Continually Adapt or Not (CAN)? A Continual Learning Benchmark of Camera Trap Species Classification over Time

Anonymous Author(s)
Affiliation
Address
email

Abstract

Camera traps offer an effective, non-invasive approach to wildlife monitoring. However, substantial variations in image style across camera setups, combined with temporal shifts in image content, pose significant challenges to developing accurate and robust image recognition models. We present a novel benchmark for these challenges, leveraging data from 546 camera traps across 17 LILA BC datasets. We introduce a systematic data preparation pipeline inspired by the FAIR principles and formulate the task as an instance of online continual learning to better reflect the practical usage of camera traps. This approach sharply contrasts with prior studies that typically disregard the chronological structure of the data. Our study reveals several critical insights. First, using the latest vision foundation model for biological domains, BioCLIP 2, we observe a long-tailed accuracy distribution across the 546 camera traps, highlighting the persistent need for model adaptation. Second, continual adaptation is generally necessary to address temporal shifts, but the required adaptation frequency may decrease over time. Third, we identify several unresolved machine learning challenges from a practical standpoint and suggest directions for future research.

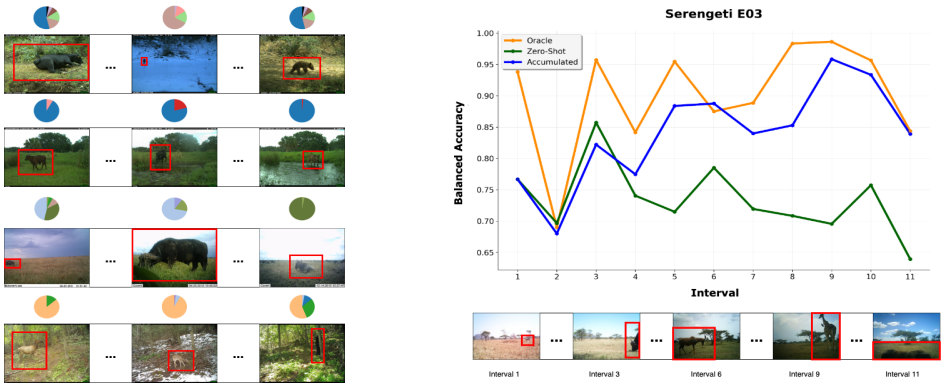
1 Introduction

Camera traps are a critical tool in ecological and wildlife research for non-invasively capturing large volumes of time-stamped images in natural habitats (Pollock et al., 2025; Tuia et al., 2022). These images support biodiversity monitoring, behavior analysis, and conservation planning but also vary greatly across space, time, hardware, and deployment strategies (Koh et al., 2021; Beery et al., 2021), creating major challenges for automated analysis.

Prior work typically frames this as domain adaptation or generalization—transferring models from certain source domains to unseen targets (Sagawa et al., 2021; Zhou et al., 2022)—yet this overlooks the practical needs of ecological practitioners. In the field, the central questions are: *Will the model work at a new location? How much data is needed to adapt it? Must it be continually updated?* These challenges are intensified by passive, slow data collection and incomplete species coverage (Tu et al., 2023).

To address these needs, we introduce the **Continually Adapt or Not (**CAN**)** benchmark for **camera-trap species classification over time**. We split each camera's image stream into sequential **time intervals** and pose the task as **online continual learning** (Mai et al., 2022): at interval j, a model may be updated on training data and then evaluated on test data from interval $j+1$, mimicking real deployments.

Submitted to 38th Conference on Neural Information Processing Systems (NeurIPS 2024). Do not distribute.



(a) (b)

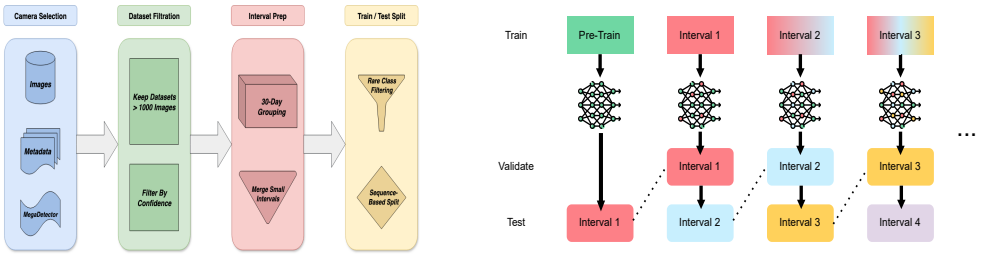
Figure 1: **(a)** Illustration of the variability of camera trap images across space (rows) and time (columns), with pie charts representing the image distributions across species. **(b)** Illustration of the three baselines: zero-shot, oracle, and accumulated. The accumulated model is trained on data from all intervals before the j-th interval and evaluated on the j-th interval.

We provide a reproducible, FAIR-compliant pipeline to prepare the benchmark and derive metrics such as temporal shift and class imbalance. We then evaluate three baselines—zero-shot BioCLIP 2, an upper-bound oracle, and an accumulated continual model—yielding insights into temporal drift, imbalance, and the limits of naïve fine-tuning.

2 "Continually Adapt or Not" Benchmark

2.1 Motivation

As shown in Figure 1a and Figure 4, camera trap images show substantial variability. Image style and quality vary widely across locations—some are blurry, others capture animals at close range, and lighting or resolution may differ. Even within one location, seasonal and temporal shifts can change both background appearance and species distribution. These variations make it challenging to develop accurate, robust classifiers and raise practical concerns for end-users: *Will the model generalize to my setting, or require further adaptation?* To address this, we introduce the **Continually Adapt or Not** (CAN) benchmark—a curated testbed for evaluating pre-trained models and fostering adaptation algorithms in the camera trap domain. For overall details, refer Appendix B.



(a) (b)

Figure 2: **(a)** Data processing pipeline (details in Appendix B). **(b)** Online continual learning setup.

2.2 Data Source and Processing Pipeline

CAN is built on the LILA BC repository (LILA BC), which aggregates dozens of camera-trap datasets (e.g., Ohio Small Animals, Snapshot Karoo) collected by hundreds of stationary cameras deployed worldwide. For our benchmark we select 17 datasets that meet minimum size and duration criteria and then process them using a standardized pipeline designed to support FAIR principles—Findable, Accessible, Interoperable, and Reusable. This high-level process is summarized in Figure 2a, and the resulting camera-trap coverage and dataset characteristics are illustrated in Figure 3a and Figure 3b.

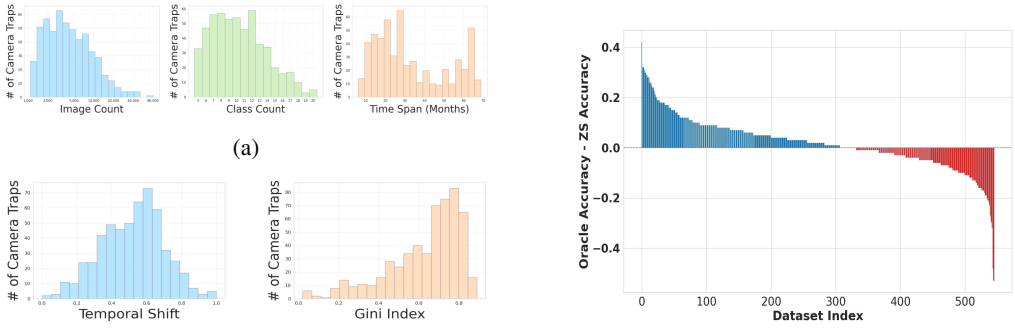


Figure 3: **(a)** Camera trap data statistics (histograms). **(b)** Histograms characterizing temporal shift and class imbalance. **(c)** Oracle vs. ZS performance gap across datasets.

2.3 Online Continual Learning Task

Unlike conventional domain adaptation, where data from the new, target domain is available all at once Gong et al. (2012); Singhal et al. (2023), CAN adopts an **online continual learning** setting Mai et al. (2022) to better reflect the practical deployment of camera traps, where new data arrives sequentially over time. This setup—illustrated in Figure 2b—evaluates models sequentially on each new time interval and then updates them before the next interval, mimicking real-world use.

2.4 Baseline Methods

Setting We adopt a *closed-set* setting in which the species expected at each camera trap are assumed known from local or historical data. Model performance is measured as balanced accuracy—per-class accuracy within each interval, averaged across all intervals.

Baseline methods. We evaluate three baselines: the *Zero-shot Model*, which uses BioCLIP 2 directly without additional training to match images to species text embeddings; the *Oracle Model*, which combines the BioCLIP 2 vision encoder with a linear classifier trained on data from all intervals simultaneously; and the *Accumulated Model*, which follows the oracle setup but is trained incrementally using only past intervals at each step (see Appendix C for details).

2.5 Results and Analysis

We begin by comparing the zero-shot and oracle performance across all 546 camera traps. Figure 3c summarizes the results, revealing three key observations.

First, with BioCLIP 2, over 170 camera traps (31%) achieve accuracy above 90%, underscoring the strong potential of foundation models in wildlife monitoring. Second, 169 camera traps (31%) still fall below 80% accuracy, highlighting the need for adaptation to improve model performance in more challenging scenarios. Third, the oracle model does not consistently outperform the zero-shot model—even when the latter underperforms—suggesting that effective adaptation hinges on a deeper understanding of the intrinsic properties of each camera trap dataset.

Why does BioCLIP2 underperform on some datasets? To understand the failure cases, we examine camera traps where BioCLIP 2 performs poorly (see Figure 4). While some performance gaps may stem from the model's difficulty in recognizing certain species, many of these traps also suffer from poor image quality—low resolution, poor lighting, heavy occlusion, or motion blur. In some cases, animals appear too close to the camera, resulting in crops that capture only partial views of the subject. These factors likely hinder recognition. Analyzing such low-performing cases can offer practical insights into how practitioners might better deploy camera traps in the field.

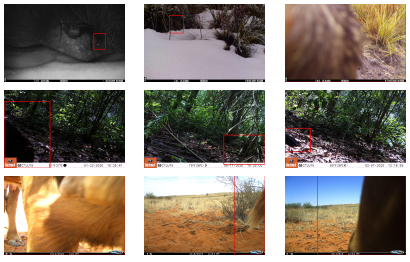


Figure 4: Very challenging cases for the zero-shot model.

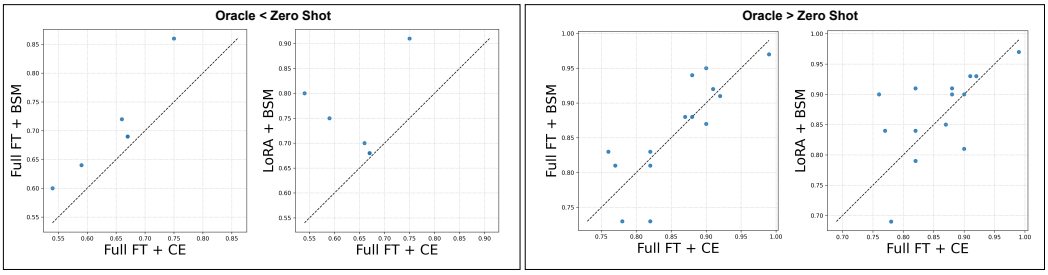
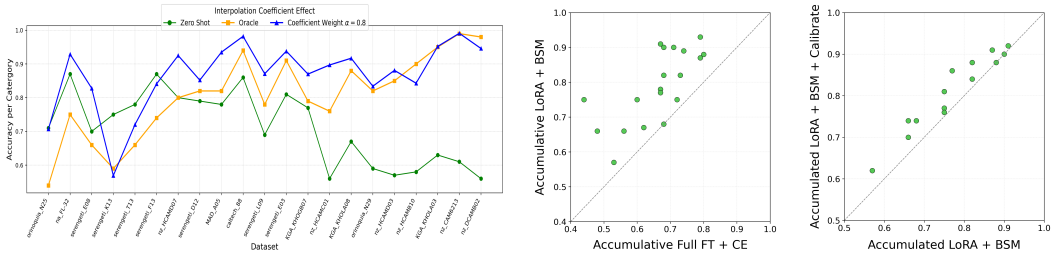


Figure 5: Oracle model improvement by the BSM loss and LoRA. As shown, BSM loss and LoRA consistently improve the oracle when it is outperformed by zero-shot.

Why does the oracle model underperform on some datasets? Although the oracle model is trained on all available data, it sometimes fails to outperform the zero-shot BioCLIP 2 baseline. This may be due to a combination of factors, including severe class imbalance and limited training samples for certain categories. In such cases, fine-tuning with standard cross-entropy loss can lead the model to drift away from the generalizable representations learned by BioCLIP 2. These observations suggest that full fine-tuning alone may not be sufficient for effective adaptation and highlight the need for more robust strategies tailored to low-data or imbalanced regimes.

3 Going Deep into the CAN Benchmark

We now take a closer look at model adaptation within CAN from two complementary viewpoints. First, we adopt an *algorithm developer's perspective*, highlighting approaches for improving and adapting models across camera traps. Second, we take an *end-user's perspective*, focusing on practical questions about when zero-shot predictions suffice and when continued adaptation may be needed.



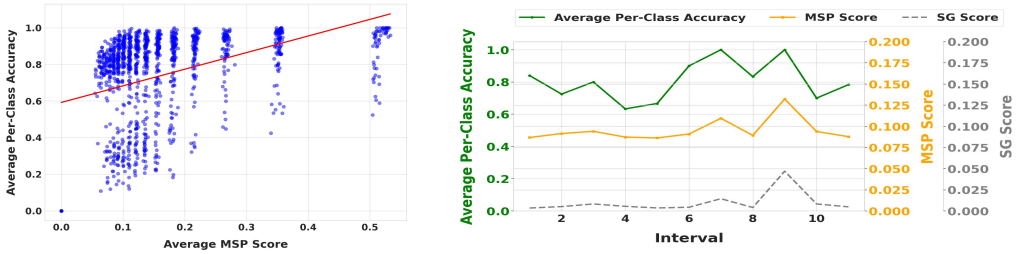
(a) (b) (c)

Figure 6: **(a)** Weight interpolation between the zero-shot and oracle models shows consistent improvement. **(b)** Accumulated model accuracy. BSM loss with LoRA outperforms CE loss with full fine-tuning. **(c)** Calibration consistently improves accumulated models.

3.1 An Algorithm Developer's Perspective

From an algorithm developer's point of view, we summarize general strategies for adapting models within CAN. This includes high-level ideas for improving oracle and accumulated models, with full methodology and empirical details referred to Appendix D.

Oracle Model Improvements. We explore three main approaches to strengthen the oracle model: (i) *Balanced Softmax (BSM) loss* (Ren et al., 2020); (ii) *parameter-efficient fine-tuning (LoRA)* (Mai et al., 2025; Hu et al., 2022); and (iii) *weight interpolation (WiSE)* (Wortsman et al., 2022). As shown in Figure 5, both BSM and LoRA generally improve oracle models, especially when naïve fine-tuning underperforms the zero-shot baseline. However, when the oracle already surpasses zero-shot, BSM can occasionally degrade performance, underscoring that class-balancing strategies are not universally beneficial and require dataset-specific tuning. In Figure 6a, WiSE consistently improves results, often outperforming both the fine-tuned oracle and the zero-shot baseline. Since it is a post-hoc method requiring no extra training, WiSE provides a simple and practical way to boost performance.



(a) (b)

Figure 7: **(a)** Non-OOD scores correlate with ZS accuracy.). **(b)** Non-OOD scores (MSP or SG) correlate with accumulated model accuracy (across intervals of a single camera trap)

Accumulated Model Improvements. Apply the same techniques (BSM loss, LoRA) during continual adaptation to mitigate imbalance and data scarcity; add post-hoc calibration to restore recognition of absent classes; and use WiSE interpolation to further boost performance. We evaluate this approach using the optimal γ (Mai et al., 2024). As illustrated in Figure 6b and Figure 6c, applying BSM loss and LoRA consistently improves the performance of accumulated models, while calibration further boosts their accuracy in the early intervals.

3.2 An End-User's Perspective

From an end-user's point of view, the key question is when a zero-shot model is "good enough" and when continual adaptation is worthwhile. Below we highlight the main considerations, while all numerical analyses and implementation details appear in Appendix E.

When is the zero-shot model sufficient? Pre-trained models generally perform well when the test data resembles their training data, but their predictions degrade under unseen or highly shifted conditions. In Figure 7a we show that simple confidence or non-OOD scores correlate with zero-shot accuracy across camera traps, suggesting a practical way to anticipate whether zero-shot predictions will be reliable before deployment.

Do we need to continually adapt? We examine whether adaptation must occur after every interval or if it can be paused. Table 1 shows that accumulated models often retain strong performance after a few updates but eventually lag behind continually adaptive models, highlighting the long-term benefit of ongoing updates.

When should we adapt? We further test whether easily computed confidence scores can signal when adaptation will pay off. As illustrated in Figure 7b, both the Maximum Softmax Probability (MSP) and the Softmax Gap (SG) scores tend to rise and fall with accumulated-model accuracy, with SG showing stronger correlations. These results point to a promising direction for using lightweight indicators to guide fine-tuning decisions in practice.

Accummulated	Accuracy
up to 33%	0.582
up to 67%	0.686
up to 100%	0.761

Table 1: Accuracy of models continually trained up to certain intervals (average over 15 camera traps).

4 Conclusion and Discussion

We introduce a novel continual learning benchmark that reflects the real-world challenges of adapting visual recognition models to camera trap deployments. Our empirical studies demonstrate that successful adaptation relies on the thoughtful application of targeted machine learning techniques, yielding valuable insights for system-level adaptation in dynamic environments.

Looking ahead, we hope this benchmark will serve as a catalyst for advancing adaptive machine learning at the system level. Rather than assuming full access to labeled data at each interval for fine-tuning, future approaches should incorporate mechanisms to actively select both intervals and instances for human annotation and storage—enabling scalable and sustainable continual learning. We also encourage the evaluation of future vision foundation models on this benchmark to assess their robustness and applicability in real-world, evolving settings.

References

Rahaf Aljundi, Min Lin, Baptiste Goujaud, and Yoshua Bengio. Gradient based sample selection for online continual learning. *Advances in neural information processing systems*, 32, 2019.

Reza Averly and Wei-Lun Chao. Unified out-of-distribution detection: A model-specific perspective. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1453–1463, 2023.

Sara Beery. The megadetector: Large-scale deployment of computer vision for conservation and biodiversity monitoring. *California Institute of Technology, Pasadena, CA, USA*, 2023.

Sara Beery, Grant Van Horn, Oisin Mac Aodha, and Pietro Perona. The iwildcam 2018 challenge dataset. *arXiv preprint arXiv:1904.05986*, 2019.

Sara Beery, Arushi Agarwal, Elijah Cole, and Vighnesh Birodkar. The iwildcam 2021 competition dataset. *arXiv preprint arXiv:2105.03494*, 2021.

Peggy A Bevan, Omiros Pantazis, Holly Pringle, Guilherme Braga Ferreira, Daniel J Ingram, Emily Madsen, Liam Thomas, Dol Raj Thanet, Thakur Silwal, Santosh Rayamajhi, et al. Deep learning-based ecological analysis of camera trap images is impacted by training data quality and quantity. *arXiv preprint arXiv:2408.14348*, 2024.

Luigi Boitani. *Camera trapping for wildlife research*. Pelagic Publishing Ltd, 2016.

Yin Cui, Menglin Jia, Tsung-Yi Lin, Yang Song, and Serge Belongie. Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9268–9277, 2019.

Matthias De Lange, Rahaf Aljundi, Marc Masana, Sarah Parisot, Xu Jia, Aleš Leonardis, Gregory Slabaugh, and Tinne Tuytelaars. A continual learning survey: Defying forgetting in classification tasks. *IEEE transactions on pattern analysis and machine intelligence*, 44(7):3366–3385, 2021.

Zalan Fabian, Zhongqi Miao, Chunyuan Li, Yuanhan Zhang, Ziwei Liu, Andrés Hernández, Andrés Montes-Rojas, Rafael Escucha, Laura Siabatto, Andrés Link, et al. Multimodal foundation models for zero-shot animal species recognition in camera trap images. *arXiv preprint arXiv:2311.01064*, 2023.

Abolfazl Farahani, Sahar Voghoei, Khaled Rasheed, and Hamid R Arabnia. A brief review of domain adaptation. *Advances in data science and information engineering: proceedings from ICDATA 2020 and IKE 2020*, pages 877–894, 2021.

Valentin Gabeff, Marc Rußwurm, Devis Tuia, and Alexander Mathis. Wildclip: Scene and animal attribute retrieval from camera trap data with domain-adapted vision-language models. *International Journal of Computer Vision*, 132(9):3770–3786, 2024.

Boqing Gong, Yuan Shi, Fei Sha, and Kristen Grauman. Geodesic flow kernel for unsupervised domain adaptation. In *2012 IEEE conference on computer vision and pattern recognition*, pages 2066–2073. IEEE, 2012.

Jianyang Gu, Samuel Stevens, Elizabeth G Campolongo, Matthew J Thompson, Net Zhang, Jiaman Wu, Andrei Kopanev, Zheda Mai, Alexander E White, James Balhoff, et al. Bioclip 2: Emergent properties from scaling hierarchical contrastive learning. *arXiv preprint arXiv:2505.23883*, 2025.

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *arXiv preprint arXiv:1610.02136*, 2016.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, et al. Wilds: A benchmark of in-the-wild distribution shifts. In *International conference on machine learning*, pages 5637–5664. PMLR, 2021.

LILA BC. LILA BC: Labeled information library of alexandria: Biology and conservation. `https://lila.science/`. Accessed: 2025-09-27.

Zheda Mai, Ruiwen Li, Hyunwoo Kim, and Scott Sanner. Supervised contrastive replay: Revisiting the nearest class mean classifier in online class-incremental continual learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3589–3599, 2021.

Zheda Mai, Ruiwen Li, Jihwan Jeong, David Quispe, Hyunwoo Kim, and Scott Sanner. Online continual learning in image classification: An empirical survey. *Neurocomputing*, 469:28–51, 2022.

Zheda Mai, Arpita Chowdhury, Ping Zhang, Cheng-Hao Tu, Hong-You Chen, Vardaan Pahuja, Tanya Berger-Wolf, Song Gao, Charles Stewart, Yu Su, et al. Fine-tuning is fine, if calibrated. *Advances in Neural Information Processing Systems*, 37:136084–136119, 2024.

Zheda Mai, Ping Zhang, Cheng-Hao Tu, Hong-You Chen, Quang-Huy Nguyen, Li Zhang, and Wei-Lun Chao. Lessons and insights from a unifying study of parameter-efficient fine-tuning (peft) in visual recognition. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 14845–14857, 2025.

Farjad Malik, Simon Wouters, Ruben Cartuyvels, Erfan Ghadery, and Marie-Francine Moens. Two-phase training mitigates class imbalance for camera trap image classification with cnns. *arXiv preprint arXiv:2112.14491*, 2021.

Yifei Ming, Ziyang Cai, Jiuxiang Gu, Yiyou Sun, Wei Li, and Yixuan Li. Delving into out-of-distribution detection with vision-language representations. *Advances in neural information processing systems*, 35: 35087–35102, 2022.

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*, 115(25):E5716–E5725, 2018.

Laura J Pollock, Justin Kitzes, Sara Beery, Kaitlyn M Gaynor, Marta A Jarzyna, Oisin Mac Aodha, Bernd Meyer, David Rolnick, Graham W Taylor, Devis Tuia, et al. Harnessing artificial intelligence to fill global shortfalls in biodiversity knowledge. *Nature Reviews Biodiversity*, pages 1–17, 2025.

Jiawei Ren, Cunjun Yu, Xiao Ma, Haiyu Zhao, Shuai Yi, et al. Balanced meta-softmax for long-tailed visual recognition. *Advances in neural information processing systems*, 33:4175–4186, 2020.

Salim Rezvani and Xizhao Wang. A broad review on class imbalance learning techniques. *Applied Soft Computing*, 143:110415, 2023.

Shiori Sagawa, Pang Wei Koh, Tony Lee, Irena Gao, Sang Michael Xie, Kendrick Shen, Ananya Kumar, Weihua Hu, Michihiro Yasunaga, Henrik Marklund, et al. Extending the wilds benchmark for unsupervised adaptation. *arXiv preprint arXiv:2112.05090*, 2021.

Julian D Santamaria, Claudia Isaza, and Jhony H Giraldo. Catalog: A camera trap language-guided contrastive learning model. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 1197–1206. IEEE, 2025.

Dongsub Shim, Zheda Mai, Jihwan Jeong, Scott Sanner, Hyunwoo Kim, and Jongseong Jang. Online class-incremental continual learning with adversarial shapley value. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 9630–9638, 2021.

Peeyush Singhal, Rahee Walambe, Sheela Ramanna, and Ketan Kotecha. Domain adaptation: challenges, methods, datasets, and applications. *IEEE access*, 11:6973–7020, 2023.

Franck Trolliet, Cédric Vermeulen, Marie-Claude Huynen, and Alain Hambuckers. Use of camera traps for wildlife studies: a review. *Biotechnologie, Agronomie, Société et Environnement*, 18(3), 2014.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025.

Cheng-Hao Tu, Hong-You Chen, Zheda Mai, Jike Zhong, Vardaan Pahuja, Tanya Berger-Wolf, Song Gao, Charles Stewart, Yu Su, and Wei-Lun Harry Chao. Holistic transfer: towards non-disruptive fine-tuning with partial target data. *Advances in Neural Information Processing Systems*, 36:29149–29173, 2023.

Devis Tuia, Benjamin Kellenberger, Sara Beery, Blair R Costelloe, Silvia Zuffi, Benjamin Risse, Alexander Mathis, Mackenzie W Mathis, Frank Van Langevelde, Tilo Burghardt, et al. Perspectives in machine learning for wildlife conservation. *Nature communications*, 13(1):792, 2022.

Delia Velasco-Montero, Jorge Fernández-Berni, Ricardo Carmona-Galán, Ariadna Sanglas, and Francisco Palomares. Reliable and efficient integration of ai into camera traps for smart wildlife monitoring based on continual learning. *Ecological Informatics*, 83:102815, 2024.

Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, et al. Robust fine-tuning of zero-shot models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7959–7971, 2022.

Jingkang Yang, Kaiyang Zhou, Yixuan Li, and Ziwei Liu. Generalized out-of-distribution detection: A survey. *International Journal of Computer Vision*, 132(12):5635–5662, 2024.

Han-Jia Ye, Hong-You Chen, De-Chuan Zhan, and Wei-Lun Chao. Identifying and compensating for feature deviation in imbalanced deep learning. *arXiv preprint arXiv:2001.01385*, 2020.

Han-Jia Ye, De-Chuan Zhan, and Wei-Lun Chao. Procrustean training for imbalanced deep learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 92–102, 2021.

Xiaoyuan Yu, Jiangping Wang, Roland Kays, Patrick A Jansen, Tianjiang Wang, and Thomas Huang. Automated identification of animal species in camera trap images. *EURASIP Journal on Image and Video Processing*, 2013(1):52, 2013.

Yifan Zhang, Bingyi Kang, Bryan Hooi, Shuicheng Yan, and Jiashi Feng. Deep long-tailed learning: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 45(9):10795–10816, 2023.

Kaiyang Zhou, Ziwei Liu, Yu Qiao, Tao Xiang, and Chen Change Loy. Domain generalization: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 45(4):4396–4415, 2022.

Haowei Zhu, Ye Tian, and Junguo Zhang. Class incremental learning for wildlife biodiversity monitoring in camera trap images. *Ecological Informatics*, 71:101760, 2022.

The supplementary is organized as follows.

- Appendix A: Related Work
- Appendix B: Benchmark Details
- Appendix C: Baseline Methods
- Appendix D: An Algorithm Developer's Perspective
- Appendix E: An End User's Perspective
- Appendix F: Additional Analysis
- Appendix G: Statistical Measure Definition

A Related Work

Camera trap data in computer vision. Camera traps have become essential tools for biodiversity monitoring, capturing vast volumes of wildlife images that provide insights into species richness and behavior (Trolliet et al., 2014; Boitani, 2016). To automate analysis, deep learning methods have been widely adopted for species detection and classification (Norouzzadeh et al., 2018; Yu et al., 2013). A major challenge is generalization: models trained on one location often perform poorly when deployed elsewhere. The iWildCam challenges (Beery et al., 2019, 2021) address this by splitting data by camera location to assess out-of-distribution generalization (Koh et al., 2021; Mai et al., 2024). Test-time training has recently been reproduced on iWildCam for lightweight adaptation without retraining. More recently, multimodal foundation models have been applied to camera trap data for richer contextual understanding (Gabeff et al., 2024; Fabian et al., 2023; Santamaria et al., 2025). However, the temporal dynamics—e.g., seasonal shifts and animal migration—remain underexplored (Tu et al., 2023). Our benchmark incorporates temporal variability to better reflect real-world deployments.

Online continual learning. In contrast to conventional domain adaptation and continual learning (Farahani et al., 2021; De Lange et al., 2021), online continual learning assumes new data arrive sequentially in small batches (Mai et al., 2022). Models must adapt to evolving streams exhibiting non-stationarity—new classes or background changes (Aljundi et al., 2019; Mai et al., 2021; Shim et al., 2021). Existing benchmarks often lack timestamps and realistic shifts, leading to overly simple or overly complex scenarios (Mai et al., 2022). Zhu et al. studied class-incremental learning for wildlife monitoring but did not account for the real temporal order in camera-trap data (Zhu et al., 2022). Velasco-Montero et al. demonstrated how continual learning could be embedded into smart camera traps for efficient deployment, but their focus was primarily on hardware and system design (Velasco-Montero et al., 2024). In contrast, our benchmark aims to transform the valuable but less accessible and underexplored camera-trap data into a practical and standardized benchmark. By incorporating real temporal order and distribution shifts, we enable more realistic and meaningful evaluation in the camera-trap domain.

Class-imbalanced learning. Camera trap datasets often follow a long-tailed distribution, where models perform well on majority species but poorly on rare ones (Bevan et al., 2024; Malik et al., 2021). Since rare species are often of greatest ecological interest, addressing this imbalance is critical. Solutions fall into data-level and algorithm-level methods (Zhang et al., 2023; Rezvani and Wang, 2023). We focus on simple but effective approaches, such as the Balanced Softmax loss (Ren et al., 2020) in our study.

B Benchmark Details

Selection Criteria and Data Sources. The LILA BC repository currently hosts more than fifty datasets and continues to grow over time. For our benchmark we limit the selection to camera trap datasets, excluding other sources such as sea-animal imagery, drone imagery, and geological or earth-observation images. We select 17 datasets containing at least 5 species per camera and spanning at least 6 months. Detailed selection criteria are implemented as a separate preprocessing step due to the large volume of images and long processing time.

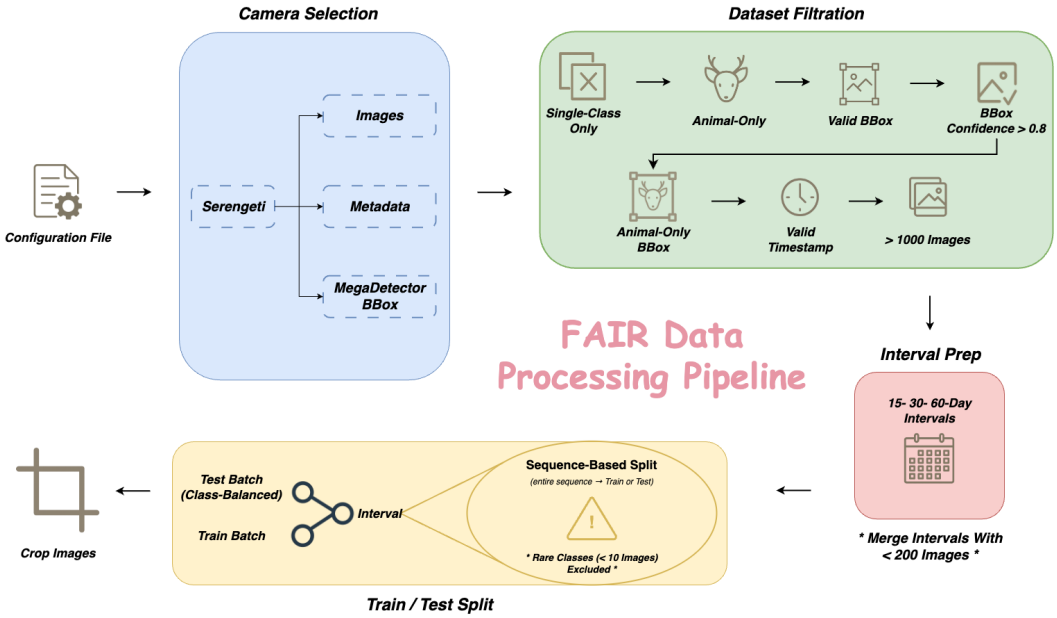


Figure 8: Detailed benchmark's data-processing pipeline.

Metadata Preprocessing. All LILA BC camera trap datasets use the COCO–camera trap format[1] which includes two optional fields—datetime and sequence—essential for our benchmark to group images into temporal intervals and to identify and handle burst images. When datetime is missing, we extract it from each image's EXIF header[2] and insert it into the metadata if available; when sequence is missing, we generate pseudo-sequences by grouping neighboring images within 3 seconds of one another; and when both fields are absent, we first recover datetime from EXIF and then apply the sequence-grouping step. After applying these criteria we retain 19 datasets, and a final filter requiring at least 5 species per camera removes 2 more, resulting in the 17 datasets used in our analyses. With the prepared metadata from all datasets, they are then passed into our benchmark data-processing pipeline to produce the final benchmark output.

Detailed Data Processing Pipeline. We assume each camera trap dataset consists of a collection of images along with associated metadata, including timestamps, image-level species labels, and bounding boxes for animal instances.[3] The finalized metadata is then fed into our data-processing pipeline, whose detailed flow is illustrated in Figure 8. Users can flexibly control the pipeline—selecting specific datasets or cameras, setting minimum image counts, bounding-box confidence thresholds, and excluding certain classes to focus on targeted taxa. Hard-case images (e.g., very low quality) may be filtered based on the pipeline configuration. In interval preparation, the default temporal interval is 30 days, but this can be adjusted (e.g., 15, 60, or custom days). Although the current benchmark does not use a fixed seed, users may specify one to ensure consistent splits. Rare-species interval data are stored alongside the training and test sets to enable separate evaluation of rare species.

Given such a dataset, we apply three filtering steps:

1. retain only images with a single species label[4] and without any humans or vehicles;
2. retain images with at least one detected animal bounding box (confidence > 0.8) and no humans or vehicles;
3. retain images that contain valid timestamps.

[1] A metadata standard maintained within the LILA BC repository, derived from the original COCO format.

[2] EXIF stands for Exchangeable Image File Format, a standard for storing metadata such as date and time inside image files.

[3] MegaDetector (Beery, 2023) detects animals without species classification.

[4] This ensures species labels can be correctly assigned to bounding boxes and removes $< 2\%$ of the images.

After filtering, we retain camera traps that still contain more than $1,000$ images and span a range of at least six months. For each retained camera trap, we divide the images into temporal chunks following the criteria above. Within each chunk, for every species category with more than 10 images, we randomly sample 10 images to form the test set and assign the remainder to the training set. To represent each image, we crop a single animal instance using the bounding box with the highest confidence score from MegaDetector. Before cropping, we enlarge the selected bounding box by 50% on each side to preserve context. A detailed illustration of the full pipeline is provided in Figure 8.

Statistical Measures. Our data processing pipeline yields 546 valid camera traps. Figure 3a presents summary statistics—each camera trap contains between 5 and 20 species, $1,000$ to $80,000$ images, and spans 6 to 70 months. We further derive metrics to quantify changes in species distribution between intervals and class imbalance both within and across intervals. (Please see the supplementary for definitions.) Figure 3b presents these statistics, revealing that many datasets exhibit substantial temporal shifts and class imbalance.

Handling Burst Images. Camera trap devices typically take bursts of images whenever motion is detected by their onboard sensors, with the number of frames per burst ranging from 3 to over 10. With preprocessed metadata, each dataset provides sequence IDs, total frame counts, and frame order. Using this information, we group all frames from the same burst into a single split (train or test) and never divide a burst across splits, preventing temporal leakage and ensuring that highly correlated frames do not appear simultaneously in training and testing. While we use this grouping for split assignment, we simplify the setting by treating each image within a burst as an independent sample for model training and evaluation. Downsampling bursts to a single frame or aggregating predictions across entire bursts are natural extensions we plan to explore to better align the benchmark with real-world deployment.

MegaDetector Bounding Box. Labeling bounding boxes requires substantial human effort and time, and not all datasets contain human-labeled bounding box data. We evaluate MegaDetector on 4 LILA BC datasets with human-labeled boxes and obtain 94.3% detection accuracy at IoU ≥ 0.5. Because we enlarge and crop each detected region by 50% of the bounding box size and apply a high-confidence threshold, we find MegaDetector to be robust for our purposes. Among the 17 datasets used in this study, 5 lack human-labeled boxes; the remaining 12 now include human-labeled boxes (this annotation is released after our benchmark preparation). For consistency, we use MegaDetector on all datasets: for the 5 without human boxes we rely entirely on MegaDetector detections, and for the 12 with human boxes we still apply MegaDetector with a high-confidence threshold to maintain uniform preprocessing. For each frame, MegaDetector provides bounding boxes for animals, humans, and vehicles, which we use to filter non-animal instances and, using metadata, to further link each captured bounding box with its species. In this work, we do not attempt to solve detection itself; instead, we use MegaDetector to build a consistent and reliable benchmark for studying classification. While detection combined with classification and even applying a classifier directly to the full image to capture broader context are directions for future research, our benchmark is designed to keep the task general and focus specifically on classification.

Online Continual Learning Task. Given time-stamped images from a camera trap, $\{(\boldsymbol{x}_i, y_i, t_i)\}$—where $\boldsymbol{x}_i$ denotes the i-th image, y_i its label, and t_i its timestamp—CAN partitions the data into a sequence of chunks, with each chunk D_j containing images captured within the j-th chronological time interval. During evaluation, we assess model performance sequentially, following the chronological order of time intervals. At each interval j, the model is evaluated on D_j. Once the evaluation is complete, D_j becomes available as training data and can be used to update the model before evaluating on the next interval, D_{j+1}.

C Baseline Methods

Zero-shot model. We use the BioCLIP 2 (Gu et al., 2025), a vision foundation model trained on biological data and capable of classifying over 800K species. For each species, we use its common name as the textual representation and apply OpenAI's prompt templates to generate the corresponding text embeddings.

Let $f_{\boldsymbol{\theta}}$ denote the pre-trained vision encoder and $\boldsymbol{w}_c$ the L2-normalized text embedding of class c. Given an image patch $\boldsymbol{x}$, the predicted class is $\hat{y} = \arg\max_c \boldsymbol{w}_c^\top f_{\boldsymbol{\theta}}(\boldsymbol{x})$.

Oracle model. We construct a reference classifier with access to training data from all time intervals simultaneously. That is, the model can preview the entire camera trap time span before training. This model serves as an upper-bound benchmark for adaptive methods over time.

We initialize the model with BioCLIP 2's vision encoder, followed by a linear classifier initialized using BioCLIP's text embeddings. We aggregate all training splits across intervals and perform full fine-tuning using standard cross-entropy loss, a cosine learning rate scheduler, and a fixed initial learning rate of 0.000025. A small validation set is held out for early stopping by sampling two frames from each class in the training data.

Accumulated model. We consider an adaptive model with access to labeled training data from all past intervals—*i.e.*, a continual learning setup with unlimited memory. This setting enables us to evaluate the difficulty of adaptation caused by temporal drift, without the added complexity of implementing specific continual learning strategies. Note that test data always comes from future intervals.

We follow the oracle model's setup for implementation, but use the test set from the immediately preceding interval for early stopping instead of a separate validation set.

D An Algorithm Developer's Perspective

Building on the previous analysis, we take a deeper look at how to apply machine learning algorithms for model adaptation. Specifically, we select 15 representative camera traps from the 546 in CAN. Among them, 5 exhibit cases where the oracle model performs significantly worse than the zero-shot BioCLIP 2 baseline, while the other 10 show strong gains from oracle fine-tuning. We begin with the first group to explore techniques that can improve performance when naive fine-tuning fails. We then examine both groups under the accumulated setting to assess the practical challenges of continual model adaptation over time.

D.1 Improving the Oracle Model

Class-imbalanced learning. We begin by investigating loss functions tailored for class-imbalanced scenarios, since standard cross-entropy tends to bias toward majority classes (Cui et al., 2019; Ye et al., 2020, 2021). Among the alternatives we experimented with, we found Balanced Softmax (BSM) loss (Ren et al., 2020) to be a simple yet effective choice.:

$$-\log\left(\frac{n_y e^{\eta_y(\boldsymbol{x})}}{\sum_c n_c e^{\eta_c(\boldsymbol{x})}}\right), \tag{1}$$

where $\eta_c(\boldsymbol{x}) = \boldsymbol{w}_c^\top f_{\boldsymbol{\theta}}(\boldsymbol{x})$ denotes the logit for class c, and n_c is the number of training instances for class c.

Parameter-efficient fine-tuning (PEFT). We explore parameter-efficient fine-tuning (PEFT) methods for model adaptation. Unlike full fine-tuning, which updates all parameters, PEFT modifies only a small subset of weights to mitigate overfitting and better preserve pre-trained knowledge (Mai et al., 2025). Among several methods we explored, we selected LoRA (Hu et al., 2022), which introduces trainable low-rank matrices into the attention projection layers, allowing efficient adaptation with minimal parameter updates.

Weight interpolation. We explore weight-space ensembles (WiSE) (Wortsman et al., 2022), a technique for enhanced model robustness. WiSE linearly interpolates the weights of the FT model with those of the pre-trained backbone. Formally, let $\boldsymbol{\theta}$ denote the pre-trained weights and $\boldsymbol{\theta}'$ the FT weights; the interpolated model weights are given by $\boldsymbol{\theta}(\alpha) = \alpha\boldsymbol{\theta} + (1-\alpha)\boldsymbol{\theta}'$ where $\alpha \in [0, 1]$ controls the interpolation ratio. This approach helps retain generalizable pre-trained features while benefiting from task-specific adaptation.

D.2 Transitioning to Accumulated Models

We now turn to our third baseline: accumulated models, which represent continually adaptive systems with access to unlimited memory. We begin by applying the techniques from subsection D.1. These adaptations show consistent improvements across different camera traps. Gains are seen not only in

cases where the oracle underperforms the zero-shot baseline, but also where it already performs well. We hypothesize that during continual adaptation, class imbalance and data scarcity are especially pronounced in early intervals, making these techniques particularly effective.

Despite these improvements, accumulated models often struggle in early intervals (Figure 1b). We suspect this is due to severe class imbalance, limited data availability, and the absence of certain categories during the early stages of adaptation. Addressing performance degradation in these intervals remains an important direction for future work.

Post-hoc calibration. When certain classes are absent during fine-tuning, the resulting model often assigns very low logits to those classes—even if the pre-trained model was originally capable of recognizing them (Tu et al., 2023; Mai et al., 2024). To address this issue, we adopt a post-hoc calibration strategy that adds a calibration factor γ to the logits of absent classes (Mai et al., 2024). The updated prediction rule becomes

$$\hat{y} = \arg\max_c \boldsymbol{w}_c^\top f_{\boldsymbol{\theta}}(\boldsymbol{x}) + \gamma \cdot \mathbf{1}[c \in \text{absent classes}]. \tag{2}$$

E An End-User's Perspective

With the proper use of machine learning techniques, we have seen that accumulated models can outperform zero-shot models after just a few intervals. In this section, we shift our perspective to address questions that a camera trap end-user might naturally ask. Specifically: When is the zero-shot model sufficient? And when is it necessary to further adapt the model?

E.1 When is the zero-shot model sufficient?

As with any machine learning model, a pre-trained model is expected to perform well when the test distribution closely aligns with its training data. However, when test data significantly deviates—becoming out-of-distribution (OOD)—the model's predictions can become unreliable. In such cases, relying on zero-shot predictions may be insufficient.

To investigate this, we explore the use of an OOD detection mechanism to estimate whether the zero-shot model is likely to be effective (Yang et al., 2024). Specifically, we adopt the Maximum Softmax Probability (MSP) (Hendrycks and Gimpel, 2016),[5] which assigns each test sample a non-OOD score defined as

$$s_{\text{MSP}}(\boldsymbol{x}) = \max_c \frac{e^{\eta_c(\boldsymbol{x})/\tau}}{\sum_{c'} e^{\eta_{c'}(\boldsymbol{x})/\tau}}, \tag{3}$$

where $\eta_c(\boldsymbol{x}) = \boldsymbol{w}_c^\top f_{\boldsymbol{\theta}}(\boldsymbol{x})$ denotes the logit for class c, and τ is a temperature parameter.

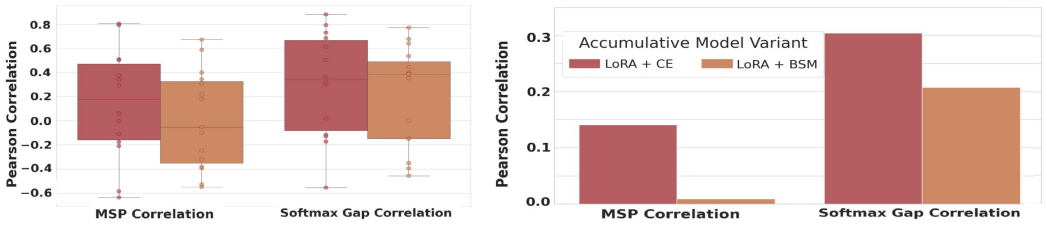
We compute the average non-OOD score for each camera trap and plot it against the corresponding zero-shot accuracy in Figure 7a. As shown, there is a clear positive correlation between the non-OOD score and zero-shot accuracy (with the red line indicating the regression fit). While some exceptions exist—where high MSP scores do not translate to high accuracy and vice versa—the overall trend highlights a promising research direction: developing reliable pre-deployment indicators of zero-shot performance.

E.2 Do We Need to Continually Adapt?

For accumulated models, a natural question is whether adaptation is needed after every interval, or if it can stop after a certain point. That is, once the model has been updated for a few intervals, might it generalize well enough to handle future data without further tuning?

To explore this, we consider a simple strategy where adaptation is halted after a fixed number of intervals—*i.e.*, the accumulated model is frozen for the remaining rounds. Table 1 shows the results. We observe that after several intervals of adaptation, the model has acquired domain-specific knowledge and can outperform the zero-shot baseline even when frozen. However, its performance gradually lags behind the continually adaptive model, highlighting the continued benefits of ongoing adaptation.

[5]Despite its simplicity, MSP has proven effective various settings(Averly and Chao, 2023), including applications with pre-trained CLIP-style models (Ming et al., 2022).



(a) (b)

Figure 9: Non-OOD scores (MSP or SG) correlate with accumulated model accuracy. **(a)** We compute the Pearson correlation within each camera trap and show the statistics across 16 traps. **(b)** The average Pearson correlation score over 16 traps. In either plot, the softmax gap (SG) score shows better correlations.

E.3 When Should We Adapt?

Given the results above, a key practical question arises: can we estimate when model adaptation is necessary? In other words, is there an effective way to decide—based on data or model signals—whether adapting at a given interval will yield meaningful performance gains?

To investigate this, we again apply OOD detection methods, this time using the accumulated models. In addition to the MSP score, we introduce a new metric called the Softmax Gap (SG), which measures the probability gap between the most likely ($c^\star$) and second most likely ($c^\dagger$) classes:

$$s_{\text{SG}}(\boldsymbol{x}) = \frac{e^{\eta_{c^\star}(\boldsymbol{x})/\tau} - e^{\eta_{c^\dagger}(\boldsymbol{x})/\tau}}{\sum_{c'} e^{\eta_{c'}(\boldsymbol{x})/\tau}}. \tag{4}$$

This score captures how confidently the model prefers its top prediction over the next best. Since FT models tend to produce high MSP scores, we hypothesize that the SG score may more sensitively reflect whether adaptation is needed.

As shown in Figure 7b (New Zealand EFD DCAME01 camera trap), both MSP and SG scores generally track with the accumulated model's accuracy, with SG often showing a better alignment. In Figure 9, we further report the Pearson correlation between accuracy and each score across datasets (one point per dataset). Results show generally positive correlations—especially for LoRA-adapted models—suggesting that non-OOD scores could serve as a useful signal for deciding when to update. Of course, further improvement and calibration of such indicators remain an important direction.

F Additional Analysis

Dataset	Camera	Model	ZS	Accum	Oracle
MAD	A05	BioCLIP 2	0.78	0.89	0.91
		SigLIP 2	0.66	0.72	0.72
KGA	KHOLA03	BioCLIP 2	0.63	0.75	0.90
		SigLIP 2	0.40	0.46	0.51

Table 2: BioCLIP 2 and SigLIP 2 comparison with BSM lss and LoRA

Foundation Model Selection. BioCLIP 2 is a recently proposed state-of-the-art model that classifies animal species with high accuracy. To further address model selection and demonstrate generality beyond BioCLIP 2, we run the same training process with SigLIP 2 (Tschannen et al., 2025) and observe that both classifiers combined with MegaDetector exhibit similar trends of temporal adaptation improvements Table 2. For comparison, we run BioCLIP 2 and SigLIP 2 using LoRA with the BSM loss.

Rare species classification. We further evaluated our trained model by performing inference on this rare-species data. To do so, we expanded our linear classifier head to include rare species not used during training, and the results are reported in Table 3. In our analysis, both the accumulated and oracle settings perform comparably well. One plausible explanation is that our benchmark processes rare species on a per-interval basis; thus, a species may be rare in an early interval but become

common in a later interval. Under this scenario, the model may learn the species during a dominant interval and subsequently infer it when it appears as rare.

Dataset	Camera	Classes	ZS	Accum	Oracle
MAD	A05	12	0.78	0.89	0.91
		10 (+ 22)	0.57	0.48	0.54
KGA	KHOLA03	7	0.63	0.75	0.90
		7 (+ 10)	0.67	0.79	0.78

Table 3: Rare classification with BSM loss and LoRA (+ values denote newly added from rare).

Dataset	Camera	Classes	ZS	Accum	Oracle
MAD	A05	12	0.78	0.89	0.91
		1096	0.38	0.78	0.90
KGA	KHOLA03	7	0.63	0.75	0.90
		1096	0.46	0.65	0.92

Table 4: Open-set classification with BSM loss and LoRA.

Open-set classification. Our benchmark can easily extend to the open-set scheme—BioCLIP 2's zero-shot ability allows recognition beyond predefined labels. Specifically, rather than assuming a fixed, closed set of species per camera trap, we include all species' common names, total 1096 classes, in our benchmark and initialize our linear classifier with the models' text embeddings, enabling zero-shot classification of previously unseen taxa. The results in Table 4 show that this approach allows open-set classification, with the oracle remaining comparable and the accumulated results approaching the original performance.

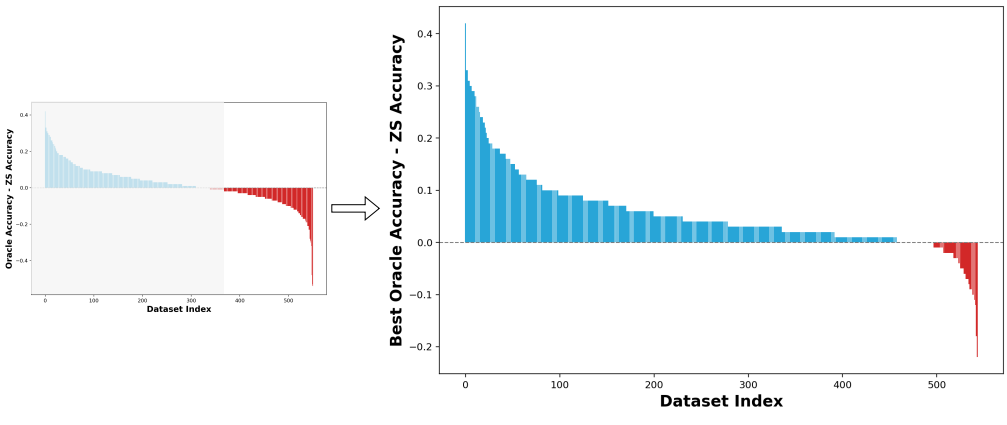


Figure 10: Overall oracle model improvement by the BSM loss and LoRA.

Improving oracle. For the 227 cameras where the base oracle (full fine-tuning with cross-entropy loss) performed worse than zero-shot in Figure 3c, we retrained the same camera traps under a setting using LoRA and the BSM loss. As shown in Figure 10, the majority of oracles that previously underperformed zero-shot now surpass the zero-shot baseline.

G Statistical Measure Definition

To quantify the diverse characteristics of each camera trap data, we provide basic statistics and measures that can quantify changes in species distribution between intervals and the degree of class imbalance.

We provide the following basic statistics for each valid camera trap in our benchmark, including: number of **species** (class), number of **images**, number of **months**. We further derive measures to quantify unique characteristics in each camera trap.

Degree of class imbalance. The Gini Index (G), used to quantify the degree of class imbalance in a camera trap dataset, is defined as:

$$G = 1 - \sum_{i=1}^{C} p_i^2, \quad \text{where} \quad p_i = \frac{n_i}{N}.$$

Here, C denotes the number of classes, n_i represents the number of samples in class i, and N is the total number of samples in the dataset. A higher value of G approaching 1 indicates a more balanced class distribution, whereas a lower value approaching 0 signifies increased class imbalance.

Degree of temporal shift. To measure the degree of class distribution shift between consecutive intervals, we define the Temporal Shift (TS) metric as follows:

1. Given a set of n intervals, for each pair of consecutive intervals $(i, i+1)$, compute the normalized class distributions:

$$p_j^{(i)} = \frac{n_j^{(i)}}{\sum_k n_k^{(i)}}, \quad q_j^{(i+1)} = \frac{n_j^{(i+1)}}{\sum_k n_k^{(i+1)}}.$$

2. Compute the pairwise L_1 shift over all classes:

$$\mathrm{TS}_{i,i+1} = \sum_j \left| p_j^{(i)} - q_j^{(i+1)} \right|.$$

3. Finally, average across all consecutive interval pairs:

$$\mathrm{TS} = \frac{1}{n-1} \sum_{i=1}^{n-1} \mathrm{TS}_{i,i+1}.$$

This metric quantifies the overall change in class distribution between consecutive intervals.