Lessons and Open Questions from a Unified Study of Camera-Trap Species Recognition Over Time

Sooyoung Jeon*[1], Hongjie Tian*[1], Lemeng Wang*[1], Zheda Mai†[1], Vidhi Bakshi[1], Jiacheng Hou[1], Ping Zhang[1], Arpita Chowdhury[1], Jianyang Gu[1], and Wei-Lun Chao[2]

[1]The Ohio State University [2]Boston University
https://jeonso0907.github.io/stream-trap/

Abstract. Camera traps are crucial for large-scale biodiversity monitoring, yet accurate automated analysis remains challenging due to diverse deployment environments. While the computer vision community has predominantly framed this challenge as cross-domain (*e.g.*, cross-site) generalization, this perspective overlooks a primary challenge faced by ecological practitioners: maintaining reliable recognition at the *fixed site over time*, where the dynamic nature of ecosystems introduces profound temporal shifts in both background and animal distributions. To bridge this gap, we present the first unified study of **camera-trap species recognition over time**. We introduce a realistic, large-scale benchmark comprising 546 camera traps with a streaming protocol that evaluates models over chronologically ordered intervals. Our end-user-centric study yields four key findings. (1) Biological foundation models (*e.g.*, BioCLIP 2) underperform at numerous sites even in initial intervals, underscoring the necessity of site-specific adaptation. (2) Adaptation is challenging under realistic evaluation: when models are updated using past data and evaluated on *future* intervals (mirrors real deployment lifecycles), naive adaptation can even degrade below zero-shot performance. (3) We identify two main drivers of this difficulty: severe class imbalance and pronounced temporal shift in both species distribution and backgrounds between consecutive intervals. (4) We find that effective integration of model-update and post-processing techniques can largely improve accuracy, though a gap from the upper bounds remains. Finally, we highlight critical open questions, such as predicting when zero-shot models will succeed at a new site and determining whether/when model updates are necessary. Together, our benchmark and analysis provide actionable deployment guidelines for ecological practitioners while establishing new directions for future research in vision and machine learning.

Keywords: Temporal Shift · Adaptation · Ecological Monitoring

1 Introduction

Camera traps are vital tools for ecology and conservation: they non-invasively capture large volumes of time-stamped images in natural habitats, supporting

*Equal contribution. †Corresponding author (mai.145@osu.edu).

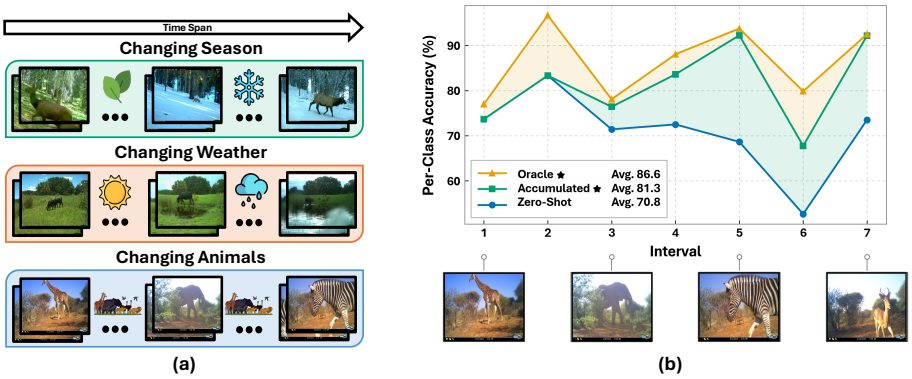

Fig. 1: (a). Temporal shifts: Even at a fixed site, foreground and background appearances evolve over time due to changing seasons, weather, and animal distribution. **(b). Streaming evaluation:** Accumulated⋆ model is fine-tuned with training data up to the current interval and evaluated on the test data of the next interval. For reference, an Oracle⋆ model is fine-tuned with the union of training data from all time intervals at a site and evaluated on the test data of all intervals. (⋆: our recipe Sec. 4.4).

biodiversity monitoring, behavior analysis, and conservation planning [41, 55]. Yet ecosystems are inherently non-stationary: seasonal shifts, weather changes, vegetation growth, and animal migration constantly reshape backgrounds and animal distributions (Fig. 1a). Such dynamic conditions introduce substantial visual variability, posing unique challenges for vision models [5, 26].

Prior studies often frame the problem as domain adaptation or generalization, aiming to transfer models trained on source domains (*e.g.*, specific cameras) to unseen target domains [45, 70]. While this perspective casts camera trap data as a crucial testbed for machine learning robustness, it diverges from the practical needs of end-users (*e.g.*, ecologists and conservation practitioners). Their dominant interest is not a model's ranking on cross-domain leaderboards, but its *ability to maintain high accuracy at a specific deployment site as dynamically changing data arrives*. Recognizing that this practical problem remains largely underexplored, we conduct a unified and comprehensive study of **camera trap species recognition over time**, reformulating the evaluation to faithfully mirror the temporal lifecycle of real-world deployments.

To facilitate this study, we introduce a realistic camera trap benchmark StreamTrap with a streaming evaluation protocol where models are evaluated as new data arrives and are incrementally updated thereafter. For each camera trap, the image stream is partitioned into consecutive *time intervals* for evaluating models chronologically. Concretely, at each interval j, a model can access cumulative, labeled training data from all intervals up to the j-th interval, undergo parameter updates, and evaluate on the unseen test data from the subsequent $(j+1)$-th interval. To construct StreamTrap, we curate data from the LILA BC repository [1], identifying **546** camera traps from 17 datasets, each

with sufficient data spanning at least **6** months. Also, we provide a modular, reproducible data pipeline that adheres to FAIR principles (Findable, Accessible, Interoperable, and Reusable), enabling practitioners to seamlessly convert their raw ecological data into standardized temporal benchmarks.

Building on StreamTrap, we conduct a unified *end-user-centric* study aiming to answer common questions faced by ecological practitioners, and distill actionable guidelines for maintaining accurate models in dynamic environments. We summarize our key analyses as follows:

Adaptation is required even with biological foundation models. Before deployment, end-users must decide whether a zero-shot biological vision foundation model is strong enough for a new camera trap, or whether site-specific adaptation is still needed. This decision is especially important as biological vision foundation models [19, 30, 49] already achieve state-of-the-art zero-shot accuracy across diverse biodiversity datasets, including subsets of the LILA BC repository. However, in StreamTrap, we observe substantial variability in BioCLIP 2's *zero-shot* accuracy across 546 camera traps: while 161 traps exceed 90% accuracy, demonstrating the promise of foundation models, another 162 fall in the 50–80% range, indicating that site-specific adaptation remains critical for many deployment sites.

Adaptation is challenging in the real world. When off-the-shelf foundation models underperform at a deployment site, fine-tuning is a natural first adaptation choice [31, 35, 69]. However, adaptation in StreamTrap is more challenging than standard in-domain fine-tuning: a model updated using data observed up to the current interval must be evaluated on the *future* interval. Under this realistic streaming protocol, we surprisingly find that naively fine-tuning on all data observed so far (**accumulated model**) can degrade performance, often below the zero-shot baseline. Our investigation identifies two primary drivers behind this adaptation failure. First, as camera traps acquire images passively and must "wait" for animals to appear, some species may not be observed within a given interval. Thus, the training data available for an interval is often extremely low-shot and severely imbalanced, causing poor generalization during naive adaptation [53]. Second, due to the inherent non-stationarity of ecosystems, both the background context and the animal population distributions undergo profound shifts between consecutive intervals. As a result, a model that achieves high accuracy on current and past intervals is not guaranteed to remain strong in the future intervals with shifted distributions.

Practical recipes for robust adaptation. To combat the low-shot and imbalance issues, we evaluate a suite of practical techniques and find that combining parameter-efficient fine-tuning [36] and class-imbalance losses [43, 61] provides a robust recipe for camera-trap model adaptation. However, even with this improved recipe, continually adapted models still exhibit a non-negligible performance gap (Fig. 1b) compared to an upper bound (**oracle*** model fine-tuned with *all* training data across all time intervals at a site). To further address the inter-interval distribution shifts and close this gap, we identify three post-

processing techniques: post-hoc logit calibration [31], weight interpolation [59], and interval model selection.

More practical considerations. Accurate real-world deployment involves more than choosing an adaptation algorithm. Through discussions with end-users, we identify practical questions frequently overlooked by the vision community: *How can practitioners predict if a foundation model's zero-shot performance will be sufficient at a new, unseen site? If not, must the model be updated at every interval, or is there an effective mechanism to determine when adapting at a given interval is actually necessary?* As these questions are largely under-explored, we provide preliminary analyses and baseline solutions.

Contributions. We present the first unified study of camera trap species recognition over time that reflects the temporal lifecycle of ecological monitoring. We systematically analyze critical deployment questions, providing actionable guidance on what works while urging the computer vision community to tackle the remaining open challenges. Our work serves two primary audiences:

- **End-users**: (1) We provide actionable guidance for leveraging foundation models and adaptation techniques in reliable camera-trap analysis; (2) Our FAIR-compliant data preparation pipeline helps practitioners convert raw ecological monitoring data into standardized, reusable benchmarks.
- **Algorithm developers**: (1) We transform the LILA BC repository, which has historically been difficult for developers to use at scale, into a standardized benchmark for evaluating robust real-world adaptation.(2) We identify critical but overlooked challenges about *what*, *how*, and *when* to update models under field constraints (*e.g.*, scarce or imbalanced data, non-stationary streams). By steering development toward these practitioner-driven questions, we aim to accelerate progress in this vital scientific application area.

2 Related Work

Camera trap data in computer vision. Camera traps are essential tools for biodiversity monitoring, producing large-scale wildlife image collections that reveal species richness and behavior [10,52]. To automate analysis, deep learning has been widely adopted for species detection and classification [39,63]. A major challenge is generalization: models trained at one location often fail when deployed at another. The iWildCam challenges [5,6] address this by splitting data by camera location to assess out-of-distribution generalization [26,31]. Recent work further explores multimodal foundation models for richer camera-trap understanding [13,17]. However, most prior work frames camera-trap analysis as a cross-domain generalization problem, whereas real deployments often require reliable species recognition at a **fixed site over time**, where seasonal changes, habitat dynamics, and animal migration continuously shift both species distributions and visual backgrounds, degrading model performance over long-term monitoring. We address this gap with a **unified, end-user-centric study** of camera-trap recognition across the temporal deployment lifecycle, characterizing adaptation challenges and deriving practical guidelines for long-term use.

Continual learning (CL). CL studies how a model learns from a non-stationary data stream while retaining acquired knowledge and improving over time without catastrophic forgetting [28,32,33]. This paradigm is closely related to our streaming evaluation protocol. However, standard CL assumes limited or no access to past data due to privacy or other constraints [9,34]. In contrast, ecological deployments treat labeled camera trap data as a valuable resource that is archived and fully accessible. To isolate temporal shift effects and avoid conflating them with artificial storage constraints, our baseline uses all historically available data at each interval, corresponding to the cumulative training upper bound in CL. While StreamTrap is explicitly tailored for studying wildlife monitoring under temporal shifts, its chronologically ordered data streams and severe, natural distribution shifts inherently provide a challenging real-world testbed for the broader continual learning community.

Class-imbalanced learning. Camera trap datasets often follow a long-tailed distribution, where models perform well on the majority species but poorly on rare ones [8,37]. As rare species are often of greatest ecological interest, addressing this imbalance is critical. Common techniques can be broadly categorized into two groups: data-level methods modifying the data distribution by over(under)sampling; algorithm-level methods adjusting the learning process with specialized losses or algorithmic modifications [44,65]. We provide more detailed related work in Sec. A.

3 StreamTrap: Benchmarking Real-World Camera Trap Deployment with Streaming Evaluation Protocol

Motivation. Camera traps are central to ecology and conservation: they non-invasively capture massive streams of time-stamped images in natural habitats, enabling biodiversity monitoring, behavior analysis, and conservation planning [41,55]. Yet the environments are inherently non-stationary: seasons, weather, vegetation, and migration continually reshape both the backgrounds and species distributions (Fig. 1a), creating persistent visual shifts for deployed vision models. Prior work typically frames this challenge as domain adaptation / generalization, focusing on transferring models from source to target domains [45,70]. However, this framing does not fully capture the needs of end-users. They do not prioritize static cross-domain leaderboards; instead, their dominant interest is whether a model can *maintain reliable accuracy at a fixed deployment site as new data arrive over time.*

Despite this practical need, existing benchmarks don't capture how camera-trap data arrive in the wild or reflect the temporal lifecycle of real deployments. To close this gap and refocus the community on deployment-critical evaluation, we introduce StreamTrap, a benchmark for camera trap species recognition over time with a streaming evaluation. By shifting the research focus from static domain transfer to long-term reliability, StreamTrap encourages the community to develop adaptive systems that remain accurate throughout real ecological monitoring deployments.

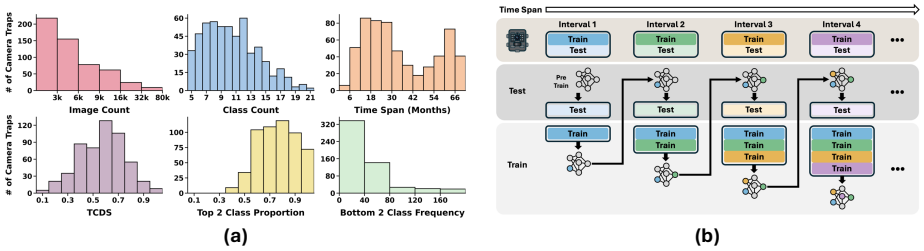

Fig. 2: (a). StreamTrap statistics. Top row: histograms of the image, class, and time span; Bottom row: histograms of temporal shift (TCDS, defined in Sec. 4.3) and class imbalance metrics (defined in Sec. 3.1). **(b). Streaming evaluation protocol:** models are continually fine-tuned on the labeled data observed so far and evaluated on future unseen data.

Streaming evaluation protocol. Unlike conventional domain adaptation, where target-domain data are often assumed to be available all at once [18,48], StreamTrap evaluates models under a realistic streaming protocol that better reflects camera-trap deployment: models are evaluated on future data as they arrive and may be updated after labeled data from previous intervals become available. 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, we partition the stream into a sequence of chronological intervals. At interval j, a model may access the accumulated labeled training data from intervals up to interval j, update its parameters, and is then evaluated on unseen test data from the subsequent (j+1)-th interval. Fig. 2b illustrates this protocol.

Also, while StreamTrap is tailored for camera traps under temporal shift, its streaming formulation also serves as an excellent, highly challenging testbed for broader machine learning paradigms, such as online continual learning [33].

3.1 Benchmark Construction and Statistics

Data source. We build StreamTrap on the LILA BC repository [1] with 17 camera-trap datasets (*e.g.*, Snapshot Serengeti) comprising hundreds of camera traps deployed across continents. Here, a camera trap refers to a stationary camera installed at a fixed location to passively monitor wildlife activity.

FAIR processing pipeline. We focus on camera traps with sufficient temporal coverage and image volume to support streaming evaluation. We first partition each camera trap stream into 30-day intervals; intervals with fewer than 200 images are merged with adjacent intervals to ensure sufficient data for both training and testing. Within each interval, we split data into training and test sets, constructing the test split to be **class-balanced**. Species with fewer than 10 samples in an interval are held out as rare species to maintain evaluation stability: they are separately evaluated and analyzed in Sec. C.1. To focus on the temporal shift problem, we follow the LILA BC protocol and use MegaDe-

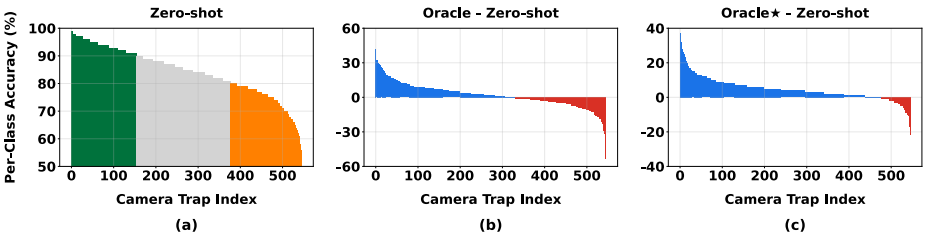

Fig. 3: (a). Zero-shot accuracy across camera traps. **Green**: $> 90\%$ accuracy; **orange**: $< 80\%$. **(b).** Accuracy difference between oracle naive fine-tuning and zero-shot. **Blue**: oracle outperforms zero-shot; **red**: zero-shot is better. **(c).** Accuracy difference between oracle fine-tuning with our adaptation recipe (Oracle⋆) and zero-shot.

tector [4] to extract image patches around detected animal instances. We enlarge each bounding box by 50% to preserve background context.

To promote the FAIR principles (Findable, Accessible, Interoperable, and Reusable) and enable future researchers to convert their raw ecological data into standardized benchmarks, we establish a modular data processing pipeline. Concretely, we preprocess each camera trap to ensure the availability of essential metadata (timestamps, image-level species labels, and animal bounding boxes) and apply two main filtering steps:

1. **Single-species filtering:** retain images containing a single species (no humans or vehicles) so that image-level labels can be unambiguously assigned to detected instances (removes $< 2\%$ of images);
2. **Detection filtering:** retain images with at least one detected animal bounding box with confidence > 0.8 to ensure reliable localization of animals.

After filtering, we retain camera traps with more than 1,000 images spanning at least **6** months. Additional details, including design choices, preprocessing and pipeline steps, MegaDetector robustness checks, and the list of camera traps included, are provided in Sec. D.

Basic statistics. Our rigorous processing pipeline yielded 546 valid camera traps. A summary of statistics (*e.g.*, image and species histogram) is provided in Fig. 2a. To quantify class imbalance for a camera trap, we report two complementary metrics: (i) the fraction of images belonging to the two most frequent classes, and (ii) the number of images in the least frequent classes.

4 A Unified End-User-Centric Study

Building on StreamTrap, we conduct the first unified *end-user-centric* study to answer common questions faced by practitioners during deployment, and distill actionable guidelines for maintaining accurate models in dynamic environments.

Setting & Model. For each camera trap, the model is initialized with the recently released BioCLIP 2 [19], a biological vision foundation model trained on multimodal biological data and capable of classifying over 800K species. To isolate the temporal dynamics we aim to study, we assume the candidate set of

species expected at each camera trap is known a priori. This assumption aligns with real-world deployments, where practitioners typically have prior knowledge of the local fauna from historical observations or metadata. To ensure the broader generalizability of our findings, we provide additional experiments in Sec. C.1 for scenarios where species are unknown beforehand.

Evaluation Metric. For each camera trap, we compute a model's average per-class accuracy within each interval (Balanced Accuracy), and then average these values across all intervals to obtain the overall performance.

4.1 Is Adaptation Still Required with Foundation Models?

Before deployment, practitioners must decide whether a zero-shot biological foundation model is reliable enough for a new camera trap, or whether site-specific adaptation is still necessary. This question is especially important as recent biological foundation models [19, 49] already achieve state-of-the-art zero-shot accuracy across biodiversity datasets, including LILA BC subsets.

Zero-shot baseline. We begin by evaluating the zero-shot performance of BioCLIP 2 for all 546 camera traps in StreamTrap. To construct the zero-shot classifier head, we utilize each species' common name as the text label and apply OpenAI's standard prompt templates to generate the text embeddings. Formally, 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 $\boldsymbol{x}$, the predicted class is $\hat{y} = \arg\max_c \boldsymbol{w}_c^\top f_{\boldsymbol{\theta}}(\boldsymbol{x})$. As illustrated in Fig. 3a, we observe significant variability in zero-shot accuracy across 546 camera traps: BioCLIP 2 exceeds 90% accuracy on 161 camera traps, highlighting its strong potential in wildlife monitoring, yet 162 camera traps fall in the 50–80% range. This wide performance spread suggests that while zero-shot models provide a strong baseline, **there remains a persistent, critical need for site-specific adaptation.** We include zero-shot results for other foundation models in Sec. B.1 for reference.

4.2 Adaptation is Challenging in the Real World

When off-the-shelf foundation models underperform at a camera trap, a natural next step is to *adapt* them using labeled data collected at that site over time. The streaming evaluation protocol in StreamTrap mirrors this deployment lifecycle (Sec. 3): for a given camera trap at interval j, the model is updated using cumulative labeled training data collected up to interval j, and is then evaluated on test data from the future interval $j{+}1$.

Preliminary study. We first examine supervised fine-tuning, arguably the most common adaptation strategy. We define the **accumulated model** as a baseline that naively fine-tunes on *all* training data collected up to interval j. The model is initialized from BioCLIP 2, consistent with the zero-shot baseline above, and fine-tuned using standard cross-entropy loss with a cosine learning-rate scheduler and a base learning rate of 2.5×10^{-5}. Importantly, evaluation is always performed on unseen data from the future interval $j{+}1$.

	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15	C16	C17	C18	C19	C20	Avg
ZS	97.7	92.3	90.9	89.1	86.1	84.3	81.5	81.2	78.3	77.1	75.8	73.8	70.1	69.8	67.1	66.0	62.2	61.5	56.3	55.8	75.8
Accum	73.6	77.1	50.0	71.7	46.7	33.3	56.7	73.0	45.0	25.0	57.0	50.0	68.2	60.8	66.7	11.4	61.7	58.0	33.3	45.7	53.2
Δ	24.1	15.2	40.9	17.4	39.4	51.0	24.8	8.2	33.3	52.1	18.8	23.8	1.9	9.0	0.4	54.6	0.5	3.5	23.0	10.1	–
Accum⋆	98.7	89.9	89.9	90.8	79.4	79.2	83.4	88.3	81.0	78.7	76.9	70.1	79.4	75.4	86.9	76.6	88.9	71.3	76.7	82.6	82.2
Δ	1.0	2.4	1.0	1.7	6.7	5.1	1.9	7.1	2.7	1.6	1.1	3.7	9.3	5.6	19.8	10.6	26.7	9.8	20.4	26.8	–
Oracle⋆	98.7	96.4	94.5	96.6	87.8	91.5	91.0	91.0	87.3	82.0	84.3	88.5	91.0	86.6	91.0	83.9	93.8	89.6	89.9	93.4	90.4
Δ	1.0	4.1	3.6	7.5	1.7	7.2	9.5	9.8	9.0	4.9	8.5	14.7	20.9	16.8	23.9	17.9	31.6	28.1	33.6	37.6	–

Table 1: Accuracy difference (Δ) between zero-shot (ZS) and three fine-tuning methods: naive accumulated fine-tuning (Accum), accumulated fine-tuning with our adaptation recipe (Accum⋆), fine-tuned on all training data at a site with our recipe (Oracle⋆). **Blue** (**red**) denotes improvement (degradation) relative to ZS.

To ensure the generalizability of our findings, we select 20 representative camera traps for this preliminary study, strategically sampled to span a wide spectrum of zero-shot accuracies. As shown in Tab. 1, naive accumulated fine-tuning *consistently underperforms* the zero-shot baseline by a large margin. This result is particularly striking because the accumulated model is already an optimistic baseline: it uses all historical labeled data at each update step and imposes no explicit computation or storage constraint. Its failure therefore highlights the **severe generalization challenge posed by realistic temporal adaptation in camera-trap deployments**, and motivates a deeper diagnosis of the factors driving adaptation failure. More experimental details and comprehensive results are provided in Sec. B and Sec. C.

4.3 Diagnosing the Drivers of Adaptation Challenges

We hypothesize that the severe adaptation difficulties may stem from two distinct but compounding factors: (1) the inter-interval distribution shifts driven by temporal non-stationarity; (2) the inherent difficulty of the camera trap data, even when temporal shift is reduced.

Intrinsic difficulty even without temporal shift. To disentangle these factors, we first consider a deliberately *upper-bound* setting that removes the effect of inter-interval shift. Specifically, we train an **oracle** model using *all* labeled training data across *all* time intervals at a site, and evaluate it on held-out test data. This setting more closely resembles standard i.i.d. training than streaming adaptation, and one might therefore expect the oracle model to consistently outperform the zero-shot baseline. Surprisingly, this is not the case: on 214/546 camera traps (∼40%), the oracle model still underperforms the zero-shot model (Fig. 3b). This result indicates that temporal shift is not the sole source of difficulty; rather, the *optimization and generalization challenges inherent to camera-trap data* are already substantial even when train and test distributions are more aligned.

A key driver of this difficulty is the passive nature of camera-trap data collection. Because camera traps must "wait" for animals to appear, many species

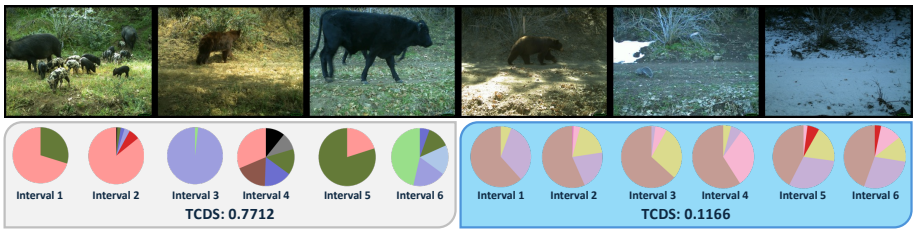

Fig. 4: Illustration of temporal shifts. (Top) Images from a camera trap showing shifting context (vegetation, lighting, weather) over time. (Bottom) TCDS comparison of two camera traps (pie chart: class frequency per interval): the left camera exhibits much higher temporal shifts (**TCDS=0.7712**) than the right one (**TCDS=0.1166**).

are observed only rarely within a given site or time period, leading to severe class imbalance. As shown in Fig. 2a, the two most frequent classes account for an average of ∼71% of all images at a camera trap. In contrast, the two least frequent classes contain only ∼49 images on average, compared to ∼4,500 images for the two majority classes. Attempting to fine-tune a foundation model on such severely imbalanced, low-shot data inevitably introduces bias, amplifies memorization, and ultimately harms generalization.

Temporal shift is pervasive. The inherent non-stationarity of natural ecosystems causes both the background context (vegetation, illumination, weather) and the animal population distributions to undergo profound shifts between consecutive intervals, as illustrated in Fig. 4 (Top). To formally quantify the *animal class distribution shift* between consecutive intervals, we propose the Temporal Class Distribution Shift (TCDS) metric for a specific camera trap. For class c at interval j, let the normalized class frequency be $p_j^c = \frac{n_j^c}{\sum_k n_j^k}$, where n_j^c is the sample count of class c in interval j. We can define:

$$\text{TCDS} := \overbrace{\frac{1}{n-1}\sum_{j=1}^{n-1}}^{\textit{Average over all intervals}} \underbrace{\sum_c \left| p_j^c - p_{j+1}^c \right|}_{\textit{Class distribution shift between intervals}}$$

TCDS captures the average magnitude of temporal shift in class distributions across all consecutive intervals for a camera trap. As illustrated in Fig. 4 (Bottom), this metric quantifies class distribution shifts across camera traps, where higher TCDS values correspond to stronger temporal shifts. Unfortunately, a substantial fraction of the camera traps exhibit exceptionally high TCDS values (Fig. 2a). Thus, a model that achieves high accuracy on the current interval can't guarantee to remain strong when confronted with the drastically shifted data distributions of future intervals.

Compounding effects. Severe data imbalance and temporal distribution shift are entangled in practice. Together, they make streaming adaptation substan-

tially more challenging than standard in-domain fine-tuning: models must learn from scarce and imbalanced observations while also generalizing to future intervals whose species distributions may differ significantly from the past. This diagnosis motivates the need for robust adaptation recipes that help end-users maintain reliable model performance in dynamic deployment environments.

4.4 Practical Recipes for Robust Adaptation

We next investigate practical techniques for addressing the adaptation challenges identified above. Our goal is to establish an effective "first-to-try" recipe that end-users can apply when adapting foundation models to camera-trap deployments.

Improving Generalization We first focus on strategies to mitigate the severe generalization degradation caused by extreme class imbalance and low-shot data. Motivated by prior literature on long-tailed recognition and efficient adaptation [36, 65], we investigate two primary directions:

Class-imbalanced loss. Standard cross-entropy fine-tuning on severely imbalanced data inherently biases the model toward majority classes, yielding poor generalization across classes [12, 61, 62]. Class-imbalanced losses offer a simple and effective mitigation. We explore several established methods, including Balanced Softmax (BSM) [43], Class-Balanced Focal Loss (CB-Focal) [12], and Class-Dependent Temperatures (CDT) [61]. As detailed in our ablation studies (Sec. B.3), we find that BSM delivers consistent improvements across camera traps without requiring hyperparameter tuning, whereas CB-Focal and CDT require more hyperparameter tuning, which is often infeasible for camera trap deployments. We therefore adopt BSM as our default imbalance-aware loss: $\mathcal{L}_{\text{BSM}}(\boldsymbol{x}, y) = -\log\left(\frac{n_y \exp(\eta_y(\boldsymbol{x}))}{\sum_c n_c \exp(\eta_c(\boldsymbol{x}))}\right)$ 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). Fully fine-tuning on imperfect data often risks corrupting the generalized representations of foundation models [31, 67]. Unlike full fine-tuning, which updates all network parameters, PEFT modifies only a minimal subset of weights, thereby preserving pre-trained knowledge [36, 64]. We evaluate three representative PEFT methods: Low-Rank Adaptation (LoRA) [22], Visual Prompt Tuning (VPT) [25, 54], and Adapter [21]. Our ablations (Sec. B.3) demonstrate that LoRA, which introduces trainable low-rank matrices into the attention projection layers, achieves superior performance without heavy hyperparameter tuning. Consequently, we adopt LoRA as our default fine-tuning mechanism.

Synergy between imbalance loss and PEFT. To rigorously assess the isolated and combined impacts of these two directions, we conduct an ablation study on a representative subset of 20 camera traps with the oracle setting (details in Sec. B.3). We observe that applying either the imbalance loss (BSM) or PEFT (LoRA) alone yields only marginal improvements (less than 1%) over naive full fine-tuning. However, combining BSM + LoRA yields a substantial performance boost, strongly suggesting their compatibility and complementarity. Motivated

by this result, we consider the **BSM + LoRA** as our recommended adaptation recipe for camera traps (denoted hereafter as $\star$). Applying this recipe to the oracle models (Oracle$\star$) across all 546 camera traps in the StreamTrap benchmark enables 474 sites (compared to only 332 for the naive oracle) to successfully outperform the zero-shot baseline (Fig. 3c). While there remain 72 camera traps where this recipe falls short, we provide a detailed qualitative and quantitative analysis of these specific failure cases in Sec. C.2.

Mitigating Temporal Shifts Our proposed adaptation recipe ($\star$) successfully transfers its effectiveness from the idealized oracle setting to the accumulated model evaluated under the realistic streaming protocol. As shown in Tab. 1, the accumulated model equipped with our recipe (denoted as Accum$\star$) consistently outperforms the zero-shot baseline. However, a non-negligible performance gap persists between Accum$\star$ and corresponding upper bounds (Oracle$\star$) on most camera traps. This discrepancy is primarily driven by inter-interval distribution shifts, which inherently complicate streaming evaluations. Mitigating such temporal shifts is notoriously difficult due to the unpredictability of future data distributions, a challenge that remains largely underexplored in the vision community. To bridge this gap, we evaluate various approaches and identify three promising post-processing techniques that can potentially improve Accum$\star$.

Post-hoc logit calibration. When some classes are absent or extremely low-shot during fine-tuning, the updated model tends to heavily suppress their predicted logits relative to the majority classes, even when the foundation model originally represented them well [31,53]. To rectify this bias, we adopt a post-hoc logit calibration [31] to boost the logits of absent or minority classes to a magnitude comparable to the majority classes by introducing a calibration factor γ: $\hat{y} = \arg\max_c \boldsymbol{w}_c^\top f_{\boldsymbol{\theta}}(\boldsymbol{x}) + \gamma \cdot \mathbf{1}[c \in \mathcal{A}]$, where $\mathcal{A}$ denotes absent/minority classes in the current interval.

Weight interpolation. While fine-tuning enables necessary task-specific adaptation, it risks forgetting of the foundation model's broad generalization capabilities. To strike a balance between generalizable pre-trained knowledge and site-specific features, we apply weight interpolation (WiSE) [59]. Let $\boldsymbol{\theta}'$ be the fine-tuned weights and $\boldsymbol{\theta}$ the pre-trained weights. The interpolated weights are given by $\boldsymbol{\theta}(\alpha) = \alpha\boldsymbol{\theta} + (1-\alpha)\boldsymbol{\theta}'$ where $\alpha \in [0, 1]$ controls the interpolation ratio.

Interval model selection. The most recently adapted model is not always the best model for the next interval. Due to recurring seasonal patterns or abrupt ecological changes, future interval $j{+}1$ may resemble an earlier interval more than the immediately preceding one. To assess the potential benefit of choosing among historical models, we conduct a diagnostic upper-bound study: for each future interval $j{+}1$, we select the historical or current model that achieves the highest accuracy on that interval's test split. This procedure is not directly deployable because it uses future labels, but it quantifies the possible gain from better model-selection mechanisms and motivates future work on practical selection criteria under streaming deployment.

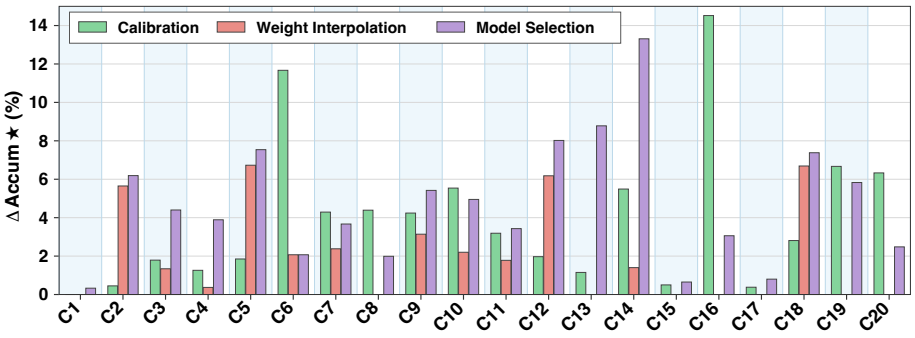

Fig. 5: Performance gain ($\Delta_{\text{Accum}\star}$) of calibration, weight interpolation, and model selection over the Accum* baseline. They are applied on top of Accum*. More results can be found in Sec. C.3.

Promises and limits of post-processing. Fig. 5 illustrates the relative gains of each post-adaptation strategy over Accum$\star$. Notably, for camera traps where Accum$\star$ initially underperforms the zero-shot baseline, at least one of these strategies can recover the performance drop. Moreover, these techniques can effectively narrow the gap between the Oracle$\star$ and Accum$\star$, underscoring their tremendous potential for mitigating temporal shift. However, we emphasize that these results are obtained under an optimistic upper-bound setting, where we select performance-maximizing hyperparameters to reveal the potential of each method. In real deployments, selecting such hyperparameters without access to future data remains an open challenge. We therefore view this as an important research opportunity and encourage future work to resolve the degradation caused by temporal shifts.

Recommended Adaptation Recipes

We recommend combining **LoRA** (PEFT) with **BSM** (imbalance loss) as a practical first-to-try adaptation strategy, augmented by post-processing techniques, such as **logit calibration**, **weight interpolation**, or **interval model selection**, to mitigate inter-interval distribution shift.

5 More Practical Considerations

Accurate real-world deployment involves more than selecting an adaptation algorithm. Practitioners must also decide when a foundation model is sufficient, whether continual adaptation is necessary, and when adaptation should be triggered. Despite their importance for deployment, these questions remain largely underexplored in the vision community. In this section, we provide preliminary analyses with simple baselines and highlight directions for future research.

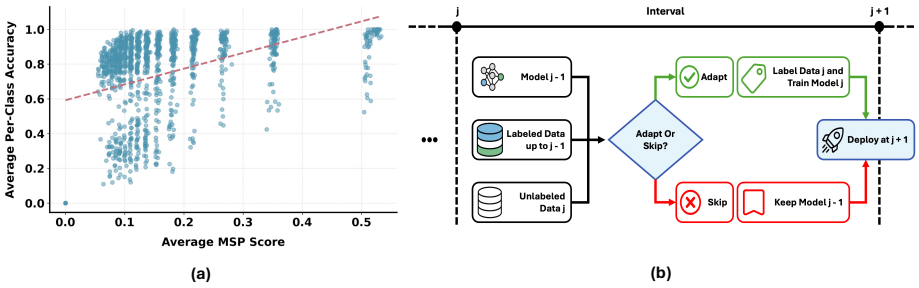

Fig. 6: (a). Non-OOD scores (MSP) moderately correlate with zero-shot accuracy. **(b).** The Adapt-or-Skip problem: should we adapt at this interval or not?.

When is the zero-shot model sufficient? In real-world deployments, practitioners need to estimate whether a foundation model's zero-shot performance is sufficient for a new site. If yes, the model can be deployed directly without data collection and adaptation. However, this decision must typically be made before labeled data from the new site are available, making zero-shot accuracy hard to assess in advance. Despite its practical importance, this problem remains largely underexplored in the vision community. As a simple baseline, we assess if out-of-distribution (OOD) signals can provide an indication of zero-shot accuracy [60]. The intuition is that if images from a new camera trap are closer to the training distribution, the model may produce higher-confidence predictions and achieve better zero-shot accuracy. Specifically, we adopt the Maximum Softmax Probability (MSP) [20], assigning each test sample a non-OOD score: $s_{\text{MSP}}(\boldsymbol{x}) = \max_c \frac{\exp(\eta_c(\boldsymbol{x})/\tau)}{\sum_{c'} \exp(\eta_{c'}(\boldsymbol{x})/\tau)}$ 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 plot the average non-OOD score for each camera trap against the zero-shot accuracy in Fig. 6a. While a moderate correlation can be observed, many camera traps deviate from this pattern, indicating that confidence signals alone are insufficient to reliably predict zero-shot performance. This highlights the difficulty of estimating zero-shot sufficiency prior to deployment, suggesting it as an open problem for future work.

Do we need continual adaptation? Another practical question raised by end-users is whether continual adaptation is necessary after a model has already been updated over several initial intervals. Once the model has learned site-specific characteristics, can it generalize to all future data without further fine-tuning? To investigate this question, we conduct a controlled study in which adaptation is deliberately halted after a fixed number of intervals. Specifically, for camera traps with long deployment durations spanning more than one year, we freeze the accumulated model at a specific interval and evaluate it on all remaining future intervals. As shown in Tab. 2, model accuracy generally improves as the model is adapted on more intervals. This trend indicates that continually incorporating additional site-specific data can provide cumulative performance gains. At the

same time, the benefit of each update is not uniform: some intervals provide substantially richer adaptation signals than others, while others offer limited marginal improvement. This observation motivates our next question: can we determine *when* an adaptation step is actually necessary?

Stage	0%	25%	50%	75%	100%
Acc.	76.2	73.1	77.8	82.7	84.9

Table 2: Accum$\star$ is frozen at a specific interval (*e.g.*, 25% means frozen after adaptation with the first 25% intervals) and evaluated on all future intervals. Averaged across 8 camera traps with long deployment durations spanning more than one year.

When should we adapt? Adaptation incurs labeling and computational costs, making it important to determine when an update is actually beneficial. However, this decision must be made before observing future performance. At every interval, the end-users will face an **Adapt-or-Skip** problem as illustrated in Fig. 6b: should we adapt at this interval or not? Despite its practical importance, this remains largely underexplored. At interval j, the user observes the deployed model f_{j-1}, prior labeled training data $\mathcal{D}^{\text{train}}_{1:j-1}$, and current unlabeled samples $\mathcal{X}_j$, and must decide whether to ADAPT or SKIP. If adapting, the user labels $\mathcal{X}_j$ and trains f_j with $\mathcal{D}^{\text{train}}_{1:j}$; if skipping, they keep deploying f_{j-1}. Retrospectively, the ground-truth decision is $a^\star_j = \text{ADAPT}$ iff $\text{Acc}(f_j, \mathcal{D}^{\text{test}}_{j+1}) > \text{Acc}(f_{j-1}, \mathcal{D}^{\text{test}}_{j+1})$, and $a^\star_j = \text{SKIP}$ otherwise. The challenge is to predict $a^\star_j$ using only information available at interval j, before observing $\mathcal{D}^{\text{test}}_{j+1}$. We construct a dataset where half of the intervals favor adapting and half favor skipping. We evaluate three simple baseline methods: random, MSP-based, and CLIP feature-based. However, both methods perform similarly to random guessing, highlighting the difficulty of the problem. For reference, we report an upper-bound that always selects the correct decision, which outperforms the three baselines by 11.34%, suggesting substantial room for future research. More details are provided in Sec. C.4.

6 Conclusion

We present the first unified study of camera-trap species recognition over time, prioritizing reliable, long-term fixed-site deployment. Our benchmark, STREAM-TRAP, uses a streaming evaluation protocol reflecting real-world data collection. Because naive foundation model fine-tuning fails under extreme data imbalance and temporal shifts, we provide robust adaptation recipes and analyze critical deployment challenges to guide practitioners and spur future research.

Acknowledgment

This research is supported by grants from the National Science Foundation (ICICLE: OAC-2112606). We are grateful for the generous support from the Ohio Supercomputer Center.

References

1. Lila bc: Labeled information library of alexandria: Biology and conservation. https://lila.science/, accessed January 2026
2. Aljundi, R., Lin, M., Goujaud, B., Bengio, Y.: Gradient based sample selection for online continual learning. Advances in neural information processing systems **32** (2019)
3. Anton, V., Hartley, S., Geldenhuis, A., Wittmer, H.U.: Monitoring the mammalian fauna of urban areas using remote cameras and citizen science. Journal of Urban Ecology **4**(1), juy002 (2018)
4. Beery, S.: The megadetector: Large-scale deployment of computer vision for conservation and biodiversity monitoring. California Institute of Technology, Pasadena, CA, USA (2023)
5. Beery, S., Agarwal, A., Cole, E., Birodkar, V.: The iwildcam 2021 competition dataset. arXiv preprint arXiv:2105.03494 (2021)
6. Beery, S., Van Horn, G., Mac Aodha, O., Perona, P.: The iwildcam 2018 challenge dataset. arXiv preprint arXiv:1904.05986 (2019)
7. Beery, S., Van Horn, G., Perona, P.: Recognition in terra incognita. In: Proceedings of the European conference on computer vision (ECCV). pp. 456–473 (2018)
8. Bevan, P.A., Pantazis, O., Pringle, H., Ferreira, G.B., Ingram, D.J., Madsen, E., Thomas, L., Thanet, D.R., Silwal, T., Rayamajhi, S., 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)
9. Bian, A., Li, W., Yuan, H., Wang, M., Zhao, Z., Lu, A., Ji, P., Feng, T., et al.: Make Continual Learning Stronger via C-Flat. NeurIPS (2024)
10. Boitani, L.: Camera trapping for wildlife research. Pelagic Publishing Ltd (2016)
11. Bothmann, L., Wimmer, L., Charrakh, O., Weber, T., Edelhoff, H., Peters, W., Nguyen, H., Benjamin, C., Menzel, A.: Automated wildlife image classification: An active learning tool for ecological applications. Ecological Informatics **77**, 102231 (2023)
12. Cui, Y., Jia, M., Lin, T.Y., Song, Y., Belongie, S.: Class-balanced loss based on effective number of samples. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9268–9277 (2019)
13. Fabian, Z., Miao, Z., Li, C., Zhang, Y., Liu, Z., Hernández, A., Montes-Rojas, A., Escucha, R., Siabatto, L., Link, A., et al.: Multimodal foundation models for zero-shot animal species recognition in camera trap images. arXiv preprint arXiv:2311.01064 (2023)
14. Farahani, A., Voghoei, S., Rasheed, K., Arabnia, H.R.: A brief review of domain adaptation. Advances in data science and information engineering: proceedings from ICDATA 2020 and IKE 2020 pp. 877–894 (2021)
15. Feng, T., Li, W., Zhu, D., Yuan, H., Zheng, W., Zhang, D., Tang, J.: Zeroflow: Overcoming Catastrophic Forgetting is Easier than You Think. In: ICML (2025)

16. Feng, T., Wang, M., Yuan, H.: Overcoming Catastrophic Forgetting in Incremental Object Detection via Elastic Response Distillation. In: CVPR (2022)
17. Gabeff, V., Rußwurm, M., Tuia, D., Mathis, A.: 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)
18. Gong, B., Shi, Y., Sha, F., Grauman, K.: Geodesic flow kernel for unsupervised domain adaptation. In: 2012 IEEE conference on computer vision and pattern recognition. pp. 2066–2073. IEEE (2012)
19. Gu, J., Stevens, S., Campolongo, E.G., Thompson, M.J., Zhang, N., Wu, J., Kopanev, A., Mai, Z., White, A.E., Balhoff, J., et al.: Bioclip 2: Emergent properties from scaling hierarchical contrastive learning. arXiv preprint arXiv:2505.23883 (2025)
20. Hendrycks, D., Gimpel, K.: A baseline for detecting misclassified and out-of-distribution examples in neural networks. arXiv preprint arXiv:1610.02136 (2016)
21. Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., de Laroussilhe, Q., Gesmundo, A., Attariyan, M., Gelly, S.: Parameter-efficient transfer learning for nlp (2019), https://arxiv.org/abs/1902.00751, accessed January 2026
22. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al.: Lora: Low-rank adaptation of large language models. ICLR **1**(2), 3 (2022)
23. Idaho Department of Fish and Game: Idaho camera traps. https://lila.science/datasets/idaho-camera-traps/, accessed January 2026
24. IUCN SSC Asian Wild Cattle Specialist Group's Saola Working Group: Northern and central annamites camera traps 2.0. Dataset (2021), sWG (2021)
25. Jia, M., Tang, L., Chen, B.C., Cardie, C., Belongie, S., Hariharan, B., Lim, S.N.: Visual prompt tuning (2022), https://arxiv.org/abs/2203.12119, accessed January 2026
26. Koh, P.W., Sagawa, S., Marklund, H., Xie, S.M., Zhang, M., Balsubramani, A., Hu, W., Yasunaga, M., Phillips, R.L., Gao, I., et al.: Wilds: A benchmark of in-the-wild distribution shifts. In: International conference on machine learning. pp. 5637–5664. PMLR (2021)
27. Lee, J., Mai, Z., Yoo, J., Fan, C., Zhang, C., Chao, W.L.: Continual unlearning for text-to-image diffusion models: A regularization perspective. arXiv preprint arXiv:2511.07970 (2025)
28. Li, W., Yuan, H., Zhao, Z., Kang, B., Liu, Z., Feng, T.: A faster path to continual learning (2026)
29. 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)
30. Mai, Z., Chowdhury, A., Wang, Z., Jeon, S., Wang, L., Hou, J., Kil, J., Chao, W.L.: Ava-bench: Atomic visual ability benchmark for vision foundation models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 25925–25937 (2026)
31. Mai, Z., Chowdhury, A., Zhang, P., Tu, C.H., Chen, H.Y., Pahuja, V., Berger-Wolf, T., Gao, S., Stewart, C., Su, Y., et al.: Fine-tuning is fine, if calibrated. Advances in Neural Information Processing Systems **37**, 136084–136119 (2024)
32. Mai, Z., Huang, Z., Jeon, S., Yoo, J., Lee, C., Zhang, G., Wang, H., Li, Y., Kong, Y., Yan, Y., et al.: A survey of continual learning for robotics in the foundation model era (2026)
33. Mai, Z., Li, R., Jeong, J., Quispe, D., Kim, H., Sanner, S.: Online continual learning in image classification: An empirical survey. Neurocomputing **469**, 28–51 (2022)

34. Mai, Z., Li, R., Kim, H., Sanner, S.: 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. pp. 3589–3599 (2021)
35. Mai, Z., Zhang, K., Wang, F.E., Wang, Z.K., Chen, A.Y., Xia, L., Sun, M., Chao, W.L., Kuo, C.H.: Revisiting model stitching in the foundation model era. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 41342–41351 (2026)
36. Mai, Z., Zhang, P., Tu, C.H., Chen, H.Y., Nguyen, Q.H., Zhang, L., Chao, W.L.: 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. pp. 14845–14857 (2025)
37. Malik, F., Wouters, S., Cartuyvels, R., Ghadery, E., Moens, M.F.: Two-phase training mitigates class imbalance for camera trap image classification with cnns. arXiv preprint arXiv:2112.14491 (2021)
38. New Zealand Trailcams: Trail camera images of new zealand animals. https://lila.science/datasets/nz-trailcams, accessed January 2026
39. 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)
40. Pardo, L.E., Bombaci, S., Huebner, S.E., Somers, M.J., Fritz, H., Downs, C., Guthmann, A., Hetem, R.S., Keith, M., Roux, A.l., et al.: Snapshot safari: A large-scale collaborative to monitor africa's remarkable biodiversity. South African Journal of Science **117**(1-2), 1–4 (2021)
41. Pollock, L.J., Kitzes, J., Beery, S., Gaynor, K.M., Jarzyna, M.A., Mac Aodha, O., Meyer, B., Rolnick, D., Taylor, G.W., Tuia, D., et al.: Harnessing artificial intelligence to fill global shortfalls in biodiversity knowledge. Nature Reviews Biodiversity pp. 1–17 (2025)
42. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision (2021), https://arxiv.org/abs/2103.00020, accessed January 2026
43. Ren, J., Yu, C., Ma, X., Zhao, H., Yi, S., et al.: Balanced meta-softmax for long-tailed visual recognition. Advances in neural information processing systems **33**, 4175–4186 (2020)
44. Rezvani, S., Wang, X.: A broad review on class imbalance learning techniques. Applied Soft Computing **143**, 110415 (2023)
45. Sagawa, S., Koh, P.W., Lee, T., Gao, I., Xie, S.M., Shen, K., Kumar, A., Hu, W., Yasunaga, M., Marklund, H., et al.: Extending the wilds benchmark for unsupervised adaptation. arXiv preprint arXiv:2112.05090 (2021)
46. Santamaria, J.D., Isaza, C., Giraldo, J.H.: Catalog: A camera trap language-guided contrastive learning model. In: 2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 1197–1206. IEEE (2025)
47. Shim, D., Mai, Z., Jeong, J., Sanner, S., Kim, H., Jang, J.: Online class-incremental continual learning with adversarial shapley value. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 35, pp. 9630–9638 (2021)
48. Singhal, P., Walambe, R., Ramanna, S., Kotecha, K.: Domain adaptation: challenges, methods, datasets, and applications. IEEE access **11**, 6973–7020 (2023)

49. Stevens, S., Wu, J., Thompson, M.J., Campolongo, E.G., Song, C.H., Carlyn, D.E., Dong, L., Dahdul, W.M., Stewart, C., Berger-Wolf, T., et al.: Bioclip: A vision foundation model for the tree of life. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 19412–19424 (2024)
50. Swanson, A., Kosmala, M., Lintott, C., Simpson, R., Smith, A., Packer, C.: Snapshot serengeti, high-frequency annotated camera trap images of 40 mammalian species in an african savanna. Scientific data **2**(1), 1–14 (2015)
51. Tabak, M.A., Norouzzadeh, M.S., Wolfson, D.W., Sweeney, S.J., VerCauteren, K.C., Snow, N.P., Halseth, J.M., Di Salvo, P.A., Lewis, J.S., White, M.D., et al.: Machine learning to classify animal species in camera trap images: Applications in ecology. Methods in Ecology and Evolution **10**(4), 585–590 (2019)
52. Trolliet, F., Vermeulen, C., Huynen, M.C., Hambuckers, A.: Use of camera traps for wildlife studies: a review. Biotechnologie, Agronomie, Société et Environnement **18**(3) (2014)
53. Tu, C.H., Chen, H.Y., Mai, Z., Zhong, J., Pahuja, V., Berger-Wolf, T., Gao, S., Stewart, C., Su, Y., Chao, W.L.H.: Holistic transfer: towards non-disruptive fine-tuning with partial target data. Advances in Neural Information Processing Systems **36**, 29149–29173 (2023)
54. Tu, C.H., Mai, Z., Chao, W.L.: Visual query tuning: Towards effective usage of intermediate representations for parameter and memory efficient transfer learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 7725–7735 (2023)
55. Tuia, D., Kellenberger, B., Beery, S., Costelloe, B.R., Zuffi, S., Risse, B., Mathis, A., Mathis, M.W., Van Langevelde, F., Burghardt, T., et al.: Perspectives in machine learning for wildlife conservation. Nature communications **13**(1), 792 (2022)
56. Van Horn, G., Mac Aodha, O., Song, Y., Cui, Y., Sun, C., Shepard, A., Adam, H., Perona, P., Belongie, S.: The inaturalist species classification and detection dataset. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 8769–8778 (2018)
57. Velasco-Montero, D., Fernández-Berni, J., Carmona-Galán, R., Sanglas, A., Palomares, F.: Reliable and efficient integration of ai into camera traps for smart wildlife monitoring based on continual learning. Ecological Informatics **83**, 102815 (2024)
58. Vélez, J., McShea, W., Shamon, H., Castiblanco-Camacho, P.J., Tabak, M.A., Chalmers, C., Fergus, P., Fieberg, J.: An evaluation of platforms for processing camera-trap data using artificial intelligence. Methods in Ecology and Evolution **14**(2), 459–477 (2023)
59. Wortsman, M., Ilharco, G., Kim, J.W., Li, M., Kornblith, S., Roelofs, R., Lopes, R.G., Hajishirzi, H., Farhadi, A., Namkoong, H., et al.: Robust fine-tuning of zero-shot models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 7959–7971 (2022)
60. Yang, J., Zhou, K., Li, Y., Liu, Z.: Generalized out-of-distribution detection: A survey. International Journal of Computer Vision **132**(12), 5635–5662 (2024)
61. Ye, H.J., Chen, H.Y., Zhan, D.C., Chao, W.L.: Identifying and compensating for feature deviation in imbalanced deep learning. arXiv preprint arXiv:2001.01385 (2020)
62. Ye, H.J., Zhan, D.C., Chao, W.L.: Procrustean training for imbalanced deep learning. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 92–102 (2021)
63. Yu, X., Wang, J., Kays, R., Jansen, P.A., Wang, T., Huang, T.: Automated identification of animal species in camera trap images. EURASIP Journal on Image and Video Processing **2013**(1), 52 (2013)

64. Zhang, P., Mai, Z., Nguyen, Q.H.P., Chao, W.L.H.: Revisiting semi-supervised learning in the era of foundation models. Advances in Neural Information Processing Systems **38**, 59295–59324 (2026)
65. Zhang, Y., Kang, B., Hooi, B., Yan, S., Feng, J.: Deep long-tailed learning: A survey. IEEE transactions on pattern analysis and machine intelligence **45**(9), 10795–10816 (2023)
66. Zhang, Y., Cai, C., Liu, F., Hamm, J.: Prime Once, then Reprogram Locally: An Efficient Alternative to Black-Box Model Reprogramming. OpenReview (2025), https://openreview.net/forum?id=Hic8PwzlBY, accessed January 2026
67. Zhang, Y., Mehra, A., Hamm, J.: Ot-vp: Optimal transport-guided visual prompting for test-time adaptation. In: IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). pp. 1122–1132. IEEE (2025), oral Presentation
68. Zhang, Y., Mehra, A., Niu, S., Hamm, J.: Dpcore: Dynamic prompt coreset for continual test-time adaptation. In: Forty-second International Conference on Machine Learning (ICML) (2025)
69. Zhang, Y., Niu, S., Liu, F., Hamm, J.: Adapting in the Dark: Towards Stable and Efficient Black-Box Test-Time Adaptation. OpenReview (2026), https://openreview.net/forum?id=KBtySUdWuH, accessed January 2026
70. Zhou, K., Liu, Z., Qiao, Y., Xiang, T., Loy, C.C.: Domain generalization: A survey. IEEE transactions on pattern analysis and machine intelligence **45**(4), 4396–4415 (2022)
71. Zhu, H., Tian, Y., Zhang, J.: Class incremental learning for wildlife biodiversity monitoring in camera trap images. Ecological Informatics **71**, 101760 (2022)