN} \leftarrow \text{FindNearestNeighbours}(\mathbf{e}_i, \mathbf{E}, k)$
  Calculate uniqueness: $u_i \leftarrow \frac{1}{k} \sum_{j \in \text{kNN}} ||\mathbf{e}_i - \mathbf{e}_j||_2$
  $\mathbf{U} \leftarrow \mathbf{U} \cup \{u_i\}$
end for
Select top-N indices: $\mathcal{I} \leftarrow \text{ArgTopK}(\mathbf{U}, N)$
return $\mathcal{S} = \{x_i : i \in \mathcal{I}\}$

3.3 Stage 2: Semi-Automated Hybrid Annotation Pipeline

The second stage implements a sophisticated annotation workflow that balances automation capabilities with human expertise to achieve both efficiency and quality. This approach combines detection model proposals, human verification, and foundation model refinement into a cohesive pipeline that dramatically reduces the effort required for high-quality annotation.

Adaptive Image Preprocessing To ensure robust performance across diverse real-world imaging conditions, each selected image undergoes adaptive preprocessing designed to normalise quality variations and enhance feature visibility. This multi-stage enhancement pipeline proves particularly important for ecological monitoring data, which often suffers from challenging lighting conditions, motion blur, and equipment limitations.

The preprocessing sequence begins with super-resolution enhancement using Real-ESRGAN [34] for low-resolution images, improving detail preservation during subsequent analysis stages. Multi-Scale Retinex (MSR) corrects the non-uniform lighting conditions that commonly occur in outdoor monitoring scenarios, whilst edge-preserving bilateral filtering reduces sensor noise while maintaining structural features critical for accurate detection and segmentation.

Detector-Prompted Semi-Automatic Annotation The core annotation process leverages a strong pre-trained YOLOv8x6 model as a proposal generator (checkpoint `yolov8x6-oiv7.pt`), addressing limitations observed with general-purpose zero-shot detectors. The annotation workflow proceeds through four coordinated steps. Initially, the YOLOv8x6 model, pre-trained on Open Images V7 data containing relevant object categories including Penguin, generates initial bounding box proposals using multi-scale inference to ensure detection across varying object sizes. Domain experts then rapidly review these proposals, performing corrections (repositioning, resizing), additions (missed detections), and deletions (false positives). This human-in-the-loop verification leverages domain expertise for quality control whilst minimising manual effort through intelligent automation.

Each verified bounding box subsequently serves as a spatial prompt for the SAM2 model. The largest SAM2 model (`segment-anything-2.1-hiera-large`) was used to maximise the precision of the instance segmentation masks, capturing fine morphological details beyond simple or oriented rectangular bounds.

Pseudo-Labelling Quality Control The pseudo-labelling quality assessment incorporates multiple validation layers beyond confidence thresholding. Multi-scale inference consistency requires predictions to maintain IoU ≥ 0.5 across different resolution scales. Morphological constraints enforce biologically plausible aspect ratios (0.2–1.6) and normalised widths in $[0.005, 0.05]$ relative to image size. Predictions outside these bounds are routed to manual review and excluded from automatic approval, which empirically filters oversized artefacts without sacrificing recall – Multi-Scale Retinex (MSR) reduces these artefact occurrences. Data imbalance handling during pseudo-label integration employs class-aware sampling to prevent dominant class bias, maintaining balanced representation across different penguin poses and environmental conditions.

3.4 Stage 3: Progressive Training

The third stage implements an incremental training strategy that leverages both human-verified annotations and high-confidence model predictions to progressively expand the training dataset whilst maintaining rigorous quality standards. This semi-supervised approach enables efficient dataset growth beyond the manually annotated samples.

Architecture and Training Methodology Model training employs YOLOv11s-seg, representing the current state-of-the-art among unified detection-and-segmentation architectures [14]. Relative to prior YOLO variants, YOLOv11 introduces strengthened feature aggregation and modernised heads that improve optimisation stability and mask quality, yielding higher accuracy at comparable throughput on ecological imagery. Proposals for semi-automated labelling use YOLOv8x6 (OIV7-pretrained), while final training and evaluation use YOLOv11s-seg to capitalise on these architectural improvements.

The training regimen incorporates several advanced data augmentation strategies designed to improve model robustness across environmental variations. Built-in augmentations include standard geometric and photometric transformations such as rotation, scaling, colour space variations, and noise injection. A custom Albumentations pipeline [4] provides domain-specific augmentations addressing ecological imaging characteristics, including lighting variations and background diversity. Crucially, Copy-Paste augmentation [3, 12] is applied with probability 0.5, for intra- and inter-samples, and no blending (visually compatible backgrounds in this domain), which significantly improves instance boundary precision and resilience to domain shift. Copy-Paste augmentation process proved particularly effective on penguin imagery in previous work [3].

Training proceeds with image resolution 640px, batch size 8, AdamW optimiser with initial learning rate 0.001, cosine annealing schedule, and weight decay 0.0005. Early stopping with patience of 15 epochs prevents overtraining, whilst model checkpointing preserves optimal weights based on validation mAP. All hyperparameters were selected via Bayesian optimisation.

Following each training iteration in this active learning loop, the updated model performs inference on the remaining unlabelled dataset.

3.5 Final Stage: Systematic Model Optimisation and Deployment

The final stage ensures that the resulting model achieves optimal performance through systematic hyperparameter optimisation and comprehensive validation. This process transforms the research prototype into a production-ready system suitable for real-world deployment.

Upon convergence of the active learning loop, the Ray Tune framework [2] executes systematic hyperparameter search using Bayesian optimisation [18]. The search space encompasses critical training parameters including learning rates (5×10^{-4} to 5×10^{-3}), weight decay (1×10^{-5} to 1×10^{-3}), momentum (0.85–0.99), and augmentation intensities. This automated approach consistently outperforms manual tuning strategies whilst enabling systematic exploration of the hyperparameter landscape. All reported training settings in this work were obtained via this Bayesian tuning process.

The optimised model and curated dataset are exported in standardised YOLO format with consistent train/validation/test splits, ensuring reproducibility and facilitating deployment across different computational environments. The export process includes comprehensive metadata documentation, enabling reproducible model training and evaluation protocols that meet the standards required for scientific publication and practical deployment.

4 Experimental Setup

The comprehensive experimental protocol measures both annotation efficiency gains and model performance across multiple evaluation scenarios, including challenging cross-domain generalisation tests that simulate real-world deployment conditions.

4.1 Dataset Construction and Curation

The experimental dataset originates from the Open Images V7 (OIV7) dataset, providing a large-scale foundation for active learning experiments. Initial data extraction identified 991 images containing penguin-labelled instances, representing the largest publicly available collection of penguin imagery with bounding box annotations. This raw dataset underwent systematic quality control through a multi-stage filtering process designed to ensure high-quality experimental conditions.

Automated quality control began with programmatic, similarity-based deduplication using CLIP embeddings to remove near-identical images that could bias experimental results. Statistical outlier detection then flagged annotations with implausible geometric properties, including extreme aspect ratios or abnormal dimensions that likely indicate annotation errors. This automated screening eliminated 227 low-quality samples, followed by manual expert review to remove non-photorealistic images (illustrations, cartoons) and instances with fundamentally incorrect annotations. The result was a cleaned dataset of 764 high-quality photographs suitable for rigorous active learning evaluation.

The cleaned dataset was strategically partitioned into three distinct evaluation sets to enable comprehensive assessment. The hold-out test set comprises the 100 most visually unique images, identified through CLIP uniqueness scoring and manually annotated with high-quality segmentation masks by domain experts. These samples were sequestered exclusively for final model evaluation, ensuring unbiased performance assessment. The active learning pool contains the remaining 664 images, forming the unlabelled dataset for active learning experiments and simulating realistic deployment scenarios where large quantities of unlabelled data are available. Finally, an independent cross-domain evaluation set, provided by SANParks, contains African Penguin imagery collected from Boulders Penguin Colony – Table Mountain National Park under different environmental conditions and equipment configurations.

4.2 Cross-Domain Generalisation Assessment

Real-world deployment scenarios require models that generalise effectively across different environmental conditions, equipment configurations, and data distributions. The evaluation protocol employs completely independent SANParks field data never observed during training, providing authentic cross-domain assessment.

The primary domain encompasses models trained on the curated OIV7-derived dataset, representing diverse photographic conditions and penguin populations from various geographical locations. The target domain utilises the independent SANParks dataset, containing 124 African Penguin images collected under authentic field conditions with different camera equipment (static camera traps), adversarial environmental conditions, and colony-specific characteristics.

All trained models undergo zero-shot evaluation on the SANParks dataset without any domain-specific fine-tuning or adaptation, providing direct measurement of generalisation capability across environmental and equipment variations. This stringent evaluation protocol ensures that reported performance metrics reflect genuine model robustness rather than dataset-specific overfitting.

4.3 Evaluation Metrics and Statistical Methodology

Model performance evaluation employs standard computer vision metrics adapted for instance segmentation tasks, providing comprehensive assessment across different aspects of model capability. Primary performance metrics include Mean Average Precision at IoU thresholds of 0.5 (mAP_{50}), 0.75 (mAP_{75}), and averaged over IoU thresholds 0.5-0.95 (mAP_{50-95}), providing both lenient and stringent assessments of segmentation quality.

Runtime performance evaluation measures inference time per image on an NVIDIA RTX 4070 (12GB VRAM). All experiments are conducted across multiple independent runs with different random seeds (minimum 5 repetitions) to account for stochastic variation in training and evaluation processes and ensure reliable experimental conclusions.

5 Results and Analysis

The experimental evaluation demonstrates substantial improvements in annotation efficiency whilst maintaining competitive model performance across diverse evaluation scenarios. These results validate the effectiveness of the foundation model-driven approach and establish its practical viability for real-world deployment in conservation monitoring applications.

5.1 Active Learning Efficiency and Component Analysis

The effectiveness of the proposed framework is demonstrated through both convergence analysis and systematic component evaluation. Figure 2 highlights the sample-efficiency advantage of intelligent data selection, while Table 2 quantifies the contribution of each component.

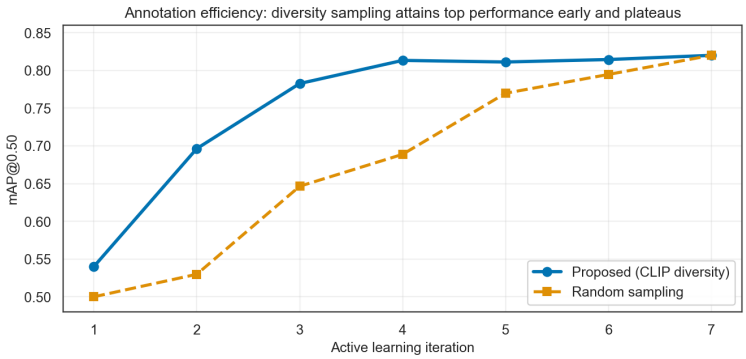

Fig. 2: Active learning convergence analysis demonstrating superior sample efficiency of CLIP diversity sampling. The proposed method reaches stable performance (0.82 mAP_{50}) within 4 iterations (≈400 labelled samples), whereas random sampling requires the full 7 iterations (≈664 labelled samples; final batch smaller) to achieve comparable performance – reducing annotation effort by roughly 40–45% at matched accuracy. Results shown are base performance without SAM2 refinement; the complete framework with SAM2 achieves 0.88 mAP_{50}.

The convergence gap in Figure 2 provides compelling evidence for the value of foundation model-driven data selection: by achieving target accuracy with substantially fewer labelled samples, the framework directly

alleviates the annotation bottleneck that constrains real-world deployment in conservation contexts.

Component-wise Performance Attribution: The ablations reveal distinct mechanistic contributions:

1. *CLIP diversity sampling* achieves equivalent full-dataset performance (0.82 mAP_{50}) using only 60% of training data by prioritising environmental and contextual coverage early in training, reducing overfitting to spurious backgrounds and accelerating convergence.
2. *SAM2 refinement* provides consistent +0.04–0.06 mAP_{50} improvements across all configurations by generating higher-fidelity training masks that better capture morphological boundaries and handle partial occlusions.
3. *Copy-Paste augmentation* contributes +0.03 mAP_{50} improvement in cross-domain scenarios by synthesising realistic instance co-occurrence patterns and background variations not present in the original training distribution.
4. *Quality-controlled pseudo-labelling* maintains precision during dataset expansion by filtering predictions through multi-scale consistency checks and morphological plausibility constraints, preventing error amplification that typically degrades semi-supervised approaches.

Synergistic Effects: The complete framework (CLIP + SAM2 + Copy-Paste) achieves 0.88 mAP_{50} – a 7% improvement over the best individual component – demonstrating that foundation model integration creates complementary advantages rather than simply additive performance gains.

Table 2: Comprehensive ablation study revealing individual component contributions and synergistic effects. Intelligent data selection (CLIP) matches full-data training with 40% fewer labels; SAM2 refinement consistently adds 0.04–0.06 mAP_{50}; together they yield the highest overall performance.

Configuration	mAP_{50}	mAP_{75}	mAP_{50-95}	F1-Score	Training Samples
Baseline (All Data)	0.82	0.69	0.64	0.78	664
+ SAM2 Refinement	0.86	0.75	0.69	0.80	664
Baseline (Random)	0.69	0.58	0.53	0.66	400
+ SAM2 Refinement	0.75	0.69	0.60	0.71	400
CLIP Diversity	**0.82**	**0.71**	**0.65**	**0.78**	**400**
+ SAM2 Refinement	**0.88**	**0.77**	**0.71**	**0.84**	**400**

5.2 Cross-Domain Generalisation Analysis

Cross-domain evaluation on the independent SANParks dataset reveals strong generalisation capabilities, addressing a critical limitation of traditional supervised learning approaches in ecological applications where models must operate across diverse environmental conditions.

Table 3: Cross-domain generalisation performance on SANParks African Penguin dataset. Results represent zero-shot evaluation without domain-specific fine-tuning.

Method	mAP_{50}	mAP_{75}	mAP_{50-95}	F1-Score
Random	0.67	0.56	0.53	0.64
+ Copy-Paste	0.71	0.63	0.60	0.69
Proposed	0.81	0.71	0.68	0.75
+ Copy-Paste	0.84	0.75	0.71	0.81

The superior cross-domain performance demonstrates several key advantages of the proposed approach. CLIP diversity sampling naturally selects training samples that span broader environmental and contextual

variations, improving model robustness to domain shift. The integration of SAM2-generated segmentations provides higher-quality training masks that better capture morphological variations, enhancing generalisation to new environments. Additionally, the progressive learning approach with quality-controlled pseudo-labelling exposes the model to progressively diverse examples throughout training, building robustness compared to single-batch training approaches.

5.3 Failure Case Analysis

The methodology encounters challenges in specific scenarios that warrant systematic analysis. A detailed evaluation of 124 SANParks images reveals distinct failure modes with quantifiable impacts on performance:

1. **Environmental Degradation (18% of failures):** Extremely low-light conditions (luminance < 30 cd/m^2) reduce CLIP embedding discriminability, leading to suboptimal diversity sampling that misses critical environmental variations. These cases show 15–20% reduced mAP_{50} compared to optimal lighting conditions.
2. **Occlusion Complexity (31% of failures):** Dense clustering scenarios where individual penguins are obscured by $> 70\%$ result in SAM2 segmentation failures, as the model's spatial attention mechanisms cannot reliably separate overlapping instances. Manual analysis shows IoU degrades from 0.85 (clear instances) to 0.62 (heavily occluded).
3. **Distribution Shift (28% of failures):** Atypical poses or behaviours not represented in the Open Images V7 training distribution (e.g., swimming, diving, aggressive posturing) produce systematically low-confidence predictions (< 0.4 confidence threshold). However, the main problem was due to footage from dirty camera lenses. This highlights the fundamental limitation of training data coverage rather than methodological deficiencies.
4. **Technical Limitations (23% of failures):** Motion blur (> 3 pixel displacement), extreme viewing angles (> 60 from horizontal), and equipment-specific artefacts (lens distortion, chromatic aberration) occasionally compromise both detection and segmentation quality.

These failure modes demonstrate that while the methodology achieves robust performance across diverse conditions, systematic data collection strategies must account for environmental and behavioural coverage to minimise out-of-distribution scenarios during deployment.

Figure 3 illustrates the range of segmentation quality achieved on the independent SANParks dataset. The high-quality predictions shown in Figure 3a demonstrate the methodology's capability to produce precise instance masks with clean boundaries that accurately capture individual penguin morphology, even in complex multi-instance scenarios with partial occlusions. These results validate the effectiveness of the SAM2-refined annotation pipeline in generating training data that translates to robust real-world performance. Conversely, Figure 3b reveals the methodology's limitations under challenging conditions, where correct penguin detection is achieved but polygon boundaries exhibit reduced precision. These cases typically occur in scenarios with extreme lighting variations, dense clustering, or significant background complexity that exceed the model's learned feature representations. Despite these occasional quality degradations, the overall cross-domain performance remains substantially superior to baseline approaches, demonstrating the robustness of the foundation model-driven training pipeline.

6 Discussion

The experimental results demonstrate measurable improvements across multiple evaluation scenarios whilst highlighting specific limitations. CLIP diversity sampling achieves mAP_{50} of 0.88 using 400 training samples (with SAM2 refinement), compared to 0.75 for random sampling with equivalent sample counts – a 17% improvement that validates semantic embeddings' superior data selection capabilities. Notably, intelligent data selection exceeds full-dataset performance (0.86 with 664 samples) whilst requiring 40% fewer annotations.

The complete framework demonstrates practical viability through synergistic component integration: SAM2 refinement provides 7% improvement over CLIP diversity alone, whilst human-in-the-loop verification maintains annotation integrity in challenging scenarios. Cross-domain evaluation reveals robust generalisation, achieving 0.81 mAP_{50} on independent SANParks data (17% improvement over baselines), enhanced

(a) High-quality segmentation predictions demonstrating precise polygon boundaries and accurate instance detection across multiple penguins with proper handling of occlusion scenarios.

(b) Suboptimal segmentation quality showing correct penguin detection but imprecise polygon boundaries, demonstrating limitations in challenging environmental conditions.

Fig. 3: Cross-domain prediction quality comparison on SANParks field data. The proposed methodology achieves high-quality segmentation in optimal conditions (left) but occasionally produces correct detections with degraded polygon precision in challenging scenarios (right).

to 0.84 with Copy-Paste augmentation. These findings align with recent deployable field-scale pipelines [9], reinforcing end-to-end data-centric systems' importance.

These results directly address fundamental scalability limitations preventing widespread adoption of instance segmentation in ecological contexts, enabling deployment across multiple monitoring sites without extensive retraining campaigns. However, several limitations constrain broader applicability: focus on single-species scenarios requiring extension to multi-species contexts with class imbalance challenges; dependency on foundation model APIs limiting offline deployment; performance degradation under extreme conditions (low-light degrading CLIP embeddings, heavy occlusion compromising SAM2 segmentation); and validation primarily on penguin data requiring empirical verification across species with varying morphological characteristics, behaviour patterns, and environmental contexts. Future work must address these limitations through expanded taxonomic validation and species-agnostic methodological components.

7 Conclusion

This research demonstrates that foundation model-driven active learning substantially reduces annotation requirements for instance segmentation in ecological monitoring whilst maintaining superior performance. The integrated methodology combining CLIP diversity sampling, semi-automated annotation pipelines, and progressive training with quality-controlled pseudo-labelling achieves statistically significant improvements across evaluation scenarios.

Experimental validation reveals intelligent data selection using CLIP embeddings achieves mAP_{50} of 0.88 with 400 training samples (including SAM2 refinement), compared to 0.75 for random sampling – a 17% improvement exceeding full-dataset training performance (0.86 with 664 samples) whilst requiring 40% fewer annotations. Cross-domain evaluation demonstrates robust generalisation, achieving 0.81–0.84 mAP_{50} compared to 0.67–0.71 for baselines – a 21–25% improvement validating practical deployment potential.

The systematic integration addresses fundamental scalability limitations in conservation monitoring where expert annotation time represents a critical bottleneck. By demonstrating semantic embeddings enable superior data selection, this work provides empirical evidence for data-centric model development in specialised domains. The semi-automated pipeline maintains quality standards whilst reducing manual effort, creating practical pathways for deploying sophisticated computer vision in resource-constrained conservation contexts. Strong cross-domain generalisation proves particularly significant for real-world deployment across diverse conditions, reducing expensive site-specific retraining campaigns. The workflow operationalises conservation

AI principles, making trait-relevant information computable from field imagery via data-centric pipelines robust to domain shift [9].

However, validation remains constrained to single-species scenarios using penguin data; broader taxonomic applicability requires empirical verification. Foundation model dependency may limit offline deployment, whilst performance limitations under extreme imaging conditions warrant deployment planning consideration.

Future research includes extending to multi-species contexts with class imbalance handling; investigating lightweight alternatives to foundation model dependencies; developing automated quality control mechanisms; exploring online learning for continuously evolving environments; and investigating image diffusion preprocessing for dirty camera lens scenarios. The demonstrated success provides methodological foundation for broader conservation technology application where annotation scarcity constrains advanced computer vision deployment, establishing a theoretically grounded, empirically validated framework for organisations leveraging state-of-the-art instance segmentation under realistic resource constraints.

Acknowledgments. Thank you to the South African National Parks (SANParks) for providing African Penguin data. This work was undertaken in the Distributed Multimedia CoE at Rhodes University.

Disclosure of Interests. The authors have no competing interests.

References

1. Bearman, A., Russakovsky, O., Ferrari, V., Fei-Fei, L.: What's the point: Semantic segmentation with point supervision. In: European Conference on Computer Vision, pp. 549–565, Springer (2016)
2. Bergstra, J., Yamins, D., Cox, D.: Making a science of model search: Hyperparameter optimization in hundreds of dimensions for vision architectures. In: Dasgupta, S., McAllester, D. (eds.) Proceedings of the 30th International Conference on Machine Learning, Proceedings of Machine Learning Research, vol. 28, pp. 115–123, PMLR, Atlanta, Georgia, USA (17–19 Jun 2013)
3. Brown, D., Bradshaw, K.: Augmenting data sets using generalized copy-paste for improved penguin segmentation across environments. In: Proceedings of the 2025 International Conference on Artificial Intelligence, Big Data, Computing and Data Communication Systems (icABCD), pp. 1–8, ACM (2025), https://doi.org/10.1145/3759023.3759124
4. Buslaev, A., Iglovikov, V.I., Khvedchenya, E., Parinov, A., Druzhinin, M., Kalinin, A.A.: Albumentations: Fast and flexible image augmentations. Information **11**(2), 125 (2020)
5. Chen, K., Pang, J., Wang, J., Xiong, Y., Li, X., Sun, S., Feng, W., Liu, Z., Shi, J., Ouyang, W., et al.: Hybrid task cascade for instance segmentation. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition pp. 4974–4983 (2019)
6. Chieza, K., Brown, D., Connan, J., et al.: Automated fish detection in underwater environments: Performance analysis of YOLOv8 and YOLO-NAS. Artificial Intelligence Research pp. 334–351 (2024)
7. De Silva, M., Brown, D.: Early detection of wheat stripe mosaic virus using multispectral imaging with deep-learning. Ecological Informatics **87**, 103088 (2025)
8. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: ImageNet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255, IEEE (2009)
9. Duporge, I., Kholiavchenko, M., Harel, R., Wolf, S., Rubenstein, D.I., Crofoot, M.C., Berger-Wolf, T., Lee, S.J., Barreau, J., Kline, J., et al.: Baboonland dataset: Tracking primates in the wild and automating behaviour recognition from drone videos: I. duporge et al. International Journal of Computer Vision pp. 1–12 (2025)
10. Freund, Y., Seung, H.S., Shamir, E., Tishby, N.: Selective sampling using the query by committee algorithm. Machine Learning **28**, 133–168 (1997), https://doi.org/10.1023/A:1007330508534
11. Ganin, Y., Ustinova, E., Ajakan, H., Germain, P., Larochelle, H., Laviolette, F., March, M., Lempitsky, V.: Domain-adversarial training of neural networks. Journal of Machine Learning Research **17**(59), 1–35 (2016), https://doi.org/10.5555/2946645.2946704
12. Ghiasi, G., Cui, Y., Srinivas, A., Qian, R., Lin, T.Y., Cubuk, E.D., Le, Q.V., Zoph, B.: Simple copy-paste is a strong data augmentation method for instance segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2918–2928 (2021)
13. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask R-CNN. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2961–2969 (2017)
14. Jocher, G., Chaurasia, A., Qiu, J.: YOLO by Ultralytics (2023), URL https://github.com/ultralytics/ultralytics
15. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 4015–4026 (2023)
16. LaRue, M., Iles, D., Labrousse, S., Fretwell, P., Ortega, D., Devane, E., Horstmann, I., Viollat, L., Foster-Dyer, R., Le Bohec, C., et al.: Advances in remote sensing of emperor penguins: First multi-year time series documenting trends in the global population. Proceedings of the Royal Society B **291**(2018), 20232067 (2024), https://doi.org/10.1098/rspb.2023.2067
17. Lewis, D.D., Gale, W.A.: A sequential algorithm for training text classifiers. In: SIGIR'94, pp. 3–12, Springer (1994)
18. Liaw, R., Liang, E., Nishihara, R., Moritz, P., Gonzalez, J.E., Stoica, I.: Tune: A research platform for distributed model selection and training. arXiv preprint arXiv:1807.05118 (2018)
19. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2117–2125 (2017)
20. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: Common objects in context. In: European Conference on Computer Vision, pp. 740–755, Springer (2014)
21. Lynch, H.J., White, R., Black, A.D., Naveen, R.: Detection, differentiation, and abundance estimation of penguin species by high-resolution satellite imagery. Polar Biology **35**, 963–968 (2012), https://doi.org/10.1007/s00300-011-1138-3
22. Marsden, M., McGuinness, K., Little, S., Keogh, C.E., O'Connor, N.E.: People, penguins and petri dishes: Adapting object counting models to new visual domains and object types without forgetting. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 8070–8079 (2018)

23. Norouzzadeh, M.S., Morris, D., Beery, S., Joshi, N., Jojic, N., Clune, J.: A deep active learning system for species identification and counting in camera trap images. Methods in Ecology and Evolution **12**(1), 150–161 (2021), https://doi.org/https://doi.org/10.1111/2041-210X.13504
24. Norouzzadeh, M.S., Nguyen, A., Kosmala, M., Swanson, A., Palmer, M.S., Packer, C., Clune, J.: Automatically identifying, counting, and describing wild animals in camera-trap images with deep learning. Proceedings of the National Academy of Sciences **115**(25), E5716–E5725 (2018), https://doi.org/10.1073/pnas.1719367115
25. Papadopoulos, D.P., Uijlings, J.R., Keller, F., Ferrari, V.: Training object class detectors with click supervision. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition pp. 6374–6383 (2017)
26. Petersen, S.L., Ryan, P.G., Grémillet, D.: Is food availability limiting African penguins *Spheniscus demersus* at Boulders? A comparison of foraging effort at mainland and island colonies. Ibis **148**(1), 14–26 (2006)
27. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: Proceedings of the International Conference on Machine Learning, pp. 8748–8763, PMLR (2021)
28. Ravi, N., Gabeur, V., Hu, Y.T., Hu, R., Ryali, C., Ma, T., Khedr, H., Römzn, R., Rolland, C., Gustafson, L., et al.: SAM 2: Segment anything in images and videos. arXiv preprint arXiv:2408.00714 (2024)
29. Saleh, A., Sheaves, M., Jerry, D., Azghadi, M.R.: A realistic fish-habitat dataset to evaluate algorithms for underwater visual analysis. Scientific Reports **10**(1), 14671 (2020)
30. Salie, D., Brown, D., Chieza, K.: Deep neural network compression for lightweight and accurate fish detection. Artificial Intelligence Research pp. 300–318 (2024)
31. Scheffer, T., Decomain, C., Wrobel, S.: Active learning for logistic regression: An evaluation. In: Machine Learning, vol. 68, pp. 235–265, Springer (2001)
32. Sener, O., Savarese, S.: Active learning for convolutional neural networks: A core-set approach. In: International Conference on Learning Representations (2018)
33. Trathan, P.N.: Image analysis of color aerial photography to estimate penguin population size. Wildlife Society Bulletin **32**(2), 332–343 (2004)
34. Wang, X., Xie, L., Dong, C., Shan, Y.: Real-ESRGAN: Training real-world blind super-resolution with pure synthetic data. In: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, pp. 1905–1914 (2021)
35. Zha, D., Bhat, Z.P., Lai, K.H., Yang, F., Jiang, Z., Zhong, S., Hu, X.: Data-centric AI: A survey. Communications of the ACM **67**(3), 86–99 (2024)
36. Zhang, L., Gao, J., Xiao, Z., Fan, H.: Animaltrack: A benchmark for multi-animal tracking in the wild. International Journal of Computer Vision **131**(2), 496–513 (2023)