Received: 1 July 2021 | Accepted: 3 December 2021

DOI: 10.1111/2041-210X.14031

Methods in Ecology and Evolution

RESEARCH ARTICLE

Can CNN-based species classification generalise across variation in habitat within a camera trap survey?

Danielle L. Norman[1,2] | Philipp H. Bischoff[1] | Oliver R. Wearn[2,3] | Robert M. Ewers[1] | J. Marcus Rowcliffe[2] | Benjamin Evans[2,4] | Sarab Sethi[5,6] | Philip M. Chapman[7] | Robin Freeman[2]

[1]Department of Life Sciences, Imperial College London, Ascot, UK; [2]Institute of Zoology, Zoological Society of London, London, UK; [3]Fauna & Flora International – Vietnam Programme, Hanoi, Vietnam; [4]Brunel University London, Kingston Lane, Uxbridge, UK; [5]Department of Plant Sciences, University of Cambridge, Cambridge, UK; [6]Centre for Biodiversity and Environment Research, University College London, London, UK and [7]BSG Ecology, Worton Park, Oxfordshire, UK

Correspondence
Danielle L. Norman
Email: d.norman18@imperial.ac.uk

Funding information
AXA Research Fund; Natural Environment Research Council, Grant/Award Number: NE/P012345/1; Sime Darby Plantation Bhd

Handling Editor: Aaron Ellison

Abstract

1. Camera trap surveys are a popular ecological monitoring tool that produce vast numbers of images making their annotation extremely time-consuming. Advances in machine learning, in the form of convolutional neural networks, have demonstrated potential for automated image classification, reducing processing time. These networks often have a poor ability to generalise, however, which could impact assessments of species in habitats undergoing change.
2. Here, we (i) compare the performance of three network architectures in identifying species in camera trap images taken from tropical forest of varying disturbance intensities; (ii) explore the impacts of training dataset configuration; (iii) use habitat disturbance categories to investigate network generalisability and (iv) test whether classification performance and generalisability improve when using images cropped to bounding boxes.
3. Overall accuracy (72.8%) was improved by excluding the rarest species and by adding extra training images (76.3% and 82.8%, respectively). Generalisability to new camera locations within a disturbance level was poor (mean F1-score: 0.32). Performance across unseen habitat disturbance levels was worse (mean F1-score: 0.27). Training the network on multiple disturbance levels improved generalisability (mean F1-score on unseen disturbance levels: 0.41). Cropping images to bounding boxes improved overall performance (F1-score: 0.77 vs. 0.47) and generalisability (mean F1-score on unseen disturbance levels: 0.73), but at a cost of losing images that contained animals which the detector failed to detect.
4. These results suggest researchers should consider using an object detector before passing images to a classifier, and an improvement in classification

might be seen if labelled images from other studies are added to their training data. Composition of training data was shown to be influential, but including rarer classes did not compromise performance on common classes, providing support for the inclusion of rare species to inform conservation efforts. These findings have important implications for use of these methods for long-term monitoring of habitats undergoing change, as they highlight the potential for misclassifications due to poor generalisability to impact subsequent ecological analyses. These methods therefore need to be considered as dynamic, in that changes to the study site would need to be reflected in the updated training of the network.

KEYWORDS
camera trap, convolutional neural network, deep learning, disturbance, generalisability, image classification, object detection

1 | INTRODUCTION

Camera traps have become an increasingly popular survey tool among ecologists and conservationists, being used in a variety of studies, including of wildlife distribution, abundance, occupancy, behaviour and community structure (Burton et al., 2015). Their two biggest advantages are that they sample a relatively broad spectrum of wildlife, making them effective for monitoring species richness, and that they can operate night-and-day for months at a time, meaning that they can produce useful data on even the rarest species (Wearn et al., 2019). They also produce thousands, or in some cases, millions, of images for analysis.

Sifting out empty images and tagging images of animals can be a very time-consuming task for researchers. Although workflow efficiency and task complexity are probably hugely variable in 'real-world' settings, our experience is that an operator can process on the order of 1000–5000 images per day (assuming a basic task of tagging species and counting individuals). Recent advances in machine learning have seen the application of neural networks to this task to reduce the burden on researchers and reduce processing time (Beery et al., 2018; Norouzzadeh et al., 2018; Swanson et al., 2015; Tabak et al., 2019; Willi et al., 2019). In the largest comparison of machine learning architectures for the task of identifying species to date—based on the 3.2 million-image Snapshot Serengeti dataset—an overall accuracy of 93.8% was achieved (Norouzzadeh et al., 2018). When restricted to only images, the network was confident of having categorised correctly, this rose to 99.3%. Overall, automating the task of identifying species could have saved over 8.4 years of manual human labelling time if implemented from the outset (Norouzzadeh et al., 2018). More recent studies have achieved even higher accuracies of 95.6% (Schneider et al., 2020) and 97.6% (Tabak et al., 2019). In their review, Wäldchen and Mäder (2018) predicted that the number of tools available, and their application to species identification tasks, will continue to increase in the future.

These high accuracy results are impressive, but do not provide the full picture since they represent performance when the network is trained and tested on images from the same camera trap locations. When networks are tested on images from camera locations unseen during training, performance invariably drops; the networks do not generalise well. Previous studies have reported varying accuracies in this case: 68.7% (Schneider et al., 2020) and 59% (Beery et al., 2018) when tested on unseen camera locations from within the camera trap dataset, and 82% (Tabak et al., 2019) when tested on camera locations from an alternative dataset. This drop in performance could be due to variables such as changes to the background scenery, lighting, camera position or average distance of subject from camera. Performance can also be impacted by variation in the distribution and density of species recorded by each camera (Wei Koh et al., 2021). The issue of poor generalisability is not unique to automated classification of camera trap images, however. In the related context of acoustic detection in birds, networks generalised poorly to new conditions including differing species balances, noise conditions or recording equipment (Stowell et al., 2019). Similarly, a 14.4% drop in marine mammal classifier accuracy occurred when testing on whistle data from a different region than that trained on (Erbs et al., 2017).

Many applications of machine learning to classification thus far have had a particular geographical focus (Weinstein, 2018), but in order for these techniques to be widely applicable and impactful, architectures are required that can be used by multiple researchers on different datasets, ideally without having to perform the time-consuming network training at each new location (Wearn et al., 2019). In a world increasingly impacted by anthropogenic activity resulting in habitat degradation and fragmentation, it is also important that we have classifiers that are robust to changes in image background to facilitate long-term monitoring of habitats undergoing change. Otherwise, when conducting analyses using images classified by a network trained on images from pristine habitat, we risk drawing wrong conclusions if the new habitat has been altered to such an extent that the image backgrounds have changed. The impact of habitat degradation on network generalisability has not been considered to date.

Here, we compare the performance of three established architectures to identify species in camera trap images taken from undisturbed and disturbed tropical forests in Borneo. We specifically aim to explore the extent to which a network is able to generalise, which we achieve by splitting the dataset into an environmental gradient of varying levels of habitat disturbance generated by historical logging. Our goals are to (1) assess the performance of established architectures and identify the best network for classifying images within our dataset; (2) explore the impacts of training dataset configuration on overall performance, specifically restricting the data to common species only or increasing the number of images per species class included; (3) use the disturbance-level categories attributed to the camera trap locations to investigate the generalisability of the chosen network within our dataset; and (4) compare generalisablity performance when images are cropped to bounding boxes. The results of this study will inform the robust application of automated image classification for monitoring biodiversity in habitats undergoing change.

2 | MATERIALS AND METHODS

The dataset comprises camera trap images taken from the Stability of Altered Forest Ecosystem (SAFE) Project (Ewers et al., 2011) in Malaysian Borneo, a subset of which form the open access BorneoCam dataset. Camera trapping took place between May 2011 and March 2018, following the sampling procedure laid out in Wearn et al. (2013). The data were originally collected under approval from the Government of Malaysia, with the following permit numbers: Economic Planning Unit 40/200/19/2656; Maliau Basin Management Committee MBMC/2010/15, and Sabah Biodiversity Council JKM/MBS.1000-2/3 (84), JKM/MBS.1000-2/2 JLD.7 (51), JKM/MBS.1000-2/2 JLD.5 (142), JKM/MBS.1000-2/2 JLD.4 (192) and JKM/MBS.1000-2/2 JLD.3 (125). This dataset makes an ideal case study since it represents a realistic ecological dataset, in terms of size and in level of imbalance between classes, and it comprises a variety of habitat disturbance levels. These images have previously been used to inform analyses of mammalian species abundance (Wearn et al., 2017), diversity (Wearn et al., 2016) and behaviour (Davison et al., 2019) across a gradient of land-use comprising unlogged forest, logged forest and oil palm plantations. Forest quality at the locations of individual camera traps has been quantified into a five-step disturbance scale: (1) undisturbed forest, (2) disturbed forest, (3) heavily disturbed forest, (4) herbaceous scrub and (5) open area (Wearn et al., 2017). (Full descriptions of disturbance categories are provided in Table S1.)

The total raw data consisted of 753,442 images from 681 camera deployments. To construct a dataset of labelled images, untagged images were removed, as well as images captured during the setup process or a camera malfunction, or containing non-target (reptile or invertebrate) or multiple species. Empty images were also removed since we were interested in classification of species and so made the explicit assumption that the step of separating images into empty and non-empty had previously taken place. Camera traps (Reconyx HC500) were programmed to take a rapid burst of 10 images, termed a capture event. Image labelling for this dataset is at the level of images, rather than events. All non-empty images from a given event were allocated together to either the training or test dataset. A small proportion (0.05% of images), where the event grouping was not recorded in the metadata, were discarded. This reduced the dataset to 378,000 images from 640 deployments. Both day and night images were included since previous studies have found this had little effect on performance (Norouzzadeh et al., 2018; Tabak et al., 2019).

A minimum of 40 images or four capture events, per species class was imposed. To include as many species as possible, species that fell below this threshold were grouped together with related species, for example, Hose's civet *Diplogale hosei* images were included within the banded civet *Hemigalus derbyanus* class. These group classes comprised between 2 and 15 species (detail provided in Table S2). To limit imbalance within the dataset, and to reduce computation time and resources required, the maximum number of images per class was restricted to 5000. We investigated an increased maximum per class (Figure 1) and found that it improved Top-1 accuracy while having a small impact on mean F1-score, suggesting that it resulted in more bias towards common classes. A 90:5:5 split for training, validation and test sets was used following Willi et al. (2019), and to ensure matching distributions across classes within the three sets (Table S4; Figure S1). This resulted in training, validation and test sets consisting of 76,637, 4290 and 4309 images, respectively, each containing images from 51 classes. Images were resized to 256 × 256 pixels before passing to the neural networks. Data augmentation was also performed, consisting of random shearing, horizontal flipping, cropping and brightness modification (Table S5). This is commonly carried out in image classification problems to bolster training data and prevent overfitting (Beery et al., 2018; Krizhevsky et al., 2012).

To identify the best network for our dataset, we compared performance of three architectures: VGG16 (Simonyan & Zisserman, 2015), Inceptionv3 (Szegedy et al., 2016) and ResNet50 (He et al., 2016). In each case, the network was pre-trained on ImageNet, which is a large database of quality-controlled, human-annotated images, including animal classes, and is commonly used to pre-train networks for image classification tasks (Deng et al., 2009). Our baseline hyperparameter settings were based on those used by Norouzzadeh et al. (2018). All models were trained for 40 epochs—more epochs and early stopping were also assessed with small changes in validation loss as the stopping criteria, but no difference was found in resulting models.

As well as optimising hyperparameter settings for our dataset, we also investigated the impact of altered dataset configurations. We created a second dataset consisting of only the most common species by restricting the baseline dataset to classes that had a minimum number of 1000 images (rather than 40), which left 21 of the original 51 classes. We also created a third dataset in which the

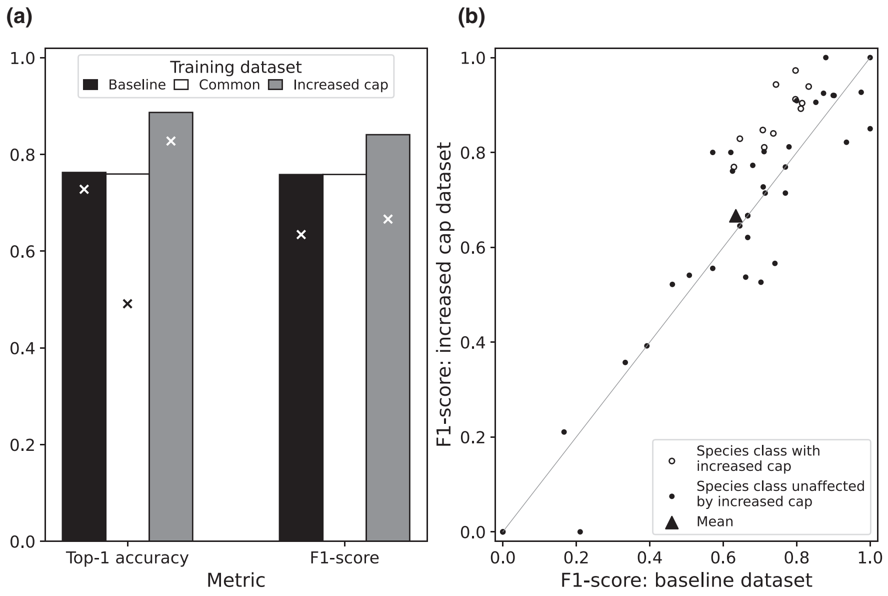

FIGURE 1 (a) Top-1 accuracy and F1-score for the Inceptionv3 network trained on the baseline dataset (max. 5000 images per class, black), common species only dataset (white) and increased cap dataset (max. 10,000 images per class, grey) when tested on common species only (bars) and when tested on the baseline test dataset (x). Note: It is not possible to calculate F1-score for the network trained on the common species only and evaluated on the full baseline test set due to model structure. (b): F1-score per species class when evaluated on the baseline test set for the network trained on the dataset with an increased cap against trained on the baseline dataset, with a 1:1 line for reference.

cap on the number of images per class was increased from 5000 to 10,000, which affected 11 of the 51 classes. In each case, performance was evaluated on the baseline and common-only test sets.

2.1 | Generalisability

To form the datasets for comparing performance across individual habitat disturbance levels, we removed all images from locations without a disturbance score (26,546 images). We then formed two datasets: one following the same procedure as above, where all images from a single event were allocated to either the training or test set following a 90–10 split (event-level), and one where 10% of the cameras within each disturbance level were withheld to form a pool for the test set, with the remaining 90% forming the pool of images for the training set (camera-level). For the event-level dataset, we imposed a minimum of four capture events per class per habitat type, leaving 14 classes. For the camera-level dataset, we restricted the data to classes which were captured on at least two cameras in all five disturbance levels, leaving 14 classes. In both cases, we imposed a cap of 5000 images per class. Only the best performing network, Inceptionv3, as identified from the initial network comparison, was used for the generalisability analysis.

To assess the effect of increasing the number of disturbance levels included in the training set on generalisability, we trained the network on images from every possible combination of disturbance level. To negate the impact of varying numbers of images across the disturbance level combinations, for each dataset configuration we fixed the total number of training images per combination to the smallest individual disturbance-level training set size, and randomly sampled images evenly across the included disturbance levels to meet this, ensuring all classes were captured. A test set was similarly formed for each individual disturbance level, ensuring consistent size and all classes included, and used to assess the performance of each combination on images from disturbance levels both seen and unseen during training.

2.2 | Bounding boxes

One suggested method for improving generalisability is to use an object detector to locate animals within the image and pass the image cropped to the resulting bounding box to the network for training (Beery et al., 2018). Here, we passed the images used for the disturbance-level combinations datasets through the Microsoft 'MegaDetector' v3 (Beery, Morris, & Yang, 2019). In most cases, a single object was identified, and the image was cropped to this bounding box and resized to 256×256 as above. Where more than one object was detected, we used the bounding box with highest confidence. In some cases, no object was detected despite being manually labelled as containing an animal. This was caused overwhelmingly by false negatives on the part of the MegaDetector, especially when animals were entering or exiting the field of view and were only partially visible (e.g. only parts of the legs or tail visible). These images were excluded from the generalisability analysis to create a fair comparison. To replicate performance in a 'real-world' scenario, however, a comparison of accuracy with and without these images included in the test sets is provided in the SI (Figure S7).

The combined disturbance-level datasets were then replicated using these bounding boxes in place of the whole images and the networks trained. We tested the networks on both the original test set for the disturbance level combinations, and the corresponding test set with images cropped to bounding boxes, for comparison.

2.3 | Metrics

Model performance was assessed against a test dataset which contained images distinct from those used to train the classifier. The performance metrics used are in line with those used in similar studies (Beery et al., 2018; Norouzzadeh et al., 2018; Tabak et al., 2019): (1) Top-1 and Top-5 accuracy: the proportion of all individual images in the test set that were correctly classified within the top 1 or 5 predictions, respectively; (2) F1-score: the harmonic mean of precision and recall, where (a) Precision is the proportion of predictions per class that were correct, that is, an indication of how reliable the predictions are for a given class; and (b) Recall is the proportion of images per class that were correctly identified, that is, how fully detected a given class is; and (3) Top-1 accuracy on an event basis: the proportion of capture events that contain at least one correctly classified image.

All metrics were evaluated for the initial network comparison, but generalisability was assessed using F1-score only. Since Top-1 accuracy is heavily influenced by the most common species, we consider F1-score to be a better metric to assess overall performance on an imbalanced dataset.

3 | RESULTS

3.1 | Network and dataset comparison

Performance of the three network architectures was comparable (Table 1), with the same pattern seen across the four metrics. All networks achieve higher Top-5 accuracy (mean 87%) and Top-1 event accuracy (mean 81%) than Top-1 accuracy on individual images (mean 73%). F1-score is consistently lowest (mean 0.62). Following optimisation (Figure S3), we chose to proceed with the Inceptionv3 network on the basis of F1-score.

When evaluated on only the common classes, overall performance was comparable for the network trained on only the common species and the network trained on all species in the baseline dataset (Top-1: 76% and F1: 0.76 in both cases; Figure 1a). Overall Top-1 accuracy and F1-score were improved by increasing the cap on the number of images per class in the training data (Top-1: 89%, F1: 0.84; Figure 1a). Including the rarer species in the test set, that is, evaluating performance on the full baseline dataset, saw lower scores for both networks trained on the baseline dataset and on the increased cap dataset (Top-1: 73% and 83%, F1: 0.63 and 0.67, respectively). There was a bigger loss of performance in terms of F1-score than Top-1 accuracy with the inclusion of the rarer species (Figure 1a), reflecting the bias towards common species in the Top-1 accuracy metric. Including all of the rarer species' test images in the evaluation of the network trained on common species only results in an absolute reduction in Top-1 accuracy of 27% (Figure 1a). Species-level F1-score and recall tended to increase with a greater number of training images available in the baseline dataset (Figure S5). This was again demonstrated in the increased cap dataset where all classes that benefitted from extra images saw an increase in F1-score, although the overall mean was only slightly higher than that trained on the baseline dataset (mean F1: 0.67 and 0.63, respectively; Figure 1b). This highlights the trade-off with increasing the imbalance within the training data, where some of the classes that did not have any additional training images saw a decrease in F1-score.

TABLE 1 All metrics for the best configurations of VGG16, ResNet50 and Inceptionv3 evaluated on the baseline dataset

Architecture	Top-1 accuracy	Top-5 accuracy	Top-1/event	F1-score
Inceptionv3	72.8%	86.5%	79.8%	0.63
ResNet50	73.5%	88.6%	81.2%	0.62
VGG16	73.2%	87.1%	80.8%	0.61

3.2 | Generalisability

For the dataset split at event-level only, peaks in F1-score occurred where the network was trained and tested on images from the same habitat disturbance level (mean: 0.76), while performance dropped substantially on disturbance levels not present in the training data (mean: 0.30; Figure 2a). Although the distribution across classes for each disturbance level was roughly even, heavily disturbed forest had the greatest number of cameras and images (Figure S2), which may have contributed to it achieving the highest F1-score (0.46) on an unseen disturbance level.

As the number of disturbance levels included within the training dataset was increased from one to four, the mean F1-score on the unseen habitats also increased (0.32, 0.41, 0.45, 0.49, respectively), that is, the network generalised better (Figure 2b). Conversely, F1-score for the disturbance levels seen during training tended to decrease (0.77, 0.76, 0.75, 0.74, respectively; Figure 2b). Performance on unseen habitats was still relatively poor, however, when only one disturbance level was omitted from the training dataset (Figure 2b).

Using images cropped to bounding boxes in both the training and test sets improved both the overall mean F1-score (0.87) and generalisability (mean F1 score on unseen disturbance levels when trained on a combination of 1, 2, 3 and 4 disturbance levels, respectively: 0.72, 0.80, 0.81, 0.84; Figure 2c). Training on bounding boxes and testing on whole images showed a large drop in performance (Figure 2c).

For the dataset split at camera level, and network trained on whole images, performance on seen disturbance levels was slightly better than on unseen disturbance levels (Figure 2d). As with the event-level dataset, an improvement was seen when the number of disturbance levels included within the training data was increased (mean F1-score on seen disturbance level combinations of one, two, three and four levels, respectively: 0.32, 0.40, 0.42, 0.44; unseen: 0.27, 0.36, 0.41, 0.41, Figure 2d). Performance was best when the network was trained on all disturbance levels (mean F1-score: 0.47).

Using bounding boxes on this dataset again improved performance, but overall F1-score was lower than that achieved with the event-level dataset (mean F1 on all five disturbance levels when trained and tested on cropped images: 0.77 versus 0.87, Figure 2e).

In a 'real-world' scenario, in which the images containing an animal undetected by the MegaDetector are included in the test set, we can see that Top-1 accuracy drops by 5% for the network

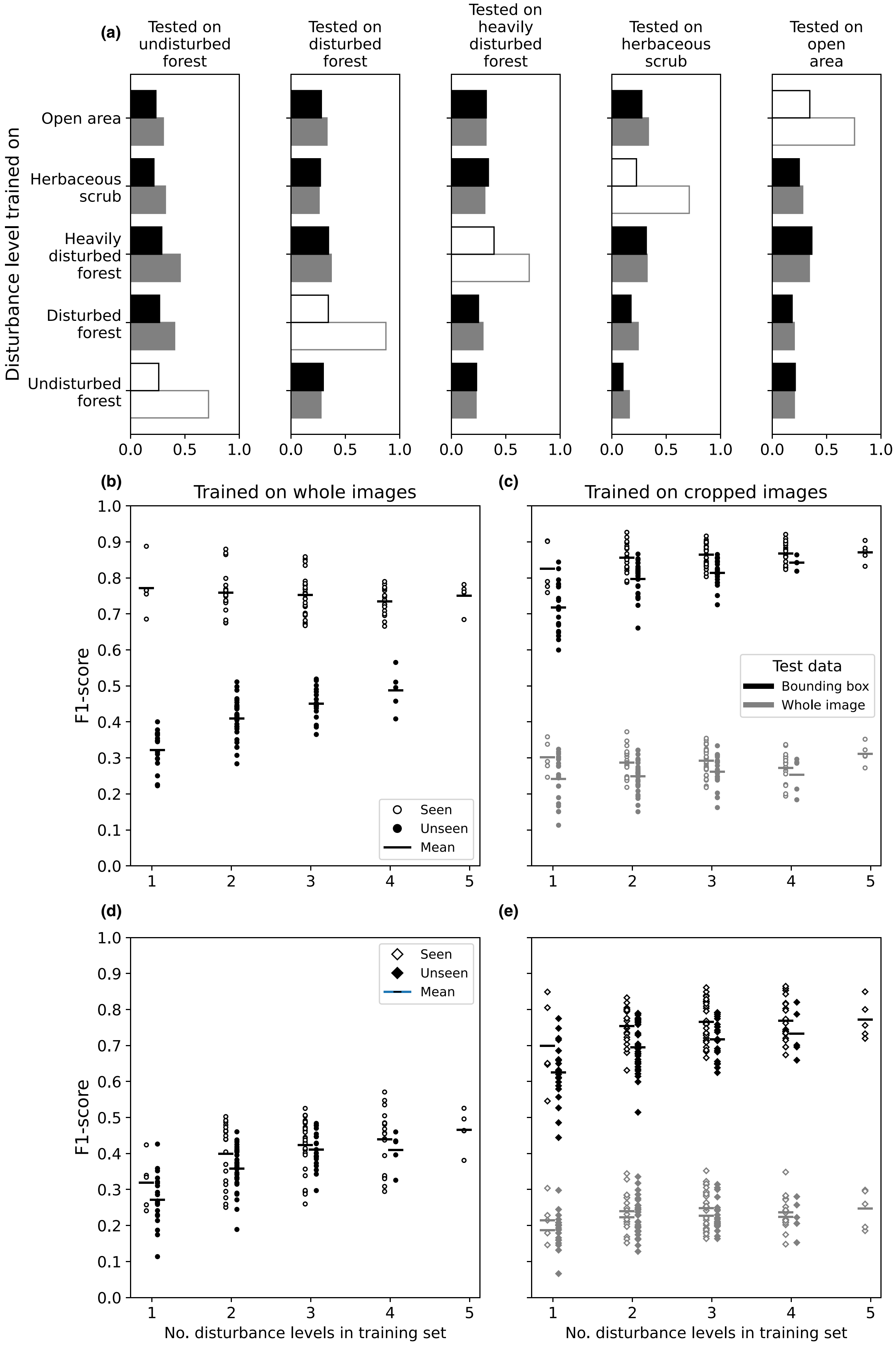
(a)
Tested on undisturbed forest
Tested on disturbed forest
Tested on heavily disturbed forest
Tested on herbaceous scrub
Tested on open area
Disturbance level trained on
Open area
Herbaceous scrub
Heavily disturbed forest
Disturbed forest
Undisturbed forest
0.0
0.5
1.0
(b)
Trained on whole images
(c)
Trained on cropped images
F1-score
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
1
2
3
4
5
Seen
Unseen
Mean
Test data
Bounding box
Whole image
(d)
(e)
Seen
Unseen
Mean
No. disturbance levels in training set

FIGURE 2 Network generalisability. (a) Disturbance-level comparison: F1-score per individual disturbance levels. White bars denote F1-score for the same disturbance level used in both training and testing. Filled bars denote F1-score for disturbance levels not seen during training. Results for the dataset split at event level are shown in black and at camera level in grey. (b and d): Disturbance-level combinations for the dataset split at event level (circles) and camera level (diamonds): F1-score for each disturbance level tested following training on a combination of disturbance levels. Every combination of disturbance level was included. Disturbance levels seen during training are denoted by a white marker, while those unseen are denoted by a black marker. The mean F1-score for each number of combinations is also marked. (c and e) disturbance level combinations using bounding boxes: as for (b and d), respectively, but training was performed on images cropped to bounding boxes. Black symbols denote testing on images cropped to bounding boxes, while grey represent performance when tested on whole images.

trained on cropped images, since all of these images are deemed to be wrongly classified (Figure S7d). For the network trained on whole images, performance does decrease, but to a lesser extent (1%) since the network has the opportunity to classify these images (Figure S7c).

4 | DISCUSSION

This is the first study to assess the application of automated image classification methods and, more specifically, the implications of poor generalisability of CNNs, when considered across a gradient of habitat disturbance in tropical rainforest. Our results highlight the ongoing issues with poor generalisability to unseen camera locations in camera trap image classification, as well as the additional problem of generalisability to changes in background associated with varying levels of habitat disturbance within a single camera trap dataset (Figure 2). Training across multiple disturbance levels improved generalisability, suggesting that these differences can be mitigated. Our results demonstrate that an awareness of variation in habitat backgrounds is required when planning a camera trap survey intended for automated classification, and when training the classifier.

We found that in addition to classification accuracy being lower at unseen camera locations within a disturbance level, performance was worse in unseen disturbance levels. One important implication of a lack of generalisability across levels of habitat disturbance, particularly in the context of increasing levels of habitat change, is that classifiers should not be considered 'static'. If a habitat changes over time, naturally or through anthropogenic impacts, new data and additional computer power may be needed to ensure derived classifications and ecological estimates are correct. Alternatively, a dataset comprising images from across the range of possible disturbances should be sought for training at the outset. Field ecologists wanting to use automated classification should therefore consider the generalisability issue when designing future camera trap surveys by stratifying their sites, a priori, by broad background types.

Although we fixed the overall number of training images per disturbance-level combination, the number of images per class was allowed to vary—by sampling in this way we aimed to replicate the abundance distribution of species within each habitat. Our analysis evaluated performance across all possible combinations of disturbance levels, which should mitigate the impacts of particularly distinct distributions of species in some disturbance levels. We additionally explored how our results changed when we only applied our classifier to the most abundant species within our study (Figure 1a; Figure S6), and found little difference. We also ensured that every class included occurred in both training and test sets to avoid differences in class distribution. Our results suggest that researchers working on smaller camera trap studies might see an improvement in classification performance if labelled images from other studies from similar habitat were to be added to their training data. Furthermore, these results support the aggregation of images from across studies on platforms such as Wildlife Insights (Ahumada et al., 2020) to enhance available training data and improve classification.

Other researchers have found that a detector–classifier combination was more generalisable than a classifier alone when applied to their dataset (Beery et al., 2018). Our results from the bounding box analysis support this, showing that by focussing images on the animals present and reducing the amount of background, the network was better able to identify species across all disturbance levels (Figure 2c,e). The results also highlight the need to test on bounding boxes rather than the whole image (Figure 2c,e). This is important from a practical perspective, since all images would need to be passed through a detector before being classified, including any new test images from ongoing projects; this adds computational time. We note, however, that even with images cropped to bounding boxes, the impact of background differences across disturbance levels is still evident, with mean F1-score on unseen disturbance levels rising from 0.62, when a classifier is trained on a single disturbance level, to 0.73 for a classifier trained on four levels (Figure 2e).

The use of an object detector highlighted some discrepancy between the expert labellers and the detector in identifying images as being empty. As a result, object detection may miss subjects that could have been classified correctly. In our data, the missed detections equated to a 5% reduction in Top-1 accuracy (Figure S7d). Although not the focus of this study, a review highlighted that in cases where only a very small part of the animal is visible, or the animal is mostly obscured by vegetation, a human has been able to identify that an animal is present using the visual aid of the whole event sequence, whereas a detector, without that context, could not. Image metadata have been found to improve both automated classification (Terry et al., 2020) and per-species detection performance (Beery et al., 2020) thus could similarly improve detection here.

Since classification performance could have a significant impact on the outcomes of ecological studies, the choice of metric used is important. Our Top-1 accuracy scores (overall: 73%; generalisability analysis: 68%) were slightly lower than one other similar sized study (79% on seen locations, Beery et al., 2018), but much lower than reported by others (>93.8%, Norouzzadeh et al., 2018, 98%, Tabak et al., 2019). Part of these differences might be due to the difficulty of the classification in different datasets, for example caused by the extent of the background noise that needs to be overcome to detect animals. Images from dense forest environments, as used here, might be expected to present a harder task than open grassland environments (e.g. Snapshot Serengeti as used in Norouzzadeh et al. (2018)). In addition, Top-1 accuracy is naturally dominated by the more common species and can therefore be a misleading metric. We chose to mostly report results as F1-scores here as this combines precision and recall, and can more accurately reflect classification performance across classes. In practice, Top-5 accuracy can be useful for shortlisting possible species for manual classification, and Top-1 event accuracy can identify events for manual review, both resulting in time-savings. Future work might look into how network performance is assessed in the context of the ecological questions we wish to ask with the data. In particular, an important extension could explore the impact of biases arising from poor generalisability across disturbance levels on resulting ecological studies (e.g. on bias and precision of state variable estimates, such as animal density and occupancy, or on the statistical power to detect differences between experimental treatments).

Although the composition of training data was shown to influence image classification accuracy, including rarer classes did not compromise performance on more common classes, which is important for the continued inclusion of data on rare species to inform conservation efforts. The increased performance achieved through additional training images supports existing studies that have concluded that, although good results can be achieved on smaller training datasets, classification accuracy is generally improved by a larger training dataset (Willi et al., 2019). Here, we saw increased classification performance when the cap on common species was increased, which, through the use of the F1-score metric, we can confidently say was not driven by a simple numerical increase in correctly identified common species.

Class imbalance is common within camera trap datasets, and has been shown to impair the performance of neural networks (Buda et al., 2018). The resulting difficulty in training neural networks on rare species is a known problem (Beery, Liu, et al., 2019). In practice, if a network is able to satisfactorily classify and remove common species such that manual classification is reduced to the rare species, this would still result in substantial time savings. For conservation projects, however, where rare species are the main interest, we might require better performance, especially in recall (since false negatives likely have a higher conservation cost than false positives). Specifically improving classification on rare species was not the focus of this study, but oversampling and weighted loss methods have been tried elsewhere with some success (Norouzzadeh et al., 2018; Terry et al., 2020). Others have tried generating artificial images containing the rare species or incorporating images from other datasets (Beery, Liu, et al., 2019; Schneider et al., 2018). Future work could therefore include applying these methods to this dataset and assessing performance.

Despite the inherent difficulties with training on rarer species, and the general trend seen for increased F1-score with number of training images, we did see instances of high F1-score for relatively rare species (Figure S5). There was also variation in performance on the most common species. Future work will need to consider the degree of morphological variation within and among species as a possible contributing factor as to why networks are able to learn some species better than others.

5 | CONCLUSIONS

This study highlights the ongoing issue of poor performance of automated species classifiers across unseen locations in camera trap studies. Importantly, it also demonstrates that unseen backgrounds (here disturbance levels) can further impair classification performance. Unseen locations in novel habitat disturbance levels had poorer classification performance than those from unseen locations in habitat levels seen during training. Generalisability can be improved by the use of bounding-box object detection prior to species classification, but the use of bounding boxes did not completely eliminate the problem. As camera trap datasets become more abundant, and the use of machine learning for automated classification becomes more commonplace, it will be critically important to ensure that estimation of changes in ecosystem function and composition are not biased by methodological choices in detection and identification of species. This is particularly important in the context of current global biodiversity loss, for monitoring the impacts of anthropogenic activities on ecosystems and mitigating further declines.

AUTHOR CONTRIBUTIONS

Philipp H. Bischoff, Oliver R. Wearn, Sarab Sethi, Danielle L. Norman, Robin Freeman, Robert M. Ewers and J. Marcus Rowcliffe conceived the ideas and designed the methodology. Oliver R. Wearn and Philip M. Chapman collected camera trap images. Benjamin Evans provided bounding boxes for the BorneoCam dataset. Danielle L. Norman programmed the models. Danielle L. Norman, J. Marcus Rowcliffe, Robin Freeman and Robert M. Ewers analysed the data. Danielle L. Norman led the writing of the manuscript. All authors contributed critically to the drafts and gave final approval for publication.

ACKNOWLEDGEMENTS

Fieldwork to collect the original image dataset would not have been possible without the collective effort of the SAFE Project's field staff. We thank Sui Heon and Henry Bernard in particular, and also Jessica Haysom and Jack Thorley for assisting with the image tagging. We thank James Rosindell for his supervision of preliminary work.

FUNDING INFORMATION

This work was supported by a Sime Darby Foundation grant to the SAFE Project. DN was supported by the QMEE CDT, funded by NERC grant number NE/P012345/1. ORW was supported by an AXA Research Fellowship.

CONFLICT OF INTEREST

None of the authors has a conflict of interest.

DATA AVAILABILITY STATEMENT

The downsized SAFE Project images and code used in this analysis are available from The Safe Project community data repository on Zenodo at https://zenodo.org/record/6627707# (Norman et al., 2022). This contains all of the data needed to replicate this analysis. The full-size BorneoCam imagery, including full metadata in COCO format, will be available through LILA BC (the Labelled Information Library of Alexandria: Biology and Conservation); see lila.science.

ORCID

Danielle L. Norman https://orcid.org/0000-0002-8228-9973
Oliver R. Wearn https://orcid.org/0000-0001-8258-3534
J. Marcus Rowcliffe https://orcid.org/0000-0002-4286-6887
Sarab Sethi https://orcid.org/0000-0002-5939-0432

REFERENCES

Ahumada, J. A., Fegraus, E., Birch, T., Flores, N., Kays, R., O'Brien, T. G., Palmer, J., Schuttler, S., Walter, J., Kinnaird, M., Kulkarni, S., Lyet, A., Thau, D., Duong, M., Oliver, R., & Dancer, A. (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

Beery, S., Liu, Y., Morris, D., Piavis, J., Kapoor, A., Meister, M., Joshi, N., & Perona, P. (2019). Synthetic examples improve generalization for rare classes. *ArXiv Preprint ArXiv*, *1904*, 05916. http://arxiv.org/abs/1904.05916

Beery, S., Morris, D., & Yang, S. (2019). Efficient pipeline for camera trap image review. http://arxiv.org/abs/1907.06772

Beery, S., Van Horn, G., & Perona, P. (2018). Recognition in Terra Incognita. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *11220 LNCS*, 472–489. https://doi.org/10.1007/978-3-030-01270-0_28

Beery, S., Wu, G., Rathod, V., Votel, R., & Huang, J. (2020). Context R-CNN: Long term temporal context for per-camera object detection. *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13072–13082. https://doi.org/10.1109/CVPR42600.2020.01309

Buda, M., Maki, A., & Mazurowski, M. A. (2018). A systematic study of the class imbalance problem in convolutional neural networks. *Neural Networks*, *106*, 249–259. https://doi.org/10.1016/j.neunet.2018.07.011

Burton, A. C., Neilson, E., Moreira, D., Ladle, A., Steenweg, R., Fisher, J. T., Bayne, E., & Boutin, S. (2015). Wildlife camera trapping: A review and recommendations for linking surveys to ecological processes. *Journal of Applied Ecology*, *52*(3), 675–685. https://doi.org/10.1111/1365-2664.12432

Davison, C. W., Chapman, P. M., Bernard, H., Ewers, R. M., & Wearn, O. R. (2019). Shifts in the demographics and behavior of bearded pigs (*Sus barbatus*) across a land-use gradient. *Biotropica*, *51*(6), 1–11. https://doi.org/10.1111/btp.12724

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database. *IEEE Conference on Computer Vision and Pattern Recognition*, 248–255.

Erbs, F., Elwen, S. H., & Gridley, T. (2017). Automatic classification of whistles from coastal dolphins of the southern African subregion. *The Journal of the Acoustical Society of America*, *141*(4), 2489–2500. https://doi.org/10.1121/1.4978000

Ewers, R. M., Didham, R. K., Fahrig, L., Ferraz, G., Hector, A., Holt, R. D., Kapos, V., Reynolds, G., Sinun, W., Snaddon, J. L., & Turner, E. C. (2011). A large-scale forest fragmentation experiment: The stability of altered forest ecosystems project. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *366*(1582), 3292–3302. https://doi.org/10.1098/rstb.2011.0049

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, *2016 December*, 770–778. https://doi.org/10.1109/CVPR.2016.90

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems 25*, 1097–1105.

Norman, D. L., Wearne, O. R., Chapman, P. M., Heon, S. P., & Ewers, R. M. (2022). Downsized camera trap images for automated classification. *Zenodo*, https://zenodo.org/record/6627707#

Norouzzadeh, M. S., Nguyen, A., Kosmala, M., Swanson, A., Palmer, M., Packer, C., & Clune, J. (2018). Automatically identifying, counting, and describing wild animals in camera-trap images with deep learning. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(25), E5716–E5725. https://doi.org/10.1073/pnas.1719367115

Schneider, S., Greenberg, S., Taylor, G. W., & Kremer, S. C. (2020). Three critical factors affecting automated image species recognition performance for camera traps. *Ecology and Evolution*, *10*(7), 3503–3517. https://doi.org/10.1002/ece3.6147

Schneider, S., Taylor, G. W., & Kremer, S. (2018). Deep learning object detection methods for ecological camera trap data. *Proceedings - 2018 15th Conference on Computer and Robot Vision, CRV 2018*, 321–328. https://doi.org/10.1109/CRV.2018.00052

Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. *3rd international conference on learning representations, ICLR 2015 - conference track proceedings*, 1–14.

Stowell, D., Wood, M. D., Pamuła, H., Stylianou, Y., & Glotin, H. (2019). Automatic acoustic detection of birds through deep learning: The first bird audio detection challenge. *Methods in Ecology and Evolution*, *10*(3), 368–380. https://doi.org/10.1111/2041-210X.13103

Swanson, A., Kosmala, M., Lintott, C., Simpson, R., Smith, A., & Packer, C. (2015). Snapshot Serengeti, high-frequency annotated camera trap images of 40 mammalian species in an African savanna. *Scientific Data*, *2*, 150026. https://doi.org/10.1038/sdata.2015.26

Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., & Wojna, Z. (2016). Rethinking the inception architecture for computer vision. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, *2016 December*, 2818–2826. https://doi.org/10.1109/CVPR.2016.308

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

Terry, J. C. D., Roy, H. E., & August, T. A. (2020). Thinking like a naturalist: Enhancing computer vision of citizen science images by harnessing contextual data. *Methods in Ecology and Evolution*, *11*(2), 303–315. https://doi.org/10.1111/2041-210X.13335

Wäldchen, J., & Mäder, P. (2018). Machine learning for image based species identification. *Methods in Ecology and Evolution*, *9*(11), 2216–2225. https://doi.org/10.1111/2041-210x.13075

Wearn, O. R., Carbone, C., Rowcliffe, J. M., Bernard, H., & Ewers, R. M. (2016). Grain-dependent responses of mammalian diversity to land use and the implications for conservation set- aside. *Ecological Applications*, *26*(5), 1409–1420. https://doi.org/10.1890/15-1363

Wearn, O. R., Freeman, R., & Jacoby, D. M. P. (2019). Responsible AI for conservation. *Nature Machine Intelligence*, *1*(2), 72–73. https://doi.org/10.1038/s42256-019-0022-7

Wearn, O. R., Rowcliffe, J. M., Carbone, C., Bernard, H., & Ewers, R. M. (2013). Assessing the status of wild felids in a highly-disturbed commercial forest reserve in Borneo and the implications for camera trap survey design. *PLoS ONE*, *8*(11), e77598. https://doi.org/10.1371/journal.pone.0077598

Wearn, O. R., Rowcliffe, J. M., Carbone, C., Pfeifer, M., Bernard, H., & Ewers, R. M. (2017). Mammalian species abundance across a gradient of tropical land-use intensity: A hierarchical multi-species modelling approach. *Biological Conservation*, *212*(October 2016), 162–171. https://doi.org/10.1016/j.biocon.2017.05.007

Wei Koh, P., Sagawa, S., Marklund, H., Xie, S. M., Zhang, M., Balsubramani, A., Hu, W., Yasunaga, M., Lanas Phillips, R., Gao, I., Lee, T., David, E., Stavness, I., Guo, W., Earnshaw, B., Haque, I., Beery, S. M., Leskovec, J., Kundaje, A., … Liang, P. (2021). WILDS: A benchmark of in-the-wild distribution shifts. *Proceedings - International Conference on Machine Learning*, *PMLR*, *139*, 5637–5664.

Weinstein, B. G. (2018). A computer vision for animal ecology. *Journal of Animal Ecology*, *87*(3), 533–545. https://doi.org/10.1111/1365-2656.12780

Willi, M., Pitman, R. T., Cardoso, A. W., Locke, C., Swanson, A., Boyer, A., Veldthuis, M., & Fortson, L. (2019). Identifying animal species in camera trap images using deep learning and citizen science. *Methods in Ecology and Evolution*, *10*(1), 80–91. https://doi.org/10.1111/2041-210X.13099

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Norman, D. L., Bischoff, P. H., Wearn, O. R., Ewers, R. M., Rowcliffe, J. M., Evans, B., Sethi, S., Chapman, P. M., & Freeman, R. (2023). Can CNN-based species classification generalise across variation in habitat within a camera trap survey? *Methods in Ecology and Evolution*, *14*, 242–251. https://doi.org/10.1111/2041-210X.14031