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

Sooyoung Jeon, Zheda Mai, Hongjie Tian, Vidhi Bakshi,
Lemeng Wang, Jiacheng Hou, Ping Zhang, Arpita Chowdhury, Wei-Lun Chao
The Ohio State University

Abstract

Camera traps enable scalable wildlife monitoring but large variations across sites, seasons, and sensors undermine long-term reliability. We introduce **Continually Adapt or Not (CAN)**, a benchmark that evaluates models under real temporal evolution across 546 cameras from 17 LILA BC datasets by framing recognition as *online continual learning*: models update in chronological order. We build a FAIR-compliant pipeline and study three settings: (1) zero-shot BioCLIP 2, (2) an oracle trained on all data, and (3) an accumulated model trained sequentially. Continual adaptation generally helps, with diminishing returns as systems stabilize. Techniques such as Balanced Softmax, LoRA, and WiSE interpolation further improve robustness in long-tailed, low-data regimes. CAN offers concrete guidance on when and how to adapt ecological vision systems and provides a unified, reproducible testbed for sustainable real-world continual learning.

1 Introduction

Camera traps enable large-scale, non-invasive wildlife monitoring (Pollock et al., 2025; Tuia et al., 2022) but vary greatly across space, time, and hardware (Koh et al., 2021; Beery et al., 2021), challenging reliable automation.

Prior works frame this as domain adaptation or generalization (Sagawa et al., 2021; Zhou et al., 2022), yet practitioners ask: *Will it work at my site? How much data is needed? Must it keep updating?* These issues are compounded by slow data collection and limited species coverage (Tu et al., 2023).

We present **Continually Adapt or Not** (CAN)—a benchmark for **camera-trap classification over time**. Each camera's stream is split into sequential **temporal intervals**, forming an **online continual learning** setup (Mai et al., 2022): at interval j, the model trains on current data and is evaluated on $j+1$, mirroring real deployments.

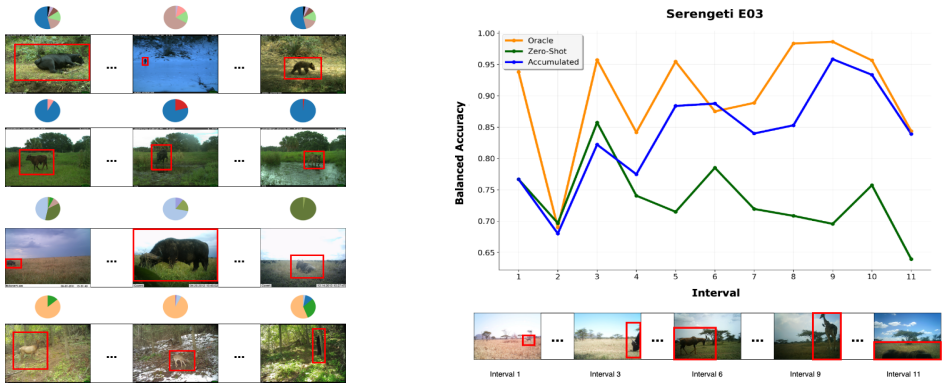
Our FAIR-compliant pipeline quantifies temporal shift and class imbalance. Three baselines—zero-shot BioCLIP 2, an all-data oracle, and an accumulated continual model—expose when fine-tuning helps or hurts. Figure 1 illustrates data variability and evaluation design.

2 "Continually Adapt or Not" Benchmark

2.1 Motivation

As shown in Figure 1a and Figure 4, camera-trap images vary widely in style and quality—some blurry, others close-up or poorly lit. Even within a single site, seasonal and temporal shifts alter both

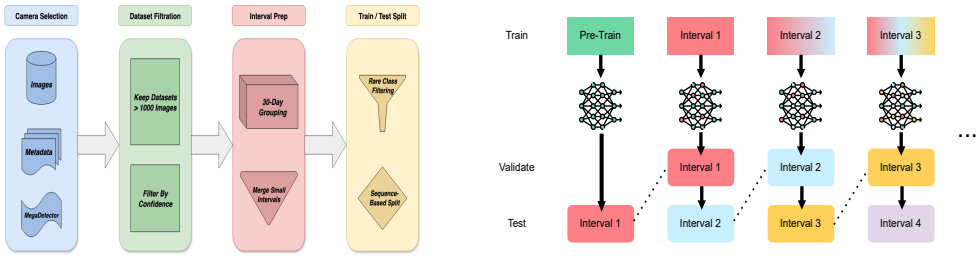
39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: The 3rd Workshop on Imageomics: Discovering Biological Knowledge from Images Using AI



(a) (b)

Figure 1: **(a)** Variability of camera-trap images across space (rows) and time (columns), with pie charts showing species distributions. **(b)** Three baselines: zero-shot, oracle, and accumulated. The accumulated model is trained on all intervals before j and evaluated on j.

background and species distribution, making it difficult to build robust classifiers. These variations raise practical questions for end-users: *Will the model generalize to my setting, or need further adaptation?* To investigate this, we introduce the **Continually Adapt or Not** (CAN) benchmark—a curated testbed to evaluate pre-trained models and foster adaptive algorithms for camera-trap analysis.



(a) (b)

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

2.2 Data Source and Processing Pipeline

CAN builds on the LILA BC repository (LILA BC), which aggregates many global camera-trap datasets (e.g., Ohio Small Animals, Snapshot Karoo) captured by hundreds of stationary cameras. We select 17 datasets meeting minimum duration and size requirements and process them using a standardized FAIR-compliant pipeline—Findable, Accessible, Interoperable, and Reusable. Figure 2a summarizes this process, while Figure 3a–Figure 3b show data coverage, temporal span, and imbalance statistics.

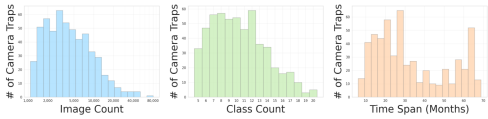
2.3 Online Continual Learning Task

Unlike conventional domain adaptation, where target data is seen all at once Gong et al. (2012); Singhal et al. (2023), CAN follows an **online continual learning** setup Mai et al. (2022) matching real deployments: data arrives sequentially, models are updated after each interval, and evaluated on the next (Figure 2b). This formulation captures temporal evolution and directly tests adaptation stability.

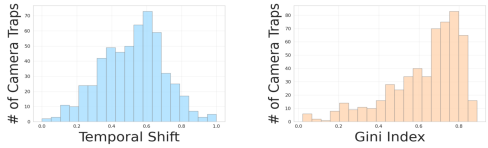
2.4 Baseline Methods

Setting. We assume a closed-set scenario where species per camera are known from historical data. Performance is measured as balanced accuracy—per-class accuracy averaged across all intervals.

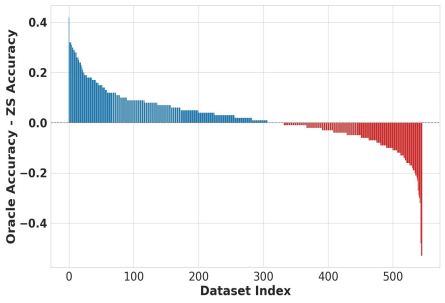
Baselines. We compare three representative strategies: (1) **Zero-shot Model**—BioCLIP 2 without training, using text-image matching. (2) **Oracle Model**—BioCLIP 2 with a linear head trained jointly



(a)



(b)



(c)

Figure 3: **(a)** Camera-trap statistics. **(b)** Temporal shift and class imbalance. **(c)** Oracle vs. ZS performance gap across datasets.

on all intervals. (3) **Accumulated Model**—same as the oracle but trained incrementally on past intervals. These configurations expose trade-offs between stability, adaptation, and overfitting.

2.5 Results and Analysis

Comparing zero-shot and oracle performance across 546 camera traps (Figure 3c) yields three insights. First, BioCLIP 2 already generalizes well—over 30% of cameras exceed 90% accuracy—showing strong zero-shot transfer. Second, another 30% remain below 80%, emphasizing persistent site-specific difficulty. Third, the oracle does not always outperform zero-shot, revealing that naïve fine-tuning can degrade generalizable features and highlighting the need for robust adaptation.

Zero-shot failure cases. Poor performance often coincides with low-quality images—blur, occlusion, poor lighting, or extreme close-ups (Figure 4). These artifacts obscure species features, suggesting that improved capture quality and pre-filtering could aid deployment.

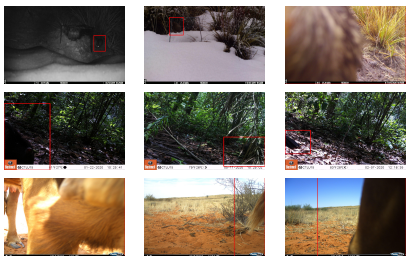


Figure 4: Very challenging cases for the zero-shot model.

Oracle underperformance. Despite full-data training, the oracle occasionally falls below zero-shot accuracy due to severe class imbalance and limited samples. Standard fine-tuning with cross-entropy can distort general representations, causing overfitting. This motivates stronger strategies such as the BSM loss and LoRA, which improve robustness under scarce or imbalanced data, as shown in Figure 5.

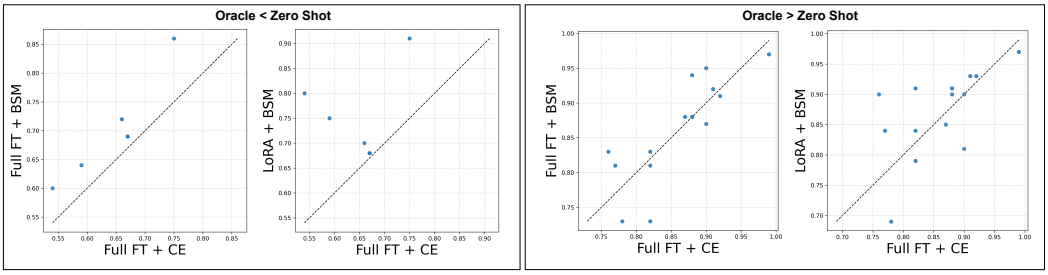
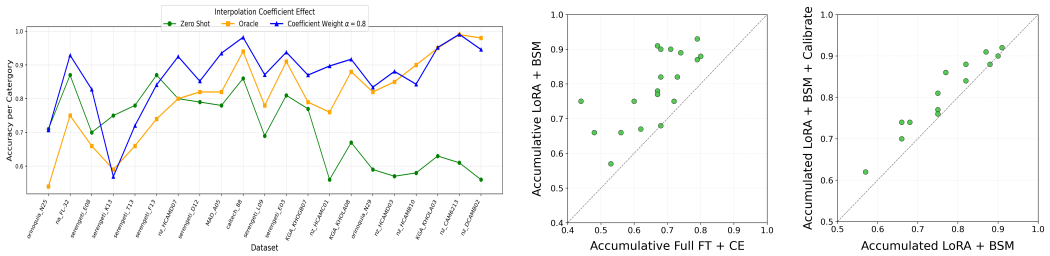


Figure 5: Oracle model improvement by the BSM loss and LoRA. BSM and LoRA consistently boost oracle performance when zero-shot outperforms fine-tuned models.

3 Going Deep into the CAN Benchmark

We analyze model adaptation in CAN from two perspectives: (1) the *algorithm developer*, focusing on improving oracle and accumulated models; and (2) the *end-user*, deciding when zero-shot suffices or continual updates are worthwhile.



(a) (b) (c)

Figure 6: **(a)** WiSE interpolation consistently improves over zero-shot and oracle. **(b)** BSM loss with LoRA surpasses CE fine-tuning for accumulated models. **(c)** Calibration yields further gains.

3.1 Algorithm Developer's Perspective

Oracle improvements. We study three techniques: (i) *Balanced Softmax (BSM)* (Ren et al., 2020), (ii) *LoRA* (Hu et al., 2022; Mai et al., 2025), and (iii) *WiSE interpolation* (Wortsman et al., 2022). BSM and LoRA help when naive fine-tuning fails, though BSM may harm strong models. WiSE interpolation (Figure 6a) reliably improves accuracy without retraining, offering a simple post-hoc boost.

Accumulated model improvements. Using BSM, LoRA, and WiSE mitigates imbalance and limited data. Post-hoc calibration further stabilizes rare-class accuracy, especially early on. Figure 6b–Figure 6c show consistent benefits during continual updates.

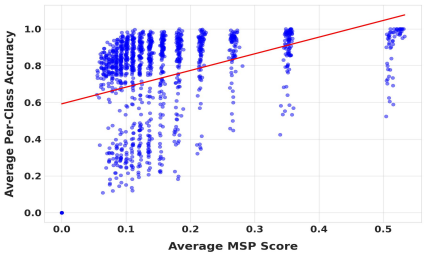


Figure 7: Non-OOD scores correlate with zero-shot accuracy.

Stage	Accuracy
33%	0.582
67%	0.686
100%	0.761

Table 1: Accuracy over accumulated training fractions (mean of 15 traps).

3.2 End-User's Perspective

When is zero-shot enough? Zero-shot performs well when test data matches pretraining domains but drops under shifts. Figure 7 shows non-OOD confidence metrics predict reliability before deployment.

Is continual adaptation needed? Table 1 shows accumulated models remain competitive for several intervals but later trail continuously adapted ones, confirming long-term benefits of regular updates.

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

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

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.

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, 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, 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.

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.

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.

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

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.

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.

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.