Target journal: (1) Methods in Ecology and Evolution, (2) Journal of Applied Ecology

Here is the link to edit figures online.

Computer vision for wildlife identification in camera trap images: Fine-tuning SpeciesNet outperforms local models for species classification

Authors

Prakash Palanivelu Rajmohan [1, 2], Renuka Sharma [1], Zachary Amir [1,3,4], Tom Bruce [1], Barry W. Brook[5], Dan Morris [6], Matthew Scott Luskin [1,4,7] *

Affiliations

[1] Wildlife Observatory of Australia (WildObs), Queensland Cyber Infrastructure Foundation (QCIF), Brisbane, Queensland, Australia

[2] School of Electrical Engineering and Computer Science, University of Queensland, Brisbane, Queensland, Australia

[3] Terrestrial Ecosystem Research Network, University of Queensland, Brisbane, Queensland 4072, Australia

[4] Centre for Biodiversity and Conservation Science, University of Queensland, Brisbane, Queensland, Australia

[5] School of Natural Sciences, University of Tasmania, Hobart, Tasmania, Australia

[6] Google, Mountain View, California, USA

[7] School of the Environment, University of Queensland, Brisbane, Queensland, Australia

*m.luskin@uq.edu.au

Abstract

Wildlife camera trap projects generate millions of images that exceed the capacity of manual processing. Computer vision (CV) is a branch of artificial intelligence (AI) and machine learning (ML) that ecologists increasingly use to increase image processing efficiently. The CV workflow generally starts with detecting animals (usually with MegaDetector) and then, for those images with animals, the snip (cropped area with the object) is passed to a classifier to identify species. SpeciesNet is an open-source AI/ML classifier capable of recognising thousands of vertebrate species globally and can therefore be referred to as a 'global model'. However, SpeciesNet has substantial geographic and taxonomic gaps. Ecologists in such 'missing' places may resort to building local classifiers for their particular sites and species. We hypothesized that a blended approach - fine-tuning SpeciesNet - could harness the respective strengths of each, with global feature representations, classes limited to those occurring at the site, and local taxonomic specialisation. Within this context, we address three questions: (i) How do local, global, and fine-tuned classifiers compare? (ii) How many training images are required? (iii) How does performance vary between random distribution and out-of-distribution testing? We used the Wildlife Observatory of Australia's (WildObs) tagged image repository for the 'Wet Tropics' rainforests (n = 454 cameras, 2,184,664 images, 121 species), and refined this to a balanced dataset of the 15 most common species for CV training and testing. We found that (i) fine-tuning SpeciesNet delivered the highest performance, often exceeding 95% accuracy, (ii) performance plateaued after 250-500 local training images per class (species), and (iii) these advantages were pronouced in out-of-distribution testing (i.e., for new cameras withheld from any model training). We conclude that fine-tuning SpeciesNet reconciles the longstanding tension between broad applicability and site-specific precision, accelerating image-to-inference workflows to achieve results within management-relevant timelines. Such advances move cameras further towards being an automated, easy, affordable, and efficient solution for wildlife monitoring, research, and conservation.

1. Introduction

There has been exponential growth in automated environmental monitoring technologies, transforming environmental monitoring and ecological research by generating unprecedented volumes of data with greater efficiency (Ahumada et al., 2020)(Bruce et al., 2025). Wildlife cameras the most developed example for detecting terrestrial vertebrates, including for cryptic mammals previously lacking significant observations (Bruce et al., 2025). Beyond species presence, camera-trap datasets can also reveal individual characteristics, health, behaviour, and even species interactions (Rowcliffe et al., 2014). However, the sheer volume of images generated by camera networks creates a significant bottleneck: manually reviewing millions of images is time- and resource-intensive, diverting effort from ecological interpretation and decision-making (Celis et al., 2024). This is a growing problem as networks of wildlife cameras are being deployed globally (Mendes et al 2024; Bruce et al. 2025). Efficiently and accurately processing and analysing images is a crucial step in extracting meaningful ecological insights, especially within timelines relevant to management, which we call the images-to-inferences pipeline. A surge in computer-vision (CV) machine-learning (ML) methods has sought to address this by automating object detection and species classification, often lumped in under the artificial intelligence (AI) umbrella (Besson et al., 2022; van Lunteren et al., 2025; Yang, 2019). However, these applications of CV are missing for locations and species lacking widely available training datasets and dedicated attention.

There are often two complementary CV steps used in ecological applications: object detectors and image classifiers. The workflow generally starts with an raw image from a camera trap trigger being sent to an object detection model, such as MegaDetector (Morris, n.d.). This identifies blanks and applies a bounding box snip to denote the area of the image containing interesting information (i.e., an animal). Such detectors are designed to efficiently locate and delineate coarse-level object categories of interest (<5 classes; often blank, humans, vehicles, and animals). In the second stage, either the entire image containing the animal and its background, or more commonly, just the cropped bounding box area(s) containing the animal snips, is passed to a classifier to identify the species. The complementarity arises from MegaDetector performing efficient object detection across a broad range of image contexts (megapixels, lighting, geographic, and taxonomic scope), whereas a classifier typically accepts its snippets but can assign dozens or even thousands of classes. The results from the classifier are then exported into standardized data formats (e.g. spreadsheets, JSON) for use in downstream ecological analysis and decisions for wildlife conservation and monitoring.

1.1 Global CV classifiers

Global species classifiers are trained on geographically and taxonomically diverse datasets spanning multiple continents. These models benefit from large sample sizes (>50M images, >10k per species) and recognize of thousands of species, providing utility across ecosystems, and even worldwide. They utilise sophisticated AI/ML/CV architectures designed for broad predictive generalisation, where the latter term refers to model performance on new, unseen data (also called 'out-of-distribution'). Their key advantage lies in scalability and broad usability; however, they may exhibit suboptimal performance in regions or taxa that are underrepresented in their training datasets (Gadot et al., 2024).

SpeciesNet (Gadot et al., 2024) has become the dominant global classifier for wildlife (Fleuré et al., 2025). SpeciesNet was trained on an extensive and diverse image dataset provided by Wildlife Insights (WI), a cloud-based platform supported by a consortium of NGOs and powered by

Google.org. Global models such as SpeciesNet are frequently adopted due to their capability to process extensive datasets spanning diverse ecological conditions. For example, it includes images from dozens of countries and has extensive taxonomic coverage (>2000 species), but is not exhaustive. SpeciesNet's training data is biased towards the northern hemisphere's developed nations, and gaps often occur in where there is slower adoption of cameras or WI (commonly in developing areas with limited internet), regions with smaller physical area like islands (often where endemism is also high), and areas with fewer research initiatives. This is the situation reported in our case study area of Australian tropical rainforests, which are relatively remote, have poor internet, and where WI has also faced limited adoption due to data-sharing hesitancy (see Methods; Bruce *et al.,* 2025). Thus, while global models offer impressive utility and scalability, their efficacy may be limited in some localised ecological and technical conditions. Finally, their black-box and un-customisable nature (e.g., inability to add species or limit classes to only those species locally present) can pose challenges in research contexts where transparency and reproducibility are required.

1.2 Local CV classifiers

Local classifiers, in contrast, are trained exclusively on datasets specific to a particular site, region, or other defined grouping. By learning from locally collected images, these models capture unique visual and ecological characteristics of their focal communities, often achieving suitable accuracy within their narrow context. Local classifiers aldo offer higher transparency and customisation to specific ecosystems. They typically achieve suitable precision for rare or locally unique species, yet their performance deteriorates markedly outside the original training environment. Local models require domain expertise and computational resources, as well as suitable amounts of labelled images and thus annotation effort (human time). There is also debate about the number of training images required for local models and the criteria for acceptable performance and perform poorly when generalised beyond conditions similar to their training. Local classifiers can be used on cloud-based camera image management platforms such as TrapTagger and Agouti, and can be trained and used offline with the Mega Efficient Wildlife Classifier (MEWC) workflow (Brook et al. 2023).

1.3 Fine-tuning global classifiers

Fine-tuning is a transfer learning method that adapts a pre-trained global classifier by refining its weights to better recognise local features, such as species traits and habitat backgrounds (Yosinski et al., 2014). This hybrid approach aims to leverage the general visual representations from broad-scale global models and refines them with local data, aiming to combine their strengths. It is expected to enhance accuracy, data efficiency, and robustness to ecological variability, thereby reconciling the longstanding tension between global applicability and local precision. Fine-tuning, in this ecological CV context, often involves two approaches: (i) pruning classes from the global model to match the focal community, and (ii) retraining the global model with supplementary local images, which can include adding entirely new classes absent from the global training set. We clearly delineate these methods to enhance clarity, noting that the MEWC system requires custom coding for class pruning.

1.4 Performance of CV approaches

Despite the increasing adoption of both global and local classifiers in ecology, and recent emergence of fine-tuning, there are few systematic comparisons. Existing research has typically benchmarked individual models within specific contexts, such as MegaDetector for object detection across global datasets (Yang, 2019) or local classifiers within targeted ecological projects (Norouzzadeh et al., 2021), without comparisons of their trade-offs (Tabak et al., 2019). Given the rapid and widespread uptake of CV methods, comparative evaluations are urgently needed to clarify the trade-offs between scalability, accuracy, and robustness under both random ‘in-distribution’ and out-of-distribtion scenarios. This is especially true for environmental managers and conservation practitioners that increasingly rely on CV-based workflows for high-stakes decisions (Christin et al., 2019).

Here, we addressing this gap using a systematic and empirical comparison of global, local, and fine-tuned models. We conduct experiments to (i) evaluate the performance of these three approaches (accuracy, precision, recall, F1-score), (ii) determine how many local training images are required to meet performance thresholds, including the point of diminishing returns, and (iii) assess differences in performance for random distribution versus out-of-distribution applications. We train and test these models using images from Australian Wet Tropics (AWT) rainforest biodiversity hotspot, where a massive wildlife camera trap dataset has been made publicly available. Our goal is to provide ecologists with practical guidance on how to process camera trap images efficently.

2. Methods

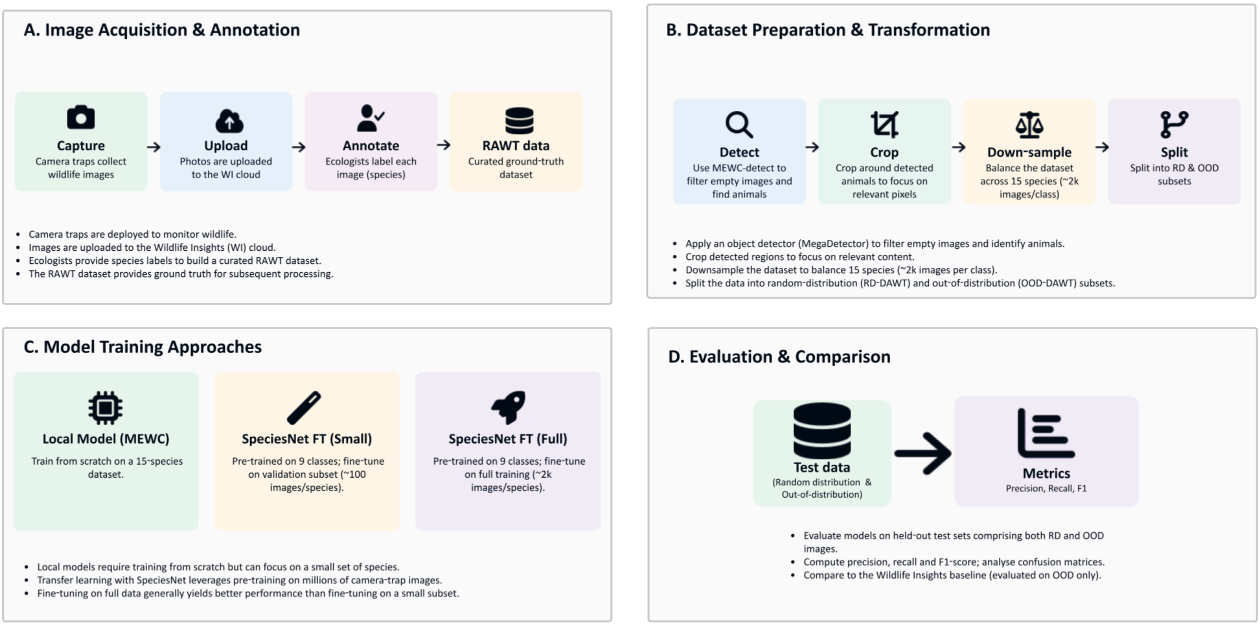


Figure 1 | Computer vision (CV) workflow for the wildlife camera case study presented here.

2.1 Assumptions and limitations

We are making the following acknowledgements or assumptions in this work:

Limited class scope: The study is constrained to a 15-class classification problem, representing a subset of species commonly found in the AWT dataset. This limitation is intended to create a manageable and controlled experimental setting rather than to cover full species diversity.

Classification task: This study focuses on a standard image classification task, where each image is assumed to contain a single identifiable species. Tasks involving multiple species per image (multi-label classification) are beyond our scope.

Methodological emphasis: Our objective is to evaluate the general effectiveness of three approaches: locally trained models, global SpeciesNet with class restriction, and SpeciesNet fine-tuned on local images. We prioritize methodological comparison over maximizing absolute metrics for any specific dataset, thereby avoiding overfitting and ad hoc tuning that could produce artificially inflated results.

2.2 Wildlife image acquisition and annotation

2.1.1 Data acquisition: camera deployment

The Wildlife Observatory of Australia (WildObs) established an extensive Australian Wet Tropics (AWT) camera-trap dataset by deploying 454 unbaited cameras between 2020 and 2024 (Anderson et al., 2025). These cameras generated 2,184,664 wildlife images of >120 vertebrate species including mammals, birds, and reptiles.

2.1.2 Data Storage, Management, and Labelling

All raw camera-trap images were uploaded to WI's Google Cloud Platform buckets. Importantly, none of the AWT images were used to train SpeciesNet, ensuring that our evaluation data remained entirely independent from any model training or pre-training process.

Images stored in WI were manually annotated by a team of trained ecologists to establish ground-truth species labels, requiring >1000 person hours. These labels served as the authoritative source for all the downstream tasks. Images containing multiple species, blanks, or vehicles were excluded from further analysis. This curated collection of annotated images constitutes the Raw Australian Wet Tropics (RAWT) dataset with a total of 412,822 images, of which 29,817 have been made publicly available in a curated format suitable for replication of our study, Wildlife Observatory of Australia (WildObs, 2025). Following annotation, all labelled images were automatically retrieved from WI using the University of Queensland's Bunya high-performance computer and stored in ARDC's Nectar system for processing.

2.2 Preparation of image dataset

2.2.1 Object detection and double filtering

We first processed the RAWT dataset using MEWC. This workflow employs MegaDetector to remove irrelevant images (blanks, vehicles) and generate bounding boxes for animal localization, producing cropped animal snips. This secondary filtering improved dataset quality by ensuring classification focused exclusively on valid animal detections.

2.2.2 Species-centred cropping

Using the bounding box outputs, we cropped each image to the detected animal snip, producing the Australian Wet Tropics Snipped (AWT-Snip) dataset. This deliberate design choice minimizes background content to prevent models from overfitting to habitat cues and camera placement, which degrades generalisation to new sites (Beery et al., 2020). Minimising background content forces models to learn morphological and textural features of the species themselves, a strategy shown to improve accuracy and robustness in ecological tasks (Norouzzadeh et al., 2021; Gadot et al 2024).

The full AWT-Snip dataset exhibited a strongly long-tailed class distribution, with most species represented by fewer than 10 images (Fig. S1). We created a balanced dataset by down-sampling to the 15 most frequent species (Fig. 2), each with at least 2,200 images, resulting in the balanced AWT dataset (~391,948 images). Six species in this balanced dataset were absent from SpeciesNet's original training data: *Casuarius casuarius*, *Heteromyia cinereifrons*, *Hypsiprymnodon moschatus*, *Megapodius reinwardt*, *Orthonyx spaldingii*, and *Uromys caudimaculatus*.

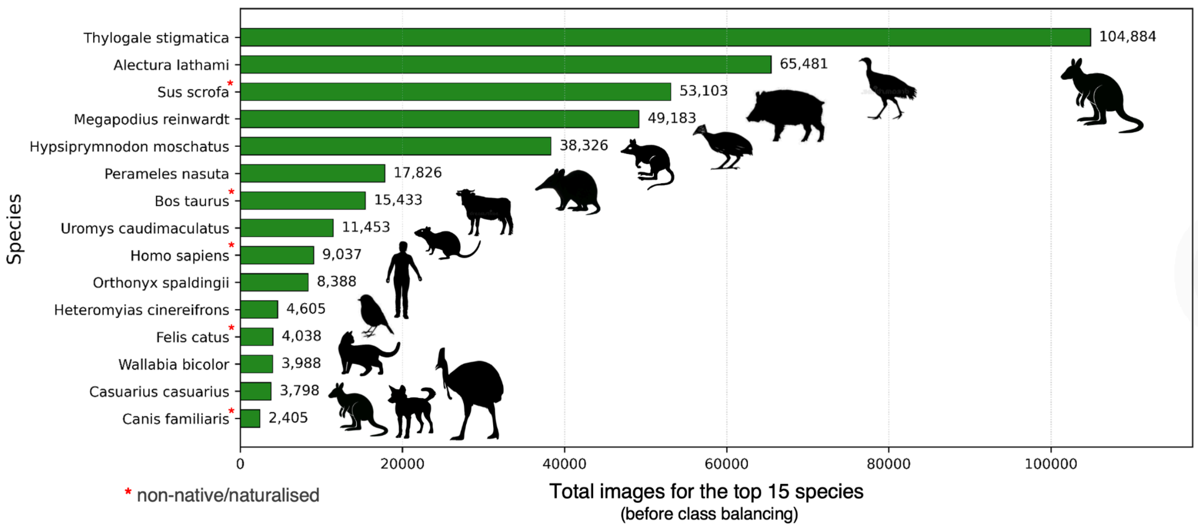


Figure 2. Total images collected for the 15 species included in the CV training and testing (prior to class balancing).

2.2.4 Partitioning for Robust Evaluation

To evaluate model performance under different deployment scenarios, downsampled and balanced was partitioned into two complementary subsets:

Random Distribution (RD): RD data refers to images that likely share similar statistical properties and environmental conditions between the training and evaluating phases (same or similar site locations, ensuring consistency in species, habitat, imaging conditions, and camera settings). RD training and test images may be from the same cameras or locations with minimal domain shift. RD is useful when the application is expected to be for new images collected from the same sites it was trained on. For instance, a model trained on data from Sites 1, 2, and 3 collected between 2020-2022, would be used to classify new images from the same sites in 2024–2025. The model is expected to generalize well because it has learned relevant features specific to these sites and their species.

Out-of-distribution (OOD): OOD data refers to images that exhibit significant differences in statistical properties and environmental conditions between the training and evaluating phases. Specifically, OOD data consists of images collected from different site locations, species, habitats, or imaging conditions, introducing a domain shift that the classifier has not been explicitly trained to handle. For example, if a model is trained using images from Sites 1 and 2 but tested on data from Site 3. This tests if there is overfitting to background features or site-specific cues and instead of species-specific visual characteristics.

We created two evaluation datasets from the balanced AWT dataset by partitioning images into training (90%), validation (5%), and testing (5%) sets with equal representation per class (Table 1). For the Random Distribution (RD) dataset, images were randomly split across all cameras. For the Out-of-distribution (OOD) dataset, we partitioned by camera deployment, ensuring no camera's images appeared in multiple splits (Figure 3). Since SpeciesNet's hyperparameters were already optimized, we repurposed the validation images for fine-tuning experiments (used it as a smaller training set).

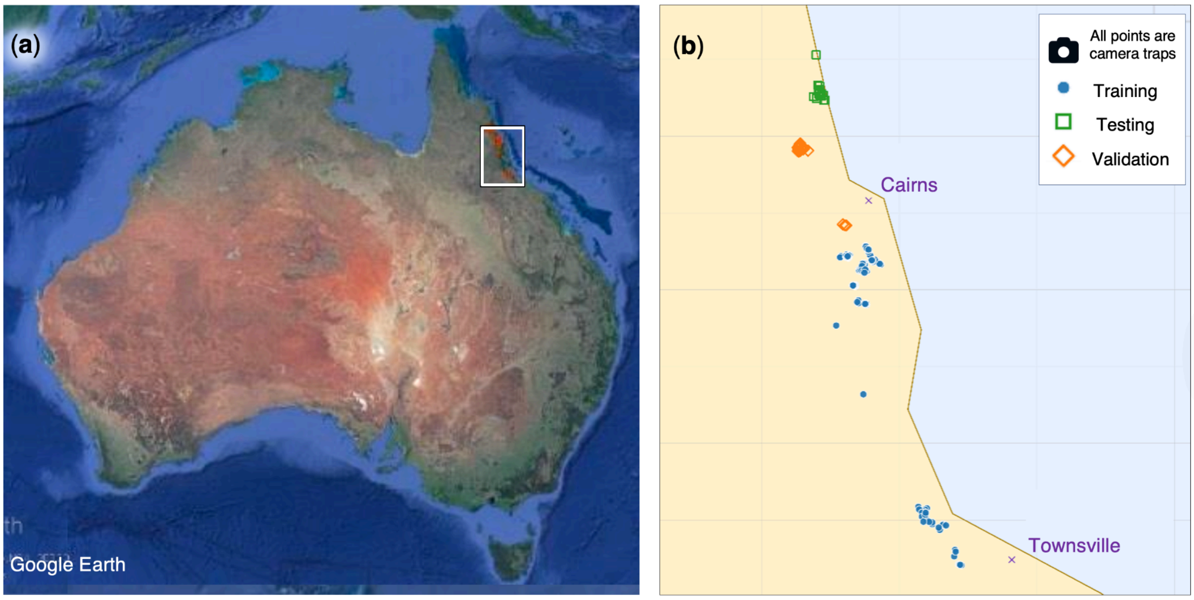


Figure 3. Study site in the Australian wet tropics rainforests (a) and specific locations of 454 cameras that produced the wildlife image dataset (b). For in-random distribution (RD) training and testing (not shown), images from any camera could be used for either training or testing. For the out-of-distribution (ODD) training and testing (b), camera deployments were grouped into three areas and exclusively assigned to training, validation, or testing sets. Major cities are labelled for geographic reference (purple).

Datasets	**Number of Observations**	**Classes (species)**	**Used for downstream CV training?**
All images	2,184,664	121	No
Species images, produce AWT-Snip	412,822	121	No
Restrict to 15 classes	391,948	15	No
Downsampled and balanced-RD	36,075	15	Yes
Downsampled and balanced-OOD	30,600	15	Yes

Table 1: Statistics of each dataset after each pre-processing step. Downsampling was used to balance images per species by removing species with less than 2200 total images

2.3 Model training approaches

Types of CV approaches

2.3.1 Local model

We trained a site-specific local model from scratch using the MEWC framework, tailored to the 15-species community without external data. The model employed an EfficientNetV2-M backbone (Tan & Le, 2021) with ImageNet initialization and a new 15-class output layer. Training used AdamW (Kingma & Ba, 2014) optimization with strong augmentation and a two-phase curriculum: initial classifier learning with frozen features, followed by progressive fine-tuning of higher layers with increasing regularization. To evaluate data requirements, we trained models across a gradient of training set sizes (1-100% of available data) with multiple random seeds, conducting this analysis under both random distribution and out-of-distribution splits. Complete training specifications, computational environment, and detailed curriculum are provided in Appendix S2.

2.3.2 SpeciesNet fine-tuned with limited local data (SNet-FT-Small)

We created SNet-FT-Small by fine-tuning SpeciesNet with minimal local data to combine global knowledge with local specialisation. Starting with pre-trained SpeciesNet weights, we pruned the output layer to our 15 target species, reusing weights for the 9 overlapping species and randomly initializing the 6 novel species. Fine-tuning used only the smaller training set (5% of images) with Adam optimization, data augmentation (Shorten & Khoshgoftaar, 2019), and a 20-epoch budget, updating only the classifier and high-level layers while preserving lower-level features. This approach demonstrates minimal adaptation of a global classifier using limited local examples. Complete implementation details are provided in Appendix S3.

2.3.4 SpeciesNet fine-tuned with all local data (SNet-FT-Full)

We developed SNet-FT-Full by comprehensively fine-tuning SpeciesNet on the complete local training set (90% of images, ~2,000 per species). Starting from the same pruned 15-class architecture used for SNet-FT-Small, we fine-tuned higher layers and the classifier head while preserving lower-level features through freezing. This approach maintains global feature representations while specializing for local conditions through extensive local examples. Complete implementation details are provided in Appendix S4.

2.3.4 SpeciesNet used directly on the Wildlife Insights platform

We evaluated the unmodified global SpeciesNet model via the Wildlife Insights (WI) platform as a baseline out-of-the-box classifier. Using the standard WI pipeline (MegaDetector detection followed by SpeciesNet classification), we submitted only the held-out OOD test images with default settings. This approach provides a reproducible baseline but offers no customization for local species, as the model uses its full taxonomy of thousands of species.

Model	Deployment	Base Model	Architecture	Detector used	Adaptation
WI SNet	Online	SpeciesNet	EfficientNet	MegaDetector	None (unmodified)
SNet-FT-Full	Offline	SpeciesNet	EfficientNet	MegaDetector	Fine-tuned (1836 images/class)

SNet-FT-Small	Offline	SpeciesNet	EfficientNet	MegaDetector	Fine-tuned (102 images/class)
Local models	Offline	ImageNet	EfficientNet	MegaDetector	Trained from scratch (1836 images/class)

Table 2: Computer vision workflows for wildlife classification, showing deployment mode, model architecture, and adaptation approach. Deployment indicates model execution via Wildlife Insights (online platform) or personal computation.

2.4 Model evaluation and comparison

2.4.1 Evaluation scenarios

We evaluated all models under both RD and OOD scenarios to assess performance in familiar versus novel conditions. RD tests model performance when deployment conditions match training data, while OOD tests generalization to new locations with different environmental contexts. Both scenarios used the 90/5/5 train/validation/test splits described in Section 2.2 and Table S0, enabling comparison of model robustness under ideal (RD) and challenging (OOD) conditions.

2.4.2 Performance metrics

We evaluated classification performance using precision, recall, and F1-score. The F1-score, as the harmonic mean of precision and recall, provides a balanced metric that accounts for both false positives (Type I errors) and false negatives (Type II errors). This balance is crucial for ecological applications where both misidentifications and missed detections carry significant costs. We report macro-averaged F1-scores in the main text, while complete per-class precision and recall values are provided in Supplementary Materials Table S2. Accuracy was not evaluated, as the MEWC framework outputs precision, recall, and F1-score by design.

Table 3. Performance metrics and their ecological relevance for wildlife classification. We report the F1-score in the main text.

Metric	Focus	Penalizes	Best use case
Accuracy	Overall correctness	None specifically	Balanced datasets
Precision	Correctness of positives	False positives (Type 1 errors)	High cost for false alarms
Recall	Coverage of true positives	False negatives (Type 2 errors)	High cost for missing detections
F1-score	Balance of precision & recall	Both false positives and negatives	Imbalanced datasets or when both matter

2.4.3 Global model evaluation via Wildlife Insights

The global SpeciesNet model was evaluated using the WI with default settings on the OOD test set, providing an out-of-the-box performance baseline. All predictions were compared against ground-truth labels to compute precision, recall, and F1-scores.

2.4.4 Handling model variability and uncertainty

We accounted for variability in model training outcomes to ensure fair comparisons. The global SpeciesNet baseline has fixed pre-trained weights and produces deterministic predictions, while the fine-tuned SpeciesNet model showed minimal run-to-run variation due to stable initialization. In contrast, the locally-trained models exhibited greater variability due to random initialization. To quantify this, we trained multiple local models with different random seeds and report mean performance with standard deviations, providing variance-aware estimates of local model performance.

3. Results or Experimentation

3.1 Random distribution (RD) performance

Under RD conditions with overlapping training and testing camera deployments, the fully fine-tuned global model (SNet-FT-Full) achieved superior performance with a macro-averaged F1-score of 0.964. This exceeded both the local model (0.937 ± 0.038) and the minimally adapted fine-tuned model (SNet-FT-Small, 0.934) (Table 4). SNet-FT-Full demonstrated balanced precision and recall (both 0.964), indicating consistent species identification with minimal false positives or missed detections. This approach exemplifies how publicly available global models can be effectively customized for regional applications using standard transfer learning techniques.

Table 4: CV model performance for random (top) and out-of-distribution (bottom), showing mean values among the 15 species of Australian rainforest wildlife (15 classes). SNet = SpeciesNet, FT=Fine-tuning, 102 and 1836 are the images used in the fine-tuning training. We provide additional stats in the appendix (Tables S3–S5).

Random distribution			
Metric	**SNet-FT-Small (120 images per class)**	**SNet-FT-Full (2030 images per class)**	**Local model (2030 images per class)**
Recall	0.933	0.964	0.937 ± 0.036
Precision	0.934	0.964	0.938 ± 0.047
F1-score	0.934	0.964	0.937 ± 0.038
Out-of-distribution			
Metric	**SNet-FT-Small (102 images per class)**	**SNet-FT-Full (1836 images per class)**	**Local model (1836 images per class)**
Recall	0.790	0.915	0.850 ± 0.199
Precision	0.784	0.925	0.849 ± 0.130
F1-score	0.764	0.909	0.839 ± 0.171

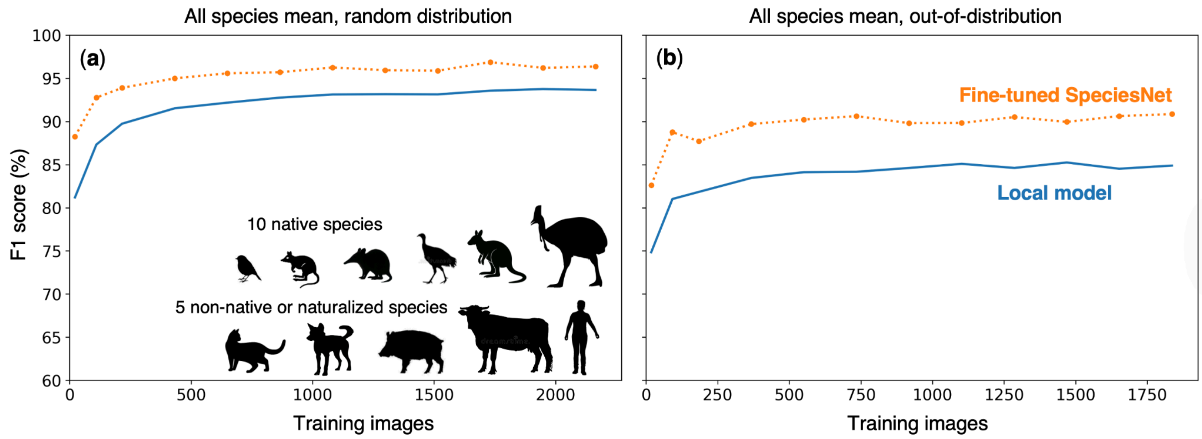


Figure 4. Performance of computer vision species classification for local models versus the fine-tuned SpeciesNet. Testing was repeated for both random distribution (a), wherein images came from the same cameras as those in training datasets, and out-of-distribution (b), in which images came from cameras that were withheld from training datasets. Averaged F1-scores (mean across 15 species) show performance as a function of local training data proportion. Fine-tuned models achieve higher performance than local models, and this advantage was larger when testing on out-of-distribution images. For comparisons with other results in which SpeciesNet was fine-tuned with just 120 images ('SNFT-Small') tuned species net model but truncated to the 120 training images

At the species level (Table S1), fine-tuning yielded large gains for several taxa, particularly *Orthonyx spaldingii*, *Uromys caudimaculatus*, and *Wallabia bicolor*, where the fine-tuned global model markedly outperformed the local model. Conversely, the local model maintained comparable precision for highly distinctive species such as *Bos taurus* and *Casuarius casuarius*, suggesting that visually distinctive taxa can be effectively learned even by locally trained models with limited data.

These results highlight two key insights: first, that global pre-training provides a strong feature foundation that can be effectively specialised to local conditions through fine-tuning, delivering performance superior to purely local training. And second, that local models, while competitive for common or visually distinctive species, struggle with more challenging taxa, likely due to limited feature diversity and training data.

3.2 Out-of-distribution performance

In the OOD evaluation, where test images came from entirely unseen camera deployments, all models exhibited performance declines, but the magnitude of these drops differed markedly between approaches (Table S3). SNet-FT-Full again outperformed all other models, achieving a macro-averaged F1-score of 0.909, compared to 0.839 ± 0.171 for the local model and 0.764 for SNet-FT-Small. This represents a 9-point F1 advantage over the local model and a 15-point advantage over the minimally fine-tuned model, underscoring the value of comprehensive fine-tuning for robust generalisation.

Precision and recall patterns reinforce this finding: SNet-FT-Full maintained high precision (0.929) and recall (0.925), while the local model exhibited greater variability (precision = 0.849 ± 0.130; recall = 0.850 ± 0.199), reflecting inconsistent performance across species. Notably, several rare or morphologically similar taxa (e.g., *Hypsiprymnodon moschatus*, *Thylogale stigmatica*) saw the greatest gains from fine-tuning, suggesting that global features enhanced recognition of difficult species in novel contexts.

Taken together, these results demonstrate that fine-tuning a globally pre-trained model on local imagery provides substantial resilience to domain shifts, delivering superior accuracy and consistency compared to purely local training.

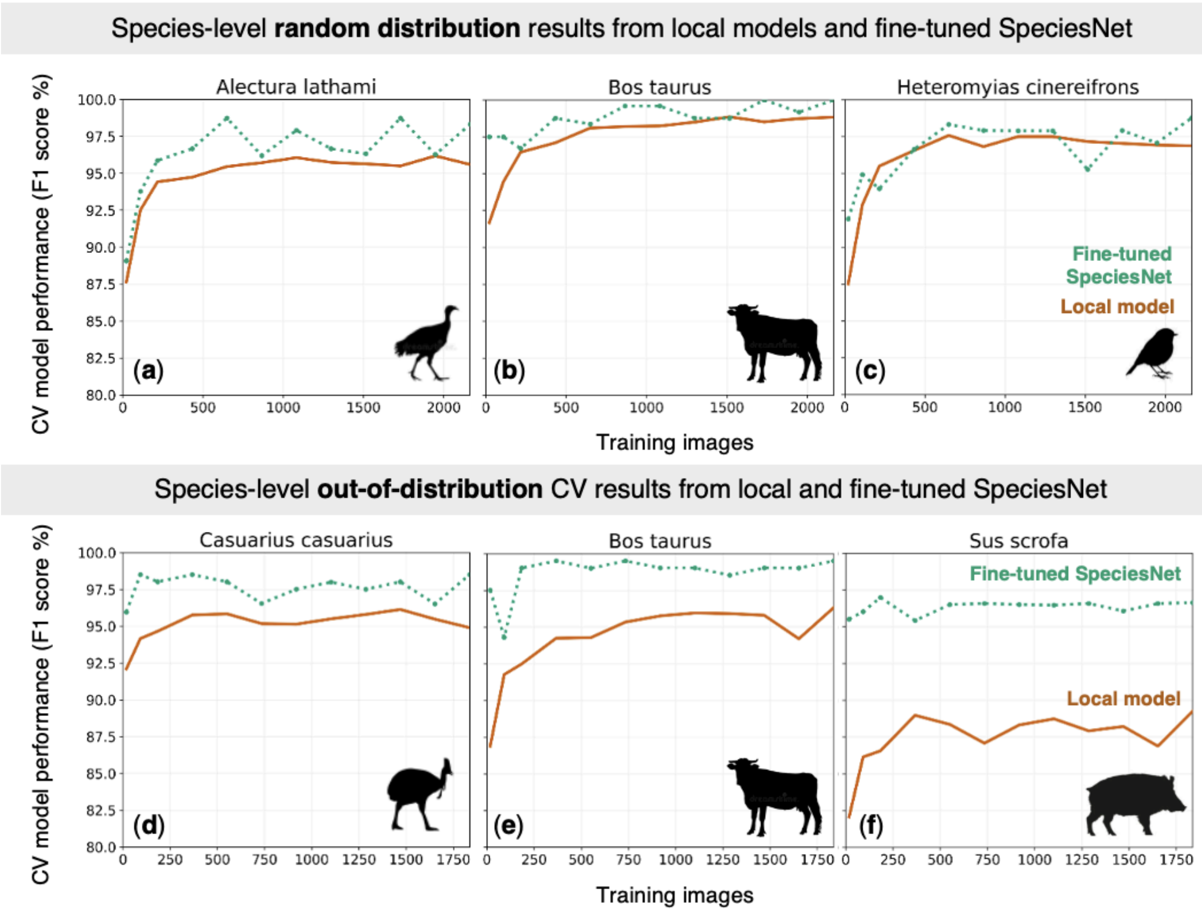


Figure 5: Species-level performance of CV models tested with random distribution (top) and out-of-distribution (bottom). Species chosen here show native and non-native species, birds and mammals, as well as a variety of body sizes (all 15 species are reported in the appendix). *Bos taurus* and *Sus scrofa* were already present in the SpeciesNet classes, while *Casuarius casuarius* and *Alextura lathami*, and *Heteromyias cinereifrons* are novel. The baseline SpeciesNet output for new species is where the green line crosses the y-axis (min value was 20 training images).

4. Discussion

Our results indicate that a fine-tuned global model with 2000 local training images (SNet-FT-Full) achieved the highest classification performance in both random-distribution (RD) and out-of-distribution (OOD) scenarios. The fine-tuned model's advantage was most pronounced under challenging OOD conditions: SNet-FT-Full attained an F1 of 0.909 on images from unseen cameras deployments, versus 0.839 ± 0.171 for the local model. In the RD test, SNet-FT-Full reached a macro

F1-score of 0.964, compared to 0.937 ± 0.038 for the local model and 0.934 for the minimally adapted global model (SNet-FT-Small). Even this 3% gain in the RD results is significant because it approaches or crosses the accuracy thresholds suggested as needing human verification (e.g., some rules-of-thumb are 95%). Notably, SNet-FT-Full excelled on species that proved difficult for the local model, underscoring its robust generalisation across taxa. Further, fine-tuning SpeciesNet performance reaches very high levels with as little as 250 images. These results align with prior findings that deep learning classifiers can achieve reasonable accuracy with as few as several hundred labelled images per class, provided they represent the range of local variation in appearance, context, and camera configurations (Raghu et al., 2019; Torralba & Efros, 2011). Taken together, fine-tuning SpeciesNet consistently yielded better outcomes than training a new model from scratch.

The fine-tuned global model likely achieves its edge by effectively integrating broad learned features (e.g., key places to search for unique differences in body shapes) and local ecological features (e.g., the body shape of southern cassowaries, *Casuarius casuarius*). It also removes thousands of irrelevant classes present in the global model species list, while introducing new local species. The result was a single, streamlined classifier that generalised well to novel conditions (as seen in the OOD test) while remaining focused on the region's species. This synergy suggests that global classifier models like SpeciesNet can serve as powerful baselines that, with minimal retraining, adapt to new ecological contexts far more effectively than either approach alone. The benefits were pronounced for hard-to-classify species such as small birds (*Orthonyx spaldingii*), rodents (*Uromys caudimaculatus)* and cryptic species such as the two macroods (*Wallabia bicolor* and *Thylogale stigmatica*). In contrast, the local model remained competitive on highly distinctive species (e.g. cattle, *Bos taurus*), suggesting that visually obvious taxa can be learned adequately even with limited data. However, the local models exhibited higher variance and degraded generalisation compared to the fine-tuned model. For example, across multiple training runs, the local model's performance fluctuated substantially (e.g. ±0.14 F1 in OOD tests), indicating sensitivity to training data variability. Meanwhile, the off-the-shelf global classifier (SpeciesNet via Wildlife Insights) often misidentified species outside its original training set, resulting in a poor OOD baseline (macro-F1 ≈ 0.1329).

A consistent source of error was the poor performance due to training images with multiple species, such *Homo sapiens* walking domestic dogs (*Canis familiaris*). This co-occurrence produces ambiguous snips and overlapping detections that inflate off-diagonal entries in the confusion matrix (Fig. S4), with many *Homo sapiens* frames assigned as dogs and vice versa. This was especially problematic with snips were dominated by the dog rather than the person and vice versa (Fig. 6). Our training was single-label by design, so these multi-object scenes were scored as errors, even when the model correctly recognises one of the co-occurring taxon. This pattern is consistent with context-induced confusions reported in camera classification (Beery et al., 2020), and suggests practical mitigations: (i) sequence-level aggregation, (ii) multi-label scoring for human–animal co-occurrence, and (iii) detector settings that favour separate, non-overlapping crops. The presence of dingoes (*Canis lupus dingo*; also called wild dogs *Canis familiaris dingo*) likely also added to misidentifications.

Figure 6: Errors often resulted from training images with multiple species. This snip is a representative frame of when a human-dog classification error occurred (the handler's leg is visible in front of the dog). This is a key limitation of the single-label evaluation framework when faced with multi-taxa occurrences.

A key practical question is how many images are needed to train a reliable local model and for fine-tuning SpeciesNet. We found that model accuracy improves rapidly with the first 20 images per class but plateaus beyond a certain threshold, often ~250 images, or even diminishing. This suggests that several hundred diverse, high-quality images per species may be sufficient (or even ideal) to approach the upper bound of local model or fine-tuning performance in this use case. Such inferences align with other deep learning models reports that achieve reasonable accuracy with as few as a few hundred labelled examples per category (Raghu et al., 2019; Torralba & Efros, 2011). This is an encouraging result – it implies that users do not necessarily need thousands of images for each species to build an effective classifier. Instead, there should be a focus on obtaining a few hundred representative images (capturing the range of individual variation, backgrounds and camera angles) for each target species. We note, however, that this threshold will vary with the number of species in the classifier and other forms of complexity; more images might be required for cryptic species that appear similar.

There are a number of practical considerations for using local and global models without fine-tuning. A practical downside to developing a local model from scratch is that it requires higher effort in data collection and annotation, as well as AI/ML expertise, and computational resources. Local models are most suitable when new cameras are placed in the same locations as the training dataset (similar to our random distribution results). On the contrary, their opposite practical advantages apply to using pre-trained global models like SpeciesNet, especially on online platforms with intuitive GUI like Wildlife Insights. If the focal site and study system are captured in SpeciesNet training, this is likely the best option. One downside to global models is when they function as plug-in-play black boxes with limited flexibility. For example, users cannot easily add new species or otherwise edit the workflow on Wildlife Insights and the lack of customization can yield unexpected outputs (e.g. predicting a species that is biogeographically implausible at the site). This can hinder interpretability and confidence.

There are also practical considerations for fine-tuning a global model. An advantage is that it is computationally efficient (since the pre-trained backbone is reused) and data-efficient (requiring fewer images to reach high performance than training from scratch). Further refining the classes based on the focal species geographical distribution ensures its outputs were both accurate and context-appropriate, improving interpretability and trust. However, this approach does require some technical capability, though we attempt to mitigate this by providing open-source code and guidelines (see Appendix). Moreover, fine-tuning still relies on a moderate amount of local training data and might need periodic updates if environmental conditions or the species community change over time. In our case, fine-tuning SpeciesNet achieved excellent species recognition for an under-represented ecosystem.

5. Conclusion

This study provides empirical guidance for ecologists using computer vision models for wildlife image classification. Fine-tuning a pre-trained global model (SpeciesNEt) emerged as a highly effective and efficient strategy, esseically a hybrid approach that reconciles the trade-off between local and global approaches. The fine-tuned model not only outperformed a local model in both familiar and novel contexts, but did so with relatively fewer training images, illustrating the power of transfer learning for ecological applications. For other practitioners, this suggests that fine-tuning global models may be the optimal path forward when faced with limited training data for a new region or species of interest. Looking ahead, we envision a growing role for global classifier models in ecology and conservation, as we demonstrated how a global model can be re-purposed to suit many different locales and taxa. This light-weight fine-tuning accelerates the development of accurate classifiers for biodiversity monitoring. This means that ecologists worldwide can harness state-of-the-art classifiers (trained on vast global data) and tweak them for their own projects without needing prohibitive data volumes or computing power, and we provide the code for users to do this.

Data availability:

Data, code and materials availability: https://doi.org/10.15468/kjqw3f (for data), https://github.com/WildObs/SpeciesNet-FineTuning (for code) and wildobs.org.au (for working on the trained model).

Appendix / Supplementary Materials:

Appendix S1:

The discrepancy in total images between the RD and OOD datasets arises from the stricter constraints imposed on OOD. In RD we sampled frames at random across all cameras, allowing an exact 90/5/5 split while maintaining per-species balance. By contrast, OOD requires that entire camera deployments be assigned to a single set (train (or) test (or) valid) to prevent spatial–temporal leakage and to evaluate the model generalizability. As a result, we prioritised disjoint deployments and balanced classes over exact initial proportions, yielding fewer total images in OOD than in RD.

Appendix S2:

Local Model Training Specifications

We implemented a site-specific local model using the MEWC framework's mewc-train step following mewc-detect and mewc-snip workflows. The architecture used an EfficientNetV2-M backbone initialized with ImageNet weights and a newly added 15-class softmax output layer. Optimization employed AdamW (weight decay 10^{-4}) with initial learning rate 10^{-4} (exponentially decayed) and batch size 32. We applied three RandAug transformations per mini-batch for augmentation. Training proceeded in two phases: Stage 1 froze all convolutional layers for 10 epochs to learn basic classifier separations; Stage 2 unfroze the top EfficientNet bottleneck block and ran four consecutive 7-epoch cycles, progressively increasing RandAug severity (0.2 to 0.8) and dropout rate (0.1 to 0.4) per cycle. This curriculum enabled gradual regularization while adapting higher-level features to local patterns.

For performance scaling analysis, we trained models on class-balanced subsets containing $p \in \{1, 5, 10, 20, 30, ..., 100\}$% of available training images per species. Each p used different random seeds and initializations, yielding 60 models per split condition. Under RD splits (2164/120/121 images per class for train/val/test) and OOD splits (1836/102/102 images per class), this produced 120 total models for variance-aware performance estimation.

All training used the RCCs Nectar HPC cluster with NVIDIA A100-PCIE-40GB GPU (CUDA 12.9, Driver 575.51.03), AMD EPYC-Rome CPUs, Conda environment (Python 3.12.4, TensorFlow 2.16.1, JAX 0.6.0, PyTorch 2.7.0 + cu126) on Ubuntu 20.04 (kernel 5.15.0-140-generic). All configurations are readily available in MEWC, significantly reducing development overhead.

Appendix S3:

We fine-tuned SpeciesNet using a small local dataset (the validation subset, ~5% of images per species) to create SNet-FT-Small. The model modification involved replacing SpeciesNet's original classification layer (>2,000 outputs) with a new 15-class layer. For the 9 species overlapping with SpeciesNet's original training, we transferred corresponding final-layer weights; for the 6 novel species, we initialized new output nodes randomly. This pruning approach restricts predictions to local taxa while retaining relevant prior knowledge.

Fine-tuning updated the new 15-class output layer and select high-level convolutional layers of the EfficientNet backbone, with lower-level features remaining frozen. We used Adam optimization with a small learning rate and data augmentation to prevent overfitting on the limited dataset. Training proceeded for 20 epochs without multiple random restarts, as the pre-trained weights provided stable initialization. This custom implementation was necessary as the Wildlife Insights platform and MEWC pipeline lack class pruning functionality for fine-tuning.

Appendix 4:

We performed comprehensive fine-tuning of SpeciesNet using the entire training split (90% of images, approximately 2,000 per species) to create SNet-FT-Full. The model initialization mirrored SNet-FT-Small, beginning with the pruned 15-class SpeciesNet architecture. During fine-tuning, we froze lower layers of the EfficientNet backbone to preserve general visual features while updating higher layers and the new classifier head to adapt to local characteristics.

Training employed RandAug data augmentation and regularization to capture environmental variation and reduce camera-specific bias. Optimization used the same settings as SNet-FT-Small. This implementation required custom code using the open-source code (available here), as the Wildlife Insights platform and SpeciesNet repository does not support the necessary class pruning and weight initialization for novel species. The resulting model integrates global feature knowledge with local specialization, achieving site-specific focus while leveraging pretrained representations.

Appendix 5:

We assessed the baseline performance of the unmodified SpeciesNet model through the Wildlife Insights online platform. This black-box global classifier, pre-trained on >65 million images covering >2,000 species, was applied using WI's standard pipeline: MegaDetector for animal detection followed by SpeciesNet for classification. No fine-tuning or customization was performed.

Evaluation used only the out-of-distribution test set (downsampled and balanced-OOD) with default confidence thresholds. The random distribution set was not evaluated through WI since SpeciesNet's training never included AWT imagery, making both splits effectively out-of-distribution. The WI interface prevented class pruning, parameter adjustment, or restriction to our 15 target species, resulting in predictions across its full taxonomy. This approach represents a minimal-effort, reproducible baseline but lacks flexibility for research customization.

Dataset	**Train (90%)**	**Validation (5%)**	**Test (5%)**
RD split	32,460 images 2,030 images/species All cameras Used for: Local model, SNet-FT-Full	1,815 images 120 images/species All cameras Used for: SNet-FT-Small	1,815 images 120 images/species All cameras
OOD split	27,540 images 1,836 images/species Separate training cameras Used for: Local model, SNet-FT-Full	1,530 images 102 images/species Unique validation cameras Used for: SNet-FT-Small	1,530 images 102 images/species Separate test cameras

Table S0: Images in each of the model trainings and testing. We used the validation set for finetuning since SpeciesNet already has suitable hyperparameters.

Table S1: F1-score of three wildlife computer vision models for RD scenario.

Species	SNet-FT-Small (F1 score)	SNet-FT-Full (F1 score)	Local Model (F1 score)
Alectura lathami	0.9465	**0.9832**	0.946
Bos taurus	0.9794	**1.0000**	0.992
Canis lupus dingo or canis familiaris*	0.9113	**0.9198**	0.914
Casuarius casuarius	0.9630	**0.9959**	0.971
Felis catus	0.9583	**0.9750**	0.971
Heteromyias cinereifrons	0.9442	**0.9876**	0.958
Homo sapiens	0.8408	**0.9048**	0.876
Hypsiprymnodon moschatus	0.9412	**0.9593**	0.918
Megapodius reinwardt	0.9540	**0.9754**	0.955
Orthonyx spaldingii	0.9317	**0.9752**	0.966
Perameles nasuta	0.9113	**0.9398**	0.892
Sus scrofa	0.9540	**0.9752**	0.953
Thylogale stigmatica	0.9205	**0.9620**	0.892
Uromys caudimaculatus	0.9016	**0.9244**	0.886
Wallabia bicolor	0.9456	**0.9791**	0.971

Table S2: F1-score of three wildlife computer vision models for OOD

Species	SNet-FT-Small	SNet-FT-Full	Local model
Alectura lathami	0.890717	**0.965275**	0.921 ± 0.022
Bos taurus	0.883208	**0.985850**	0.941 ± 0.029
Canis familiaris	0.625167	**0.710392**	0.680 ± 0.025
Casuarius casuarius	0.924567	**0.976367**	0.951 ± 0.013

Felis catus	0.838817	**0.968558**	0.907 ± 0.054
Heteromyias cinereifrons	0.926933	**0.940225**	0.908 ± 0.018
Homo sapiens	0.348050	**0.423567**	0.388 ± 0.048
Hypsiprymnodon moschatus	0.791300	**0.938833**	0.855 ± 0.045
Megapodius reinwardt	0.882825	**0.965550**	0.948 ± 0.018
Orthonyx spaldingii	0.769833	**0.896367**	0.810 ± 0.036
Perameles nasuta	0.676683	**0.904833**	0.830 ± 0.034
Sus scrofa	0.790467	**0.963342**	0.874 ± 0.024
Thylogale stigmatica	0.500450	**0.902008**	0.762 ± 0.051
Uromys caudimaculatus	0.769975	**0.893950**	0.816 ± 0.043
Wallabia bicolor	0.137250	**0.955050**	0.891 ± 0.083

Table S3: Random distribution summary statistics of all computer vision models assessing 30,000 images of 15 species of Australian rainforest wildlife (15 classes). The values are the mean performance among the 15 species. The main text only presents F1-score, while these tables include Recall and Precision

Metric	**Fine-tuned global model (val)**	**Fine-tuned global model (train)**	**Local model trained in MEWC**
Recall	0.910	0.965	0.913 ± 0.047
Precision	0.917	0.966	0.914 ± 0.049
F1-score	0.910	0.965	0.913 ± 0.046

Table S4: Precision of three wildlife computer vision models:

Species	**Precision of fine-tuned global model (val)**	**Precision of local model**	**Precision of fine-tuned global model (train)**
Alectura_lathami	0.9120	0.957	**0.9829**
Bos_taurus	0.9677	0.972	**1.0000**
Canis_familiaris	0.8682	0.88	**0.9237**
Casuarius Casuarius	0.9583	0.961	**1.0000**
Felis Catus	0.9655	0.933	**0.9833**
Heteromys_cinereifrons	**1.0000**	0.977	0.9915

Homo_sapiens	0.8468	0.857	**0.8712**
Hypsiprymnodon_moschatus	**0.9421**	0.870	0.9355
Megapodius_reinwardt	0.9746	0.957	**0.9916**
Orthonyx_spaldingii	0.7763	0.972	**0.9752**
Perameles_nasuta	**0.9691**	0.880	0.9302
Sus_scrofa	0.9508	**0.961**	0.9593
Thylogale_stigmatica	0.9320	0.933	**0.9914**
Uromys_caudimaculatus	0.8156	**0.977**	0.9661
Wallabia_bicolor	0.8731	0.857	**0.9916**

Table S5: Out-of-Distribution Performance Comparison for fine-tuned global model (SpeciesNet). Here SpeciesNet is fine-tuned only on the smaller training set.

Species	**Recall**	**Precision**	**F1-score**
Alectura_lathami	0.9216	0.9592	0.9400
Bos_taurus	0.9510	0.7348	0.8291
Canis_familiaris	0.6863	0.5469	0.6087
Casuarius_casuarius	0.9412	0.9796	0.9600
Felis_catus	0.9804	0.9804	0.9804
Heteromyias_cinereifrons	0.9412	0.9897	0.9648
Homo_sapiens	0.2157	0.5641	0.3121
Hypsiprymnodon_moschatus	0.7941	0.8710	0.8308
Megapodius_reinwardt	0.9706	0.9000	0.9340
Orthonyx_spaldingii	0.9216	0.8034	0.8584
Perameles_nasuta	0.9510	0.6831	0.7951
Sus_scrofa	0.9314	0.7252	0.8155
Thylogale_stigmatica	0.6176	0.4315	0.5081
Uromys_caudimaculatus	0.7745	0.8977	0.8316
Wallabia_bicolor	0.0686	0.7778	0.1261

Table S6: Out-of-Distribution Performance Comparison for Local model. Here, the Local (MEWC architecture) model is trained on the entire training set of AWT dataset.

Species	**F1-score (Mean ± Std)**	**Precision (Mean ± Std)**	**Recall (Mean ± Std)**
Alectura_lathami	0.921 ± 0.022	0.953 ± 0.024	0.892 ± 0.033
Bos_taurus	0.941 ± 0.029	0.964 ± 0.021	0.920 ± 0.046
Canis_familiaris	0.680 ± 0.025	0.568 ± 0.022	0.850 ± 0.038
Casuarius_casuarius	0.951 ± 0.013	0.983 ± 0.021	0.921 ± 0.015
Felis_catus	0.907 ± 0.054	0.877 ± 0.075	0.941 ± 0.035
Heteromyias_cinereifrons	0.908 ± 0.018	0.915 ± 0.026	0.902 ± 0.028
Homo_sapiens	0.388 ± 0.048	0.601 ± 0.060	0.288 ± 0.046

Hypsiprymnodon_moschatus	0.855 ± 0.045	0.872 ± 0.046	0.841 ± 0.058
Megapodius_reinwardt	0.948 ± 0.018	0.920 ± 0.029	0.977 ± 0.014
Orthonyx_spaldingii	0.810 ± 0.036	0.787 ± 0.061	0.837 ± 0.024
Perameles_nasuta	0.830 ± 0.034	0.753 ± 0.046	0.929 ± 0.047
Sus_scrofa	0.874 ± 0.024	0.868 ± 0.037	0.881 ± 0.024
Thylogale_stigmatica	0.762 ± 0.051	0.862 ± 0.050	0.686 ± 0.065
Uromys_caudimaculatus	0.816 ± 0.043	0.831 ± 0.048	0.804 ± 0.059
Wallabia_bicolor	0.891 ± 0.083	0.893 ± 0.060	0.894 ± 0.111

Table S7: Out-of-Distribution Performance Comparison for fine-tuned global model (train). Here, SpeciesNet is fine-tuned on the full training set of AWT dataset.

Species	Recall	Precision	F1-score
Alectura_lathami	0.9608	1.0000	0.9800
Bos_taurus	0.9804	0.9901	0.9852
Canis_familiaris	0.9804	0.5848	0.7326
Casuarius_casuarius	0.9510	1.0000	0.9749
Felis_catus	0.9902	0.9528	0.9712
Heteromyias_cinereifrons	0.9608	0.9074	0.9333
Homo_sapiens	0.2745	0.9032	0.4211
Hypsiprymnodon_moschatus	0.9412	0.9796	0.9600
Megapodius_reinwardt	1.0000	0.9533	0.9761
Orthonyx_spaldingii	0.8824	0.9574	0.9184
Perameles_nasuta	0.9608	0.9245	0.9423
Sus_scrofa	0.9902	0.9712	0.9806
Thylogale_stigmatica	0.9510	0.9065	0.9282
Uromys_caudimaculatus	0.9412	0.9057	0.9231
Wallabia_bicolor	0.9412	1.0000	0.9697

Experimental results (single run) for random distribution of different species

Add a bad image and then add the combined plots

Species	Precision	Recall	F1-score
Wallabia bicolor	1.0000	0.3039	0.4662
Sus scrofa	0.9870	0.7451	0.8492

Megapodius reinwardt	0.0000	0.0000	0.0000
Alectura lathami	0.9429	0.3235	0.4818
Hypsiprymnodon moschatus	0.0000	0.0000	0.0000
Homo sapiens	0.5294	0.0882	0.1513
Heteromyias cinereifrons	0.0000	0.0000	0.0000
Orthonyx spaldingii	0.0000	0.0000	0.0000
Felis catus	0.9881	0.8137	0.8925
Casuarius casuarius	0.0000	0.0000	0.0000
Canis familiaris	0.5783	0.4706	0.5189
Thylogale stigmatica	1.0000	0.1961	0.3279
Bos taurus	0.9691	0.9216	0.9447
Uromys caudimaculatus	0.0000	0.0000	0.0000
Perameles nasuta	1.0000	0.0098	0.0194

Table S reports per-species precision, recall, F1-score, and sample size for the Wildlife Insights (WI) online classifier when applied to the out-of-distribution downsampled and balanced-OOD test set. The table distinguishes species that were represented in WI's training taxonomy from those that were non-represented, the latter receiving no correct predictions (all metrics = 0). Among represented taxa, performance was highly variable: some common or distinctive species (e.g. Bos taurus, Felis catus, Sus scrofa) achieved high precision and recall, whereas many locally important taxa (e.g. Megapodius reinwardt, Hypsiprymnodon moschatus, Uromys caudimaculatus) were entirely missed. This pattern highlights the limitations of an unadapted global model in recognising region-specific fauna, even when some species are nominally included in its label set, and underscores the need for local adaptation or fine-tuning to improve coverage and consistency.


RAWT Image Dataset Distribution
Species
Thylogale stigmatica
Alectura lathami
Sus scrofa
Megapodius reinwardt
Hypsiprymnodon moschatus
Perameles nasuta
Bos taurus
Uromys caudimaculatus
Homo sapiens
Orthonyx spaldingii
Heteromyias cinereifrons
Felis catus
Wallabia bicolor
Casuarius casuarius
Canis familiaris
Isoodon macrourus
Chalcophaps longirostris
Pitta versicolor
Aepyprymnus rufescens
Melloria quoyi
Trichosurus vulpecula
Macropus giganteus
Chalcophaps indica
Tachyglossus aculeatus
Ptilonorhynchus violaceus
Psophodes olivaceus
Macropus parryi
Rallina tricolor
Ailuroedus melanotis
Burhinus grallarius
Macropus agilis
Meleagris gallopavo
Heteromyias albispecularis
Gymnorhina tibicen
Hydromys chrysogaster
Columba leucomela
Varanus varius
Dendrolagus lumholtzi
Strepera graculina
Dasyurus maculatus
Ailuroedus crassirostris
Sericornis citreogularis
Dendrolagus bennettianus
Zoothera lunulata
Intellagama lesueurii
Phaps chalcoptera
Vulpes vulpes
Rhipidura rufifrons
Dromaius novaehollandiae
Leucosarcia melanoleuca
Colluricincla harmonica
Centropus phasianinus
Thylogale thetis
Sericornis frontalis
Rhinella marina
Macropus eugenii
Colluricincla boweri
Rhipidura leucophrys
Sericornis keri
Macropus rufogriseus
Pitta sordida
Gallus gallus
Rattus fuscipes
Dacelo novaeguineae
Equus caballus
Oryctolagus cuniculus
Lophorina victoriae
Geophaps scripta
Menura novaehollandiae
Macropygia phasianella
Macropus fuliginosus
Tanysiptera sylvia
Paramelomys naso
Platycercus elegans
Pseudechis porphyriacus
Cacatua galerita
Dactylopsila trivirgata
Oreoscopus gutturalis
Podargus strigoides
Platycercus adscitus
Pachycephala rufiventris
Colluricincla megarhyncha
Pitta novaeguineae
Talegalla cuvieri
Pseudocheirus peregrinus
Rhipidura phoenicura
Pseudomys gracilicaudatus
Pachycephala pectoralis
Ptilinopus cinctus
Vombatus ursinus
Scenopoeetes dentirostris
Sturnus vulgaris
Pseudochirops archeri
Potorous tridactylus
Antechinus godmani
Rhipidura albiscapa
Turnix varius
Varanus rosenbergi
Lophophanes cristatus
Acanthorhynchus tenuirostris
Pachycephala simplex
Philemon corniculatus
Neochmia temporalis
Petaurus breviceps
Mus musculus
Xanthotis macleayanus
Amaurornis moluccana
Tregellasia capito
Rusa unicolor
Corvus orru
Pavo cristatus
Trichosurus caninus
Entomyzon cyanotis
Pica pica
Acrobates pygmaeus
Monarcha melanopsis
10^0
10^1
10^2
10^3
10^4
10^5
10^6
Number of images

Figure [S1]: Class frequencies (all species) in the complete camera corpus (horizontal lines; log-scaled x-axis), showing a strongly long-tailed distribution. Blue points mark per-species totals; the vertical dashed guide indicates the per-species threshold used to define the downsampled and balanced set (2,164 images).

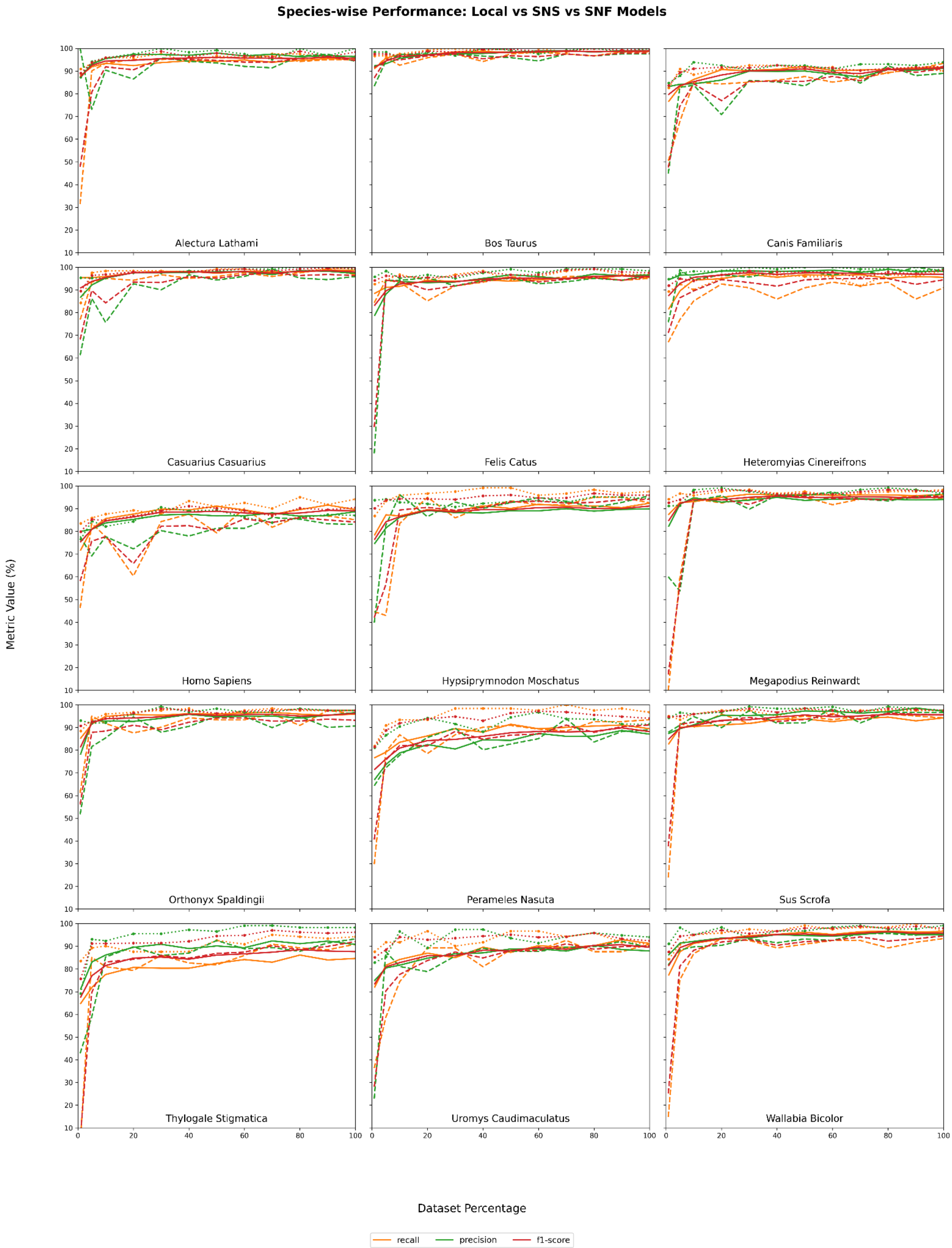


Figure S2: In distribution results for all performance metrics. The x-axis is the percentage of the total 2000 image dataset that was used for the training.

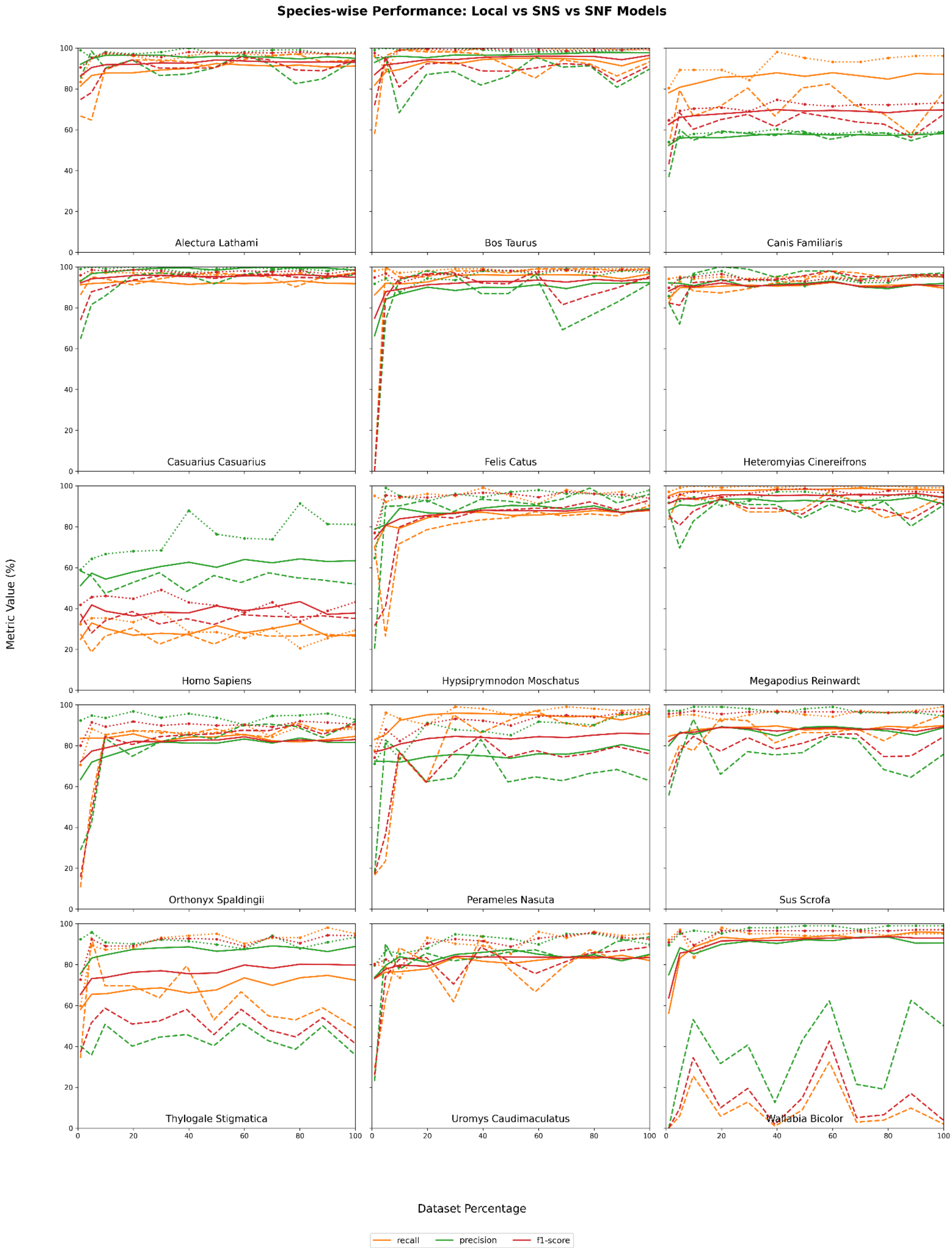
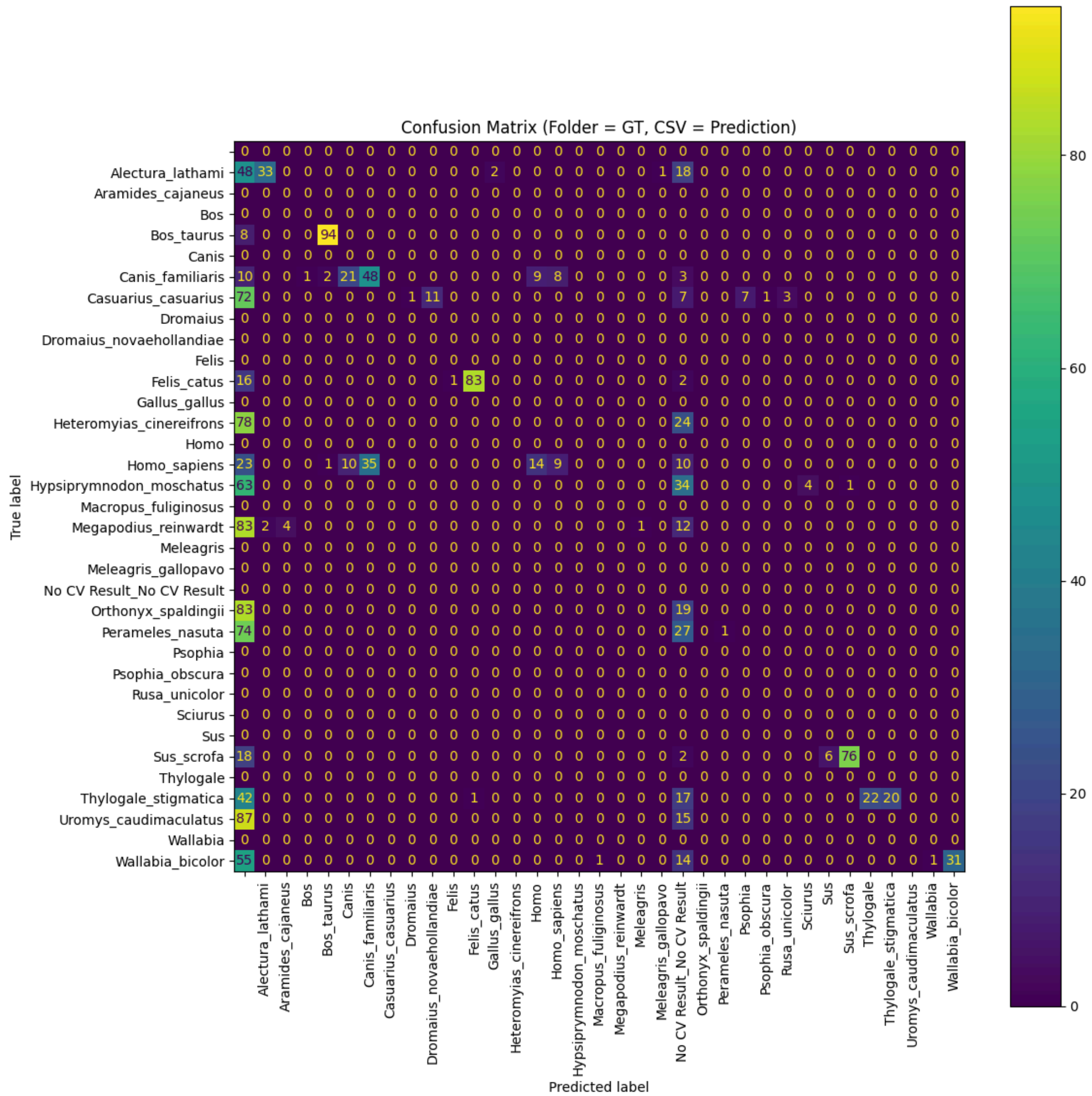


Figure S3: Out-Of-Distribution performance, noting the main text only present F1 score, so here we present the same results but with more performance metrics. The x-axis is the percentage of the total 2000 image dataset that was used for the training.



Accuracy: 25.82%, Macro-F1: 0.1329 Total: 1530 images

Figure S4. Confusion matrix with the full Wildlife Insights label space (unrestricted labels). The same downsampled and balanced-OOD images (N = 1,530) are evaluated with an unrestricted prediction vocabulary, so columns include genus-only labels, genus + species labels, and classes outside the 15-species set (e.g. additional WI taxa and "No CV Result"). Off-diagonal counts therefore include off-list assignments arising from taxonomy mismatches rather than only within-set confusions. Values are counts (not normalised). This matrix illustrates how an unadapted global classifier disperses predictions across non-target taxa, explaining reduced performance when labels are not harmonised to the 15-species evaluation set.

Removing Blanks & unknown (species):

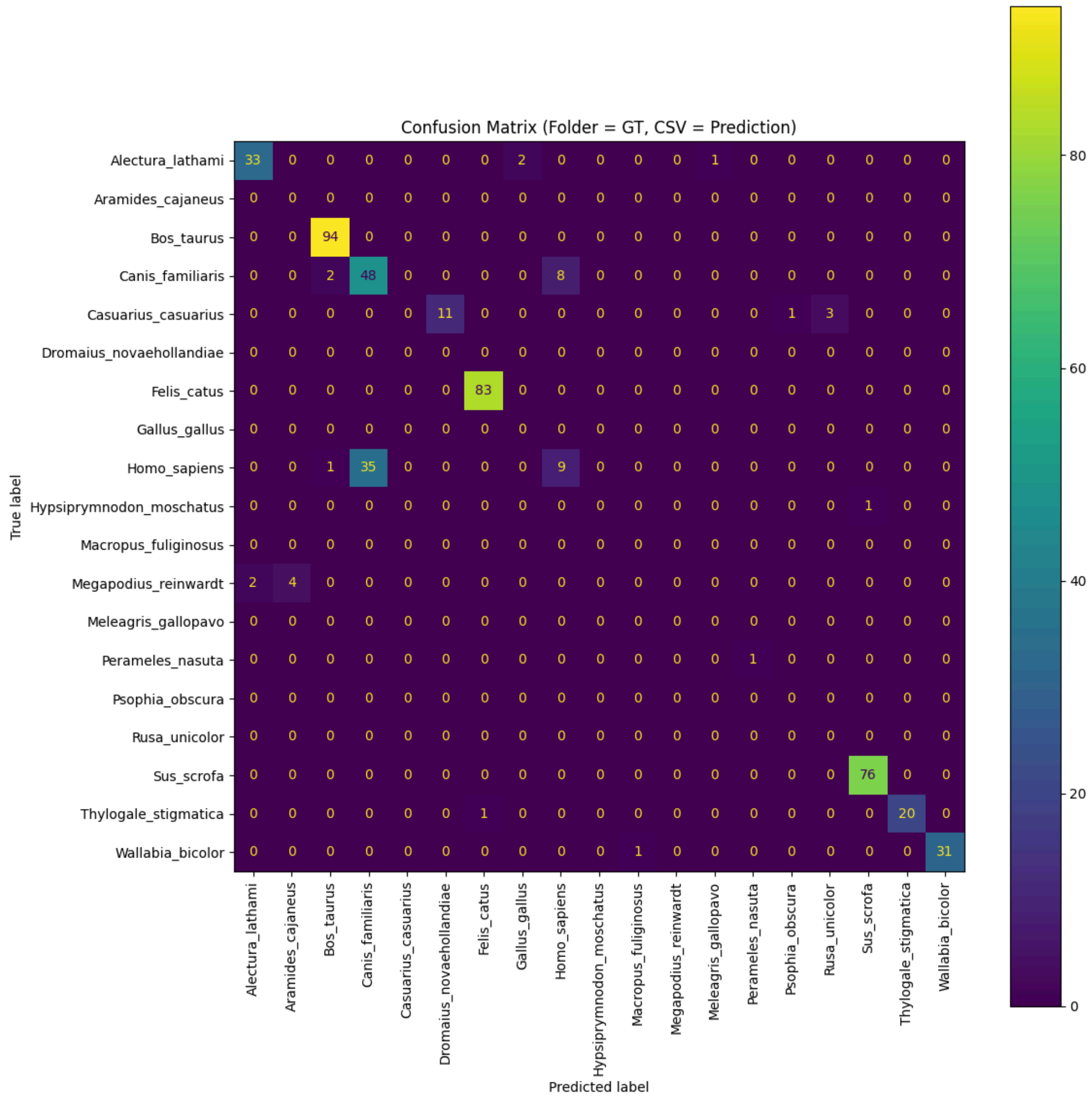


Figure SX. Confusion matrix for the 15-species benchmark (restricted label space). Rows show ground-truth labels and columns model predictions for the downsampled and balanced-OOD test set (N = 1,530; 102 images per species). The model's label space is restricted to the 15 focal species, so off-diagonal entries reflect true species-level confusions within the target community. Cell values are counts (not normalised); colour intensity indicates frequency.

Table 2: Random Distribution Performance Comparison for fine-tuned SpeciesNet (global model)

Species	Recall	Precision	F1-score
Alectura_lathami	0.9421	0.9120	0.9268
Bos_taurus	0.9917	0.9677	0.9796
Canis_familiaris	0.9256	0.8682	0.8960
Casuarius Casuarius	0.9504	0.9583	0.9544
Felis Catus	0.9256	0.9655	0.9451
Heteromyias_cinereifrons	0.7355	1.0000	0.8476
Homo_sapiens	0.8678	0.8468	0.8571
Hypsiprymnodon_moschatus	0.9421	0.9421	0.9421
Megapodius_reinwardt	0.9504	0.9746	0.9623
Orthonyx_spaldingii	0.9752	0.7763	0.8645
Perameles_nasuta	0.7769	0.9691	0.8624
Sus_scrofa	0.9587	0.9508	0.9547
Thylogale_stigmatica	0.7934	0.9320	0.8571
Uromys_caudimaculatus	0.9504	0.8156	0.8779
Wallabia_bicolor	0.9669	0.8731	0.9176

Table 3: Random Distribution Performance Comparison for local model

Species	**F1-score (Mean ± Std)**	**Precision (Mean ± Std)**	**Recall (Mean ± Std)**
Alectura_lathami	0.946 ± 0.025	0.957 ± 0.032	0.935 ± 0.021
Bos_taurus	0.973 ± 0.022	0.972 ± 0.025	0.974 ± 0.023
Canis_familiaris	0.883 ± 0.036	0.880 ± 0.033	0.887 ± 0.047
Casuarius_casuarius	0.967 ± 0.023	0.961 ± 0.035	0.974 ± 0.013
Felis_catus	0.934 ± 0.037	0.933 ± 0.053	0.936 ± 0.024
Heteromyias_cinereifrons	0.958 ± 0.029	0.977 ± 0.019	0.941 ± 0.044
Homo_sapiens	0.863 ± 0.042	0.857 ± 0.036	0.870 ± 0.057
Hypsiprymnodon_moschatus	0.880 ± 0.042	0.870 ± 0.048	0.891 ± 0.041
Megapodius_reinwardt	0.938 ± 0.030	0.929 ± 0.037	0.947 ± 0.027
Orthonyx_spaldingii	0.938 ± 0.040	0.931 ± 0.051	0.946 ± 0.034
Perameles_nasuta	0.845 ± 0.056	0.822 ± 0.067	0.871 ± 0.056
Sus_scrofa	0.932 ± 0.033	0.948 ± 0.039	0.918 ± 0.035
Thylogale_stigmatica	0.837 ± 0.059	0.880 ± 0.062	0.800 ± 0.063
Uromys_caudimaculatus	0.861 ± 0.050	0.857 ± 0.049	0.867 ± 0.058
Wallabia_bicolor	0.933 ± 0.041	0.936 ± 0.031	0.930 ± 0.055

Table 4: Random Distribution Performance Comparison for SpeciesNet + local model

Species	Recall	Precision	F1-score
Alectura_lathami	0.9504	0.9829	0.9664
Bos_taurus	0.9917	1.0000	0.9959
Canis_familiaris	0.9008	0.9237	0.9121
Casuarius_casuarius	1.0000	1.0000	1.0000
Felis_catus	0.9752	0.9833	0.9793
Heteromyias_cinereifrons	0.9669	0.9915	0.9791
Homo_sapiens	0.9504	0.8712	0.9091
Hypsiprymnodon_moschatus	0.9587	0.9355	0.9469
Megapodius_reinwardt	0.9752	0.9916	0.9833
Orthonyx_spaldingii	0.9752	0.9752	0.9752
Perameles_nasuta	0.9917	0.9302	0.9600
Sus_scrofa	0.9752	0.9593	0.9672
Thylogale_stigmatica	0.9504	0.9914	0.9705
Uromys_caudimaculatus	0.9421	0.9661	0.9540
Wallabia_bicolor	0.9752	0.9916	0.9833

References

Ahumada, J. A., Fegraus, E., Birch, T., Flores, N., Kays, R., O'Brien, T. G., Palmer, J., Schuttler, S., Zhao, J. Y., Jetz, W., & Kinnaird, M. (2020). Wildlife insights: A platform to maximize the potential of camera trap and other passive sensor wildlife data for the planet. *Environmental Conservation*, *47*(1), 1–6. https://doi.org/10.1017/s0376892919000298

Anderson, S. E., Amir, Z., Bruce, T., & Luskin, M. S. (2025). Range-wide camera trapping for the Australian cassowary reveals habitat associations with rainfall and forest quality. *Ecology and Evolution*, *15*(6), e71464. https://doi.org/10.1002/ece3.71464

Beery, S., Cole, E., & Gjoka, A. (2020). The iWildCam 2020 Competition Dataset. In *arXiv [cs.CV]*. arXiv. http://arxiv.org/abs/2004.10340

Besson, M., Alison, J., Bjerge, K., Gorochowski, T. E., Høye, T. T., Jucker, T., Mann, H. M. R., & Clements, C. F. (2022). Towards the fully automated monitoring of ecological communities. *Ecology Letters*, *25*(12), 2753–2775. https://doi.org/10.1111/ele.14123

Bruce, T., Amir, Z., Allen, B. L., Alting, B. F., Amos, M., Augusteyn, J., Ballard, G.-A., Behrendorff, L. M., Bell, K., Bengsen, A. J., Bennett, A., Benshemesh, J. S., Bentley, J., Blackmore, C. J., Boscarino-Gaetano, R., Bourke, L. A., Brewster, R., Brook, B. W., Broughton, C., … Luskin, M. S. (2025). Large-scale and long-term wildlife research and monitoring using camera traps: a continental synthesis. *Biological Reviews of the Cambridge Philosophical Society*, *100*(2), 530–555. https://doi.org/10.1111/brv.13152

Celis, G., Ungar, P., Sokolov, A., Sokolova, N., Böhner, H., Liu, D., Gilg, O., Fufachev, I., Pokrovskaya, O., Ims, R. A., Zhou, W., Morris, D., & Ehrich, D. (2024). A versatile, semi-automated image analysis workflow for time-lapse camera trap image classification. *Ecological Informatics*, *81*(102578), 102578. https://doi.org/10.1016/j.ecoinf.2024.102578

Christin, S., Hervet, É., & Lecomte, N. (2019). Applications for deep learning in ecology. *Methods in Ecology and Evolution*, *10*(10), 1632–1644. https://doi.org/10.1111/2041-210x.13256

Fleuré, V., Planolles, K., Claverie, T., Mulot, B., & Villéger, S. (2025). Impact of deep learning and post-processing algorithms performances on biodiversity metrics assessed on videos. *PloS One*, *20*(8), e0327577. https://doi.org/10.1371/journal.pone.0327577

Gadot, T., Istrate, Ștefan, Kim, H., Morris, D., Beery, S., Birch, T., & Ahumada, J. (2024). To crop or not to crop: Comparing whole-image and cropped classification on a large dataset of camera trap images. *IET Computer Vision*. https://doi.org/10.1049/cvi2.12318

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. In *arXiv [cs.LG]*. arXiv. http://arxiv.org/abs/1412.6980

Morris, D. (n.d.). *MegaDetector: MegaDetector is an AI model that helps conservation folks spend less* *time doing boring things with camera trap images*. Github. Retrieved September 5, 2025, from https://github.com/agentmorris/MegaDetector

Norouzzadeh, M. S., Morris, D., Beery, S., Joshi, N., Jojic, N., & Clune, J. (2021). A deep active learning system for species identification and counting in camera trap images. *Methods in Ecology and Evolution*, *12*(1), 150–161. https://doi.org/10.1111/2041-210x.13504

Raghu, M., Zhang, C., Kleinberg, J., & Bengio, S. (2019, February 13). Transfusion: Understanding Transfer Learning for Medical Imaging. *Advances in Neural Information Processing Systems, 2019*. https://proceedings.neurips.cc/paper_files/paper/2019/file/eb1e78328c46506b46a4ac4a1e378b91-Paper.pdf

Rowcliffe, J. M., Kays, R., Kranstauber, B., Carbone, C., & Jansen, P. A. (2014). Quantifying levels of animal activity using camera trap data. *Methods in Ecology and Evolution*, *5*(11), 1170–1179. https://doi.org/10.1111/2041-210x.12278

Shorten, C., & Khoshgoftaar, T. M. (2019). A survey on image data augmentation for deep learning. *Journal of Big Data*, *6*(1), 60. https://doi.org/10.1186/s40537-019-0197-0

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., Teton, B., Beasley, J. C., Schlichting, P. E., Boughton, R. K., Wight, B., Newkirk, E. S., Ivan, J. S., Odell, E. A., Brook, R. K., … Miller, R. S. (2019). Machine learning to classify animal species in camera trap images: Applications in ecology. *Methods in Ecology and Evolution*, *10*(4), 585–590. https://doi.org/10.1111/2041-210x.13120

Tan, M., & Le, Q. (2021). EfficientNetV2: Smaller Models and Faster Training. *International Conference on Machine Learning*, 10096–10106. https://proceedings.mlr.press/v139/tan21a.html

Torralba, A., & Efros, A. A. (2011). Unbiased look at dataset bias. *CVPR 2011*, 1521–1528. https://doi.org/10.1109/cvpr.2011.5995347

van Lunteren, P. P., Rajmohan, R. Z., Aandahl, Brook, B. W., & Buettel, J. C. P. (2025). MEWC: A user-friendly AI workflow for customised wildlife-image classification. *Peer Community Journal*, *5*(e57). https://doi.org/10.24072/pcjournal.565

WildObs. (2025). *Wildlife Observatory of Australia*. https://doi.org/10.15468/kjqw3f

Yang, S. B. D. M. (2019). Efficient Pipeline for Camera Trap Image Review. In *arXiv [cs.CV]*. arXiv. http://arxiv.org/abs/1907.06772

Yosinski, Clune, J., Bengio, Y., & Lipson, A. H. (2014). How transferable are features in deep neural networks? *arXiv [cs.LG]*. arXiv. http://arxiv.org/abs/1411.1792