Imperial College London

AUGUST 2018

DEPARTMENT OF LIFE SCIENCES

Generalisability of convolutional neural networks for automated animal identification in camera trap images

AUTHOR:
PHILIPP H. BISCHOFF

INTERNAL SUPERVISORS:
JAMES ROSINDELL

EXTERNAL SUPERVISOR:
OLIVER R. WEARN

WORD COUNT: 5,994

A thesis submitted in partial fulfilment of the requirements for the degree of Master of Science at Imperial College London

Formatted in the journal style of 'Methods in Ecology and Evolution'
Submitted for the MSc in Computational Methods in Ecology and Evolution

Declaration

Camera trap data was by provided by researchers involved with the Stability of Altered Forest Ecosystems project, namely Dr Phil Chapman (Imperial College London) and Dr Oliver Wearn (Zoological Society of London). Dr James Rosindell (Imperial College London), Sarab Sethi (Imperial College London) and Dr Oliver Wearn were involved in shaping the direction of this project. Data analysis and writing was exclusively performed by the author. References and citations were formatted using an adapted BibTex style file from Sean Anderson (Fisheries and Oceans Canada).

Abstract

1. **Remote sensor camera traps are a popular tool for biomonitoring, but their use is heavily restricted by the manual labour needed to extract information from images. Machine learning algorithms known as convolutional neural networks (ConvNets) have shown great potential for automating the identification of animals. However, little has been done to characterise the features that affect their performance. Here, I tested the robustness of three ConvNets to correctly classify animals with regards to two factors, using data from the Stability of Altered Forest Ecosystem project.**

2. **Firstly, ConvNets categorise animals into arbitrary classes, usually species, which results in few images for rare animals. I assessed how changing the taxonomic level of classes affects identification ability by comparing three groupings, each differing in their taxonomic resolution. Secondly, camera traps are deployed in a variety of habitats. I investigated the effect of background scenery on identification ability by parameterising the ConvNets to certain vegetation types, and applying the neural networks to novel backgrounds. The performance of the ConvNets was assessed using three evaluation metrics, namely F-1, Top-1 and Top-5 accuracy.**

3. **Taxonomic breadth of groups has little effect on accuracy (<4%) for all metrics used. Detrimental effects on accuracy were observed however, when neural networks were applied to vegetations they had not encountered before, with accuracy decreases of more than 50%.**

4. **This study indicates that researchers with limited camera trap data could use ConvNets as an identification tool, by classifying animals into wider taxonomic groups. My results further suggest that publicly available ConvNets will only be a widely applicable tool if parameterised on a multitude of background types.**

1 INTRODUCTION

In light of anthropogenic induced habitat destruction and climate change, real-time monitoring of biodiversity has become fundamental for wildlife ecology and conservation efforts (Steenweg et al., 2017). At its core, it allows for informed diversity loss management and mitigation through the collection of data on the state of ecological systems and the effectiveness of conservation programs (Nichols & Williams, 2006). Remotely triggered camera traps provide a non-invasive method for observing fauna in its natural habitat (Swann et al., 2011). Instead of manually tracking or capturing animals, a motion sensor in the camera detects movement and triggers when an object comes into the camera's radius. A wide range of ecological research, primarily on mammalian species, makes use of such traps including studies of presence-absence, occupancy, abundance, diversity, and behaviour (McCallum, 2013; Burton et al., 2015). Technological advancements (Kucera & Barrett, 2011) and continued decreasing costs (Marcus Rowcliffe, 2017) have made camera traps an increasingly popular tool in ecology, with the number of scientific publications utilizing the technology increasing by 10% annually since the early 1990's (McCallum, 2013; Burton et al., 2015).

Despite their increase in popularity, camera trap studies are generally limited by their ability to handle the unprecedented amount of data they generate. A major analytical bottleneck has been the time-consuming, manual process of image labelling (Spampinato et al., 2015), in which each image must be checked for false positive triggers or species identities of present animals. Traditionally performed by ecological experts, this task is increasingly accomplished using crowd-sourcing platforms (Dickinson et al., 2010; Sauermann & Franzoni, 2015), where members of the general public are given access to tag images. While citizen participation can greatly reduce labour costs and return valuable time to researchers (Sauermann & Franzoni, 2015), a major drawback is the lack of experience and expertise of citizen scientists, which can heavily affect data quality through increased error (Dickinson et al., 2010). For camera trap studies, this necessitates the use of additional validation algorithms to determine a 'census' label from a pool of individual classifications (Swanson et al., 2015). Even so, crowd-sourcing techniques are still highly time intensive. The Snapshot Serengeti project for example, which deploys 225 cameras, accumulated approximately 5.5 million images between 2010 and 2013, and it has taken citizen scientists a combined 14.6 years to label them (Swanson et al., 2015).

FIGURE 1: Sample images commonly found in camera trap datasets. *Camera trap images are often of low quality, making it difficult to identify what is in the image. Some examples of challenging images are presented here. A) A lesser mouse deer located almost out of view in the bottom left. B) A binturong photographed at a very close range. C) Fog or a dirty lense blurs the image, and heavily blends the orangutan with its surroundings. D) The reduced saturation in black and white images make it difficult to pinpoint the common palm civet at the center-right of the image.*

Automation of the classification process could greatly increase the duration, breadth and repeatability of image-based ecological studies (Weinstein, 2018) and has consequently received a lot of attention in recent years (Pimm et al., 2015). The discipline of computer vision deals with these types of tasks by processing digital images and extracting high-level data, such as taxonomic ranks. Camera trap images, nonetheless, bring with them various challenges which are difficult to overcome (Figure 1). First, species identification is a fine-grained task, as species can exhibit similar physical traits such as size, shape, and fur pattern. Secondly, environmental conditions including background scenery, light intensity and lens obstruction heavily affect image quality. Lastly, animals do not cross the field of view of camera traps in a consistent way, and can approach from a variety of angles and distances. Combined with the fact that animals are deformable objects, animals can present as highly variable features in camera trap images.

Developments in machine learning and the re-emergence of convolutional neural networks (ConvNets) have shown great promise in overcoming these issues. Loosely based on the structure of the mammalian visual cortex (Hubel & Wiesel, 1962; LeCun et al., 1995), ConvNets are a type of neural network, which can convert multiple-array data (e.g. images) into suitable internal representations (e.g. feature vectors), to ultimately detect patterns in the input data (LeCun et al., 2015) (see Methods 2.1). The first ConvNet-based species identification model achieved an accuracy level of 38% for 20 North American species (Chen et al., 2014), however more recent attempts have been more successful. To date, the highest accuracy has

been produced by Tabak et al. (2018) who built, and provided an R package for, a ConvNet that reached an accuracy of 98% for 28 classes found across the United States. In addition to species recognition, Norouzzadeh et al. (2018) were able to determine the count and behaviour for 48 species in Serengeti National Park with a 63% and 76% accuracy, respectively.

While these findings demonstrate the potential of ConvNets for large scale object recognition in ecological research, an understanding of the underlying features of image datasets affecting accuracy is still lacking. For instance, Tabak et al. (2018) observed an unexplained drop in accuracy from 94% to 82% when using images containing the same species from different locations. This presents the possibility that ConvNets may be more suited to certain sets of images. However, in order for ConvNets to become a widely applied tool in image-based ecological research, they will need to produce reliable results for a variety of wildlife datasets. Some attempts have been made to investigate such factors, like Villa et al. (2017) who demonstrated accuracy can vary depending on the position of animals in images. Additional studies found that coloured images have no notable benefit over black and white images for ConvNet model performance (Norouzzadeh et al., 2018; Tabak et al., 2018). Still, very little research has gone into the effects of particular image features on the accuracy of ConvNets, which makes it difficult to apply them as an identification tool for a variety of studies and habitats.

A widely held assumption when applying ConvNets as object classifiers is that there is a high contrast between the object of interest and its background surroundings (Zhu et al., 2014). Currently, many studies achieving high accuracies using ConvNets in the field of ecology are making use of the Serengeti National Park dataset (Gomez et al., 2016; Villa et al., 2017; Norouzzadeh et al., 2018; Tabak et al., 2018), which mostly contains images with relatively uniform backgrounds of open grasslands and savannah. However, camera traps are deployed in a wide range of environments, predominantly in structurally-complex forested habitats or in otherwise heterogeneous landscapes (McCallum, 2013), where animals may blend in with their surroundings to a greater degree. This might mean that firstly, ConvNets are less accurate in general, and secondly performance is more strongly affected by backgrounds present in training sets.

The effect of identification specificity has also been largely overlooked. Unlike the Serengeti

National Park dataset, labelled ecological datasets are usually small in size (Chao, 1989). Combined with the fact that species of interest are often rare, it is difficult to attain a sufficient number of images needed to train a ConvNet. To overcome this problem, ConvNets may potentially be trained to identify animals on a coarser scale, so that species are classified on a higher taxonomic rank for example. Images could then be manually reviewed to identify species of interest, which would still drastically reduce labelling time for researchers by segmenting data into smaller sizes. Such an approach would additionally allow ConvNets to be applied to study regions with different species compositions, but similar broad taxonomic groups.

Here, I study the robustness of machine learning learning algorithms to features in ecological images by investigating the effects of 1) classification broadness and 2) background scenery on ConvNet model performance. More specifically, I assess three different ConvNets to correctly identify animal subjects in regards to these variables, using images from the Stability of Altered Forest Ecosystem (SAFE) project, a large-scale field experiment in Borneo with over 750,000 thousand labelled camera trap images. I evaluate three types of classifications, which differ in their breadth of taxa in each class, and expect that increasing the breadth within classes will increase the accuracy of the ConvNet models. As animals are classed with other species, features unique to particular species will create noise within classes, however, more basic features such as size and shape should become more distinguishable. I further asses the effects of background environment on model performance by filtering images using a vegetation index as a scenery type proxy. Since many species blend in with their environments in more complex habitat structures, I hypothesise that ConvNet models are highly affected by background noise.

2 MATERIALS AND METHODS

2.1 Convolutional Neural Networks

While ConvNets can vary extensively in structure, their underlying principles are the same. At the fundamental level, deep ConvNets are made up of three or more convolutional and fully connected layers, which together process the large input data of an image to extract features and classify objects within an image (Figure 2).

After an image is passed into the model, two dimensional matrices known as kernels, are

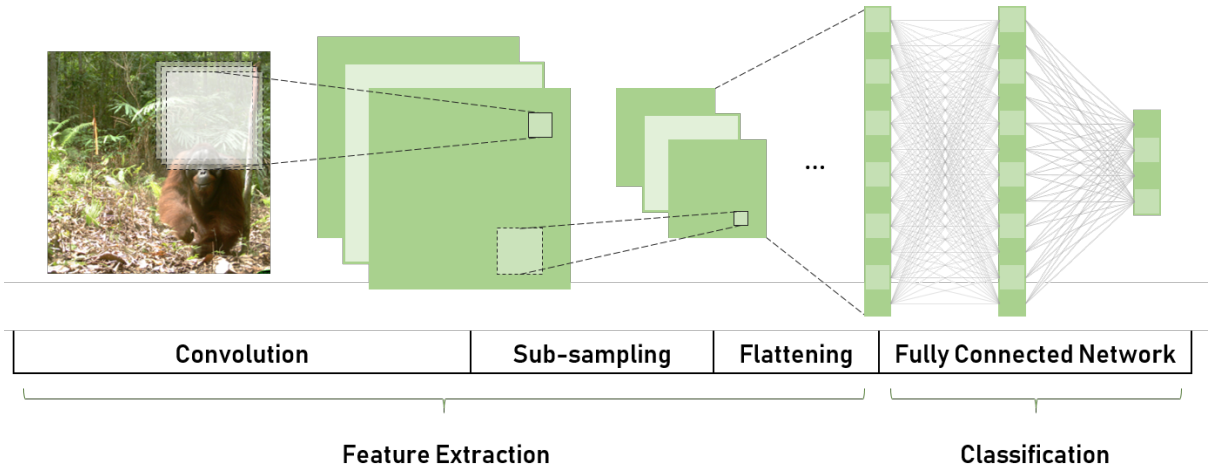

FIGURE 2: Visualisation of a basic convolutional neural network. *An image is taken as the input for the network. Multiple kernels are run over the image, which produce multiple feature maps through a process of mathematical operations. Each map highlights a specific feature of the image (contrast, borders, etc.). Sub-sampling is then applied, to reduce the size of the feature maps. This is usually done by taking the maximum or average value within a neighbourhood of pixel values. The process allows for the cutting of data, without great loss of information. These two steps, convolution and sub-sampling can then be repeated several times (ellipsis), taking feature maps as 'inputs' for convolutions. The feature extraction phase is then terminated by concatenating, or 'flattening', feature maps into a feature vector. The vector produced is used as the input for a neural net consisting of fully connected layers. The last layer consists of as many neurons as there are classes and gives a probabilistic statement for each class being in the image.*

moved across the image. During this process, values within the kernel, more commonly referred to as weights, are multiplied with pixel values to compute the sum of element-wise products known as activations. Activations collectively produce feature maps that can then in turn be further convoluted. This process allows ConvNets to extract broad features like edges in early stages, and more detailed structures like eyes in subsequent layers.

Following a series of convolutions, feature maps are transformed into a vector which is used as an input for a neural network of fully connected layers. A neural network is a collection of neurons, where each neuron is connected to all neurons of the last layer through weights. The activation of a neuron is then calculated by adding the product of the weights and activations of neurons in the previous layer. Finally, probabilities representing the chance that the input image belongs to a specific class, are returned in the final layer.

As images are passed through the model, an algorithm known as an optimiser attempts to minimise the loss function, a function which measures the error of a model's performance.

This step is referred to as training. The length of the training period is set by the number of epochs chosen, which are the number of iterations the model will be fed the complete training dataset. Weights are continually updated after a set of images, or 'batches', have been fed into the neural network. The parameter known as learning rate determines the magnitude by which weights are updated. Once the weights of the model have been tuned, the ConvNet can be evaluated by screening it against a test set, which is a collection of images the model has not seen before.

2.2 Model Training

In order to assess the generalisability of deep learning classifiers, I set out to test three frequently used ConvNet architectures. The three classifiers in question were Inception-v3 (Szegedy et al., 2015), ResNet-50 (He et al., 2015) and VGG-16 (Simonyan & Zisserman, 2014), which among other things vary in the type, number and size of layers deployed (see SI 6.1.5). The architecture was the same as the one from the original paper. I made use of transfer learning, where the weights in the models are initialised with pre-trained weights from ImageNet's ILSVRC competition (Russakovsky et al., 2015). Although the ILSVRC competition requires models to identify objects primarily irrelevant to this study, weights in early layers are already trained to distinguish broad features, and are generally thought to be better than randomised starting points. During training, I allowed all weights to be updated freely, and weights were only saved when there was an improvement on Top-1 accuracy (see Methods 2.7) on the test set. This was done to prevent the neural networks from over-fitting to the training images.

Training of ConvNets requires several parameters and hyper-parameters to be set. This was done by varying and testing different combinations until a desired result is achieved. I found the parameters summarised in Table 1 to be most suitable for this analysis. I used the Adam optimiser (Kingma & Ba, 2014) and categorical cross entropy loss function. I additionally made use of stepwise learning rate annealing, where the learning rate is reduced at specific epochs during training. This allows for better fine-tuning of the models. Models were trained and evaluated on a single NVIDIA Quadro P5000 graphic processing unit. Due to the high computational cost of training deep learning networks, models were only trained once, which generally allows weights to converge to an global optimum nonetheless (Dauphin et al., 2014; LeCun et al., 2015).

TABLE 1: Overview of hyper-parameters used for ConvNet training. *Hyper-parameter values were chosen based on experimental testing. There are no set boundaries for each parameter, creating a multitude of possible combinations. As it is impossible to manually test for all combinations, it is difficult to find optimal hyper-parameter values. Parameter were, therefore, based on values used in similar research (Norouzzadeh* et al.*, 2018), and optimised for this study.*

Hyper-parameter	Value	Details
Batch Size	64	Smaller batch sizes allow for increased generalisability at the cost of computational time.
Crop Size	244 $\times$ 244	A standard image size for deep learning models to avoid computational issues with memory allocation.
Epochs	40	A saturating effect is observed towards 40 epochs even with learning rate reduction.
Learning rate	1×10^{-5} / 5×10^{-6} / 1×10^{-6} / 5×10^{-7} / 1×10^{-7}	Learning rates are altered at epoch, 10, 20, 25, 30 and 35 respectively to fine tune weights.
β_1 & β_2	0.9 & 0.999	Controls the decay of the moving averages calculated by Adam optimser, which eases computation for model weights.

2.3 Dataset and Image Allocation

The SAFE project camera trap dataset, labelled by ecological researchers and experts, contained 750,988 images. Metadata was extracted from each image to obtain information on 1) the camera from which the image was taken, 2) the animal present in the image and 3) when the image was taken. Both coloured and black & white images were utilised in the analysis, as this is common for camera trap datasets. I discarded any images unsuitable for the automatic identification process, which included images with set-up/pick-up, malfunctions or non-species-specific tags (see SI 6.1.1). In order to compare single-species identification models, I additionally removed multi-species images. Lastly, images generated from camera traps which had no associated vegetation score were excluded from the analysis on background effects (see Methods 2.6). Images without labels were taken to be empty.

Camera traps take images in events known as trigger events, where a burst of images are taken whenever a motion is detected. As well as capturing animals in similar positions, the background and brightness is highly conserved in images from a single trigger event. I there-

fore allocated randomly selected images to the training and test sets, instead of assigning trigger events to respective sets. This prevents ConvNets from learning to identify species from a certain angle or within a certain lighting, and improves generalisability. The total image dataset was divided by a modified 75-25% split to create the training and test set, respectively (see SI 6.1.2). Due to the high imbalance of images between taxonomic groups, common for camera trap datasets, I limited the maximum number of images per group to 5000. ConvNets, and machine learning algorithms in general, can be highly affected by imbalance, as it causes accuracy biases towards highly representative groups. The minimum was set at 60 images per class, which, while low, was chosen to include as many animals as possible (see SI 6.1.2).

2.4 Pre-processing and Data Augmentation

The ConvNets used here are not capable of handling the initial Reconyx HC500 camera trap output image size of 2048 $\times$ 1536 pixels and had to be pre-processed (Figure 3). I scaled images down to a 256 $\times$ 256 resolution, which distorts the images but is a common practice in deep learning community (Goodfellow et al., 2016). Pixel values were then standardised to a [0,1] range. This is part of a normalisation strategy that facilitates training of models due to reduced computational intensity (Goodfellow et al., 2016). During training the images are further augmented by randomly shearing images by 0.2 radians and flipping them horizontally, which provides ConvNets with slightly different images during each epoch, thereby preventing over-fitting (Krizhevsky et al., 2012).

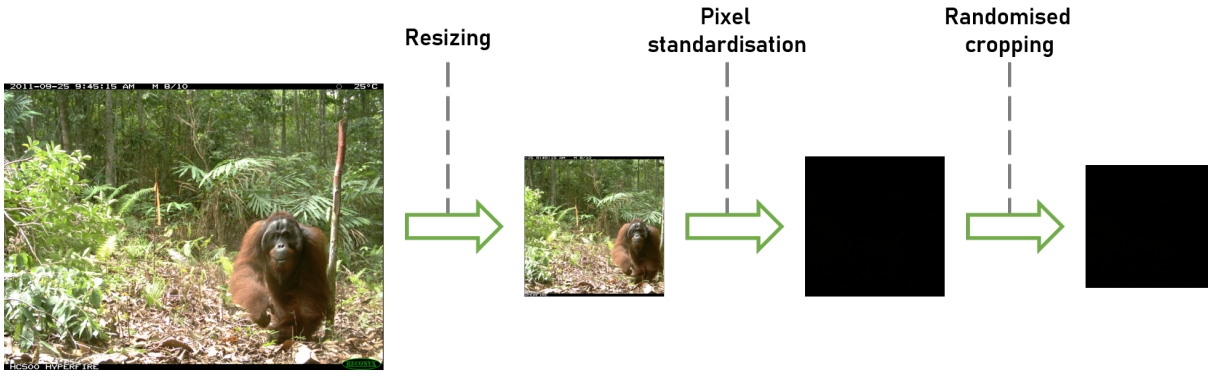

FIGURE 3: Image pre-processing workflow. *To facilitate training images are commonly processed before passing them to deep learning algorithms. Images are first resized by reducing the resolution of the image. Pixel values are then standardised to be between 0 and 1, which makes the image appear black to the human eye, but is preferred for deep learning. For each epoch, images are then further transformed with randomised cropping.*

2.5 Grouping and Classification Analysis

I worked under a one-stage identification process, where empty images are classed as a 'species' for the model to identify. The effect of classification specificity was assessed by testing and evaluating deep learning models on three separate levels of groupings. The first categorisation was done on the species level. The second classification was performed on the family level, which groups generally similar looking species. The last classification scenario involved a more morphological approach to grouping, that grouped phenotypically similar species into ecologically interesting groups (McCallum, 2013; Burton et al., 2015). This included grouping species by broad features, predominantly on outline, size, overall stance and time of general activity.

Birds are often considered undesired bycatch data of camera traps, and previous research tends to group them as a distinct class (Gomez et al., 2016; Tabak et al., 2018). I therefore classed birds on the order level for the first two groupings (species- and family level), and as a single class for the morphological grouping. If there were less than 60 images for a given class, I grouped them with their nearest relative (see SI 6.1.3). Taxonomic information was acquired using the ITIS database (Roskov et al., 2013). When this provided no viable results, the taxonomic NCBI database (Federhen, 2011) or Encyclopedia of Life (Parr et al., 2014) was consulted. Relatedness was assessed using OneZoom's Tree of Life (Rosindell & Harmon, 2012; Hinchliff et al., 2015).

2.6 Background Analysis

Background effects were tested using the type of vegetation in the image as a proxy for background types. The vegetation types were taken from Wearn et al. (2017), who assessed the habitat structure surrounding the camera traps deployed by the SAFE project on a scale from one to five. Open areas were given a score of 1 and pristine, thick rainforest a score of 5 (see SI 6.1.4).

The VGG-16 architecture was visibly the best performing model in the grouping analysis (Figure 4), and I conducted the background analysis solely on VGG-16. Models were trained using images from one vegetation and evaluated against images from all vegetations. I used the fam-

ily level grouping for this analysis, as the groups are generally larger, the grouping exhibited a high accuracy (Figure 4) and simulates the grouping of other related studies (Norouzzadeh et al., 2018; Tabak et al., 2018). Since the animals present differ between the vegetation types, the models were only trained and evaluated on taxonomic groups present in all vegetations (Table S7).

2.7 Evaluation Metrics

The ability of a deep learning algorithm to correctly classify data can be assessed using various metrics. Two commonly reported metrics are Top-1 accuracy and Top-5 accuracy, which describe the proportion of images a model correctly identified by its top one and top five guesses, respectively (Table 2). In imbalanced datasets, as is the case here, these metrics can be misleading, as a model's inability to classify rare groups can be masked by the ability to correctly identify highly representative classes. Precision and recall can be used to investigate this. Precision assesses a model's ability to avoid labelling negatives as false positives, while recall informs whether the model can correctly detect all positive samples (Table 2). Hence, I report three metrics Top-1 accuracy, Top-5 accuracy and F-1 score, a harmonic mean of precision and recall (Table 2), and a suggested metric for imbalanced datasets (Sokolova & Lapalme, 2009). All metrics can range from 0 to 1, with a score of 1 being desirable.

TABLE 2: Summary of evaluation metrics. *Model performance can be measured by a variety of metrics. I made use of the F-1, Top-1 and Top-5 metrics, which assess model accuracies in slightly different manners. The equations for each metric are given.* tp *are true positives,* tn *are true negatives,* i *refers to an individual class, and* l *is the number of classes.*

Evaluation Metric	Formula	Description / Purpose
Top accuracy	$\frac{tp+tn}{tp+tn+fp+fn}$	Effectiveness of the model to correctly predict class within its first or five best guess
Precision	$\frac{\sum_{i=1}^{l} \frac{tp_i}{tp_i+fp_i}}{l}$	Ability to find only the relevant class
Recall	$\frac{\sum_{i=1}^{l} \frac{tp_i}{tp_i+fn_i}}{l}$	Ability to identify all the relevant class
F-β score	$(\beta^2+1) \times \frac{Precision \times Recall}{Precision+Recall}$	Harmonic mean with a β score of 1 to equally weight precision and recall.

3 RESULTS

3.1 Grouping analysis

A total number of 457,296 images were suitable for the classification specificity analysis. Of these, 61,534 were empty. The number of images for the training and test set were 84,281 and 6,918 for the species, 61,414 and 4,582 for the family and 42,073 and 2,158 for the morphology classification, respectively. There were 49 classes for the species classification, 27 for the family and 12 for the morphological grouping. SI 6.2.1 has a detailed breakdown of the number of images per class for each type of classification.

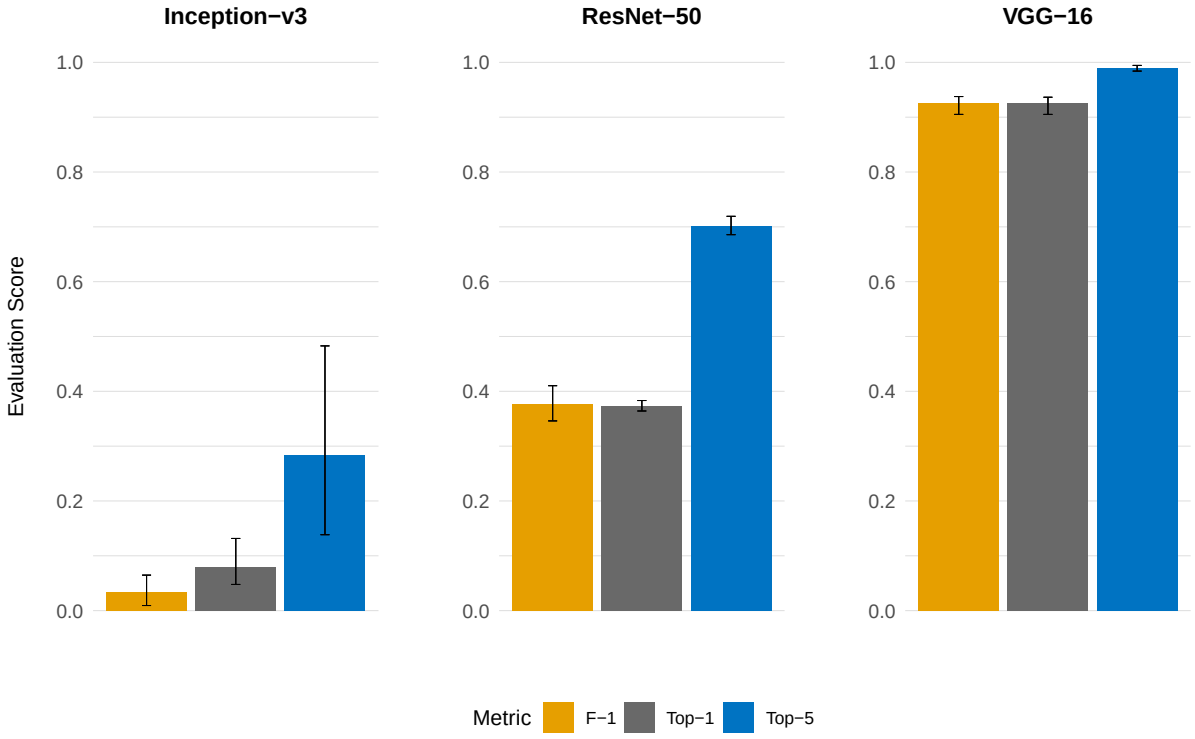

FIGURE 4: Performance of three convolutional neural network architectures. *Performance was assessed by three evaluation metrics, visualised by the bars. The scores displayed are averages of all three grouping types. Error bars signify the range achieved by the three classification types. Large variation was observed for Top-5 accuracy using the Inception-v3 architecture, likely due to the reduction of classes for broader taxonomic groupings.*

As shown in Figure 4, the VGG-16 model clearly outperformed the Inception-v3 and ResNet-50 architectures for all accuracy metrics in all grouping types. Averaging the accuracies of each grouping type VGG-16 reached accuracies of 0.926, 0.925 and 0.990 for the F-1, Top-1 and Top-5 accuracies, respectively. Inception-v3 did not perform notably higher than random in any of the groupings. ResNet-50 exhibited low-to-moderate accuracies, and was on average

0.594%, 0.598%, and 0.292% worse when compared to the F-1, Top-1 and Top-5 scores of the VGG-16, respectively. Based on these results, VGG-16 was chosen as the only architecture suitable to the dataset used here, and subsequent presented analyses were solely conducted on this model.

The VGG-16 achieved the highest F-1 score and Top-1 accuracy on the species level with 0.936 and 0.938, respectively. These two metrics decreased as the breadth of the classification type increased (Figure 5). Compared to the species classification, both the F-1 score and Top-1 accuracy decreased by less than 1% for the family classification and approximately 3.5% for the morphology classification. The Top-5 accuracy peaked at 0.994 on the morphology class, decreasing by less than 1% moving from the morphological to species grouping.

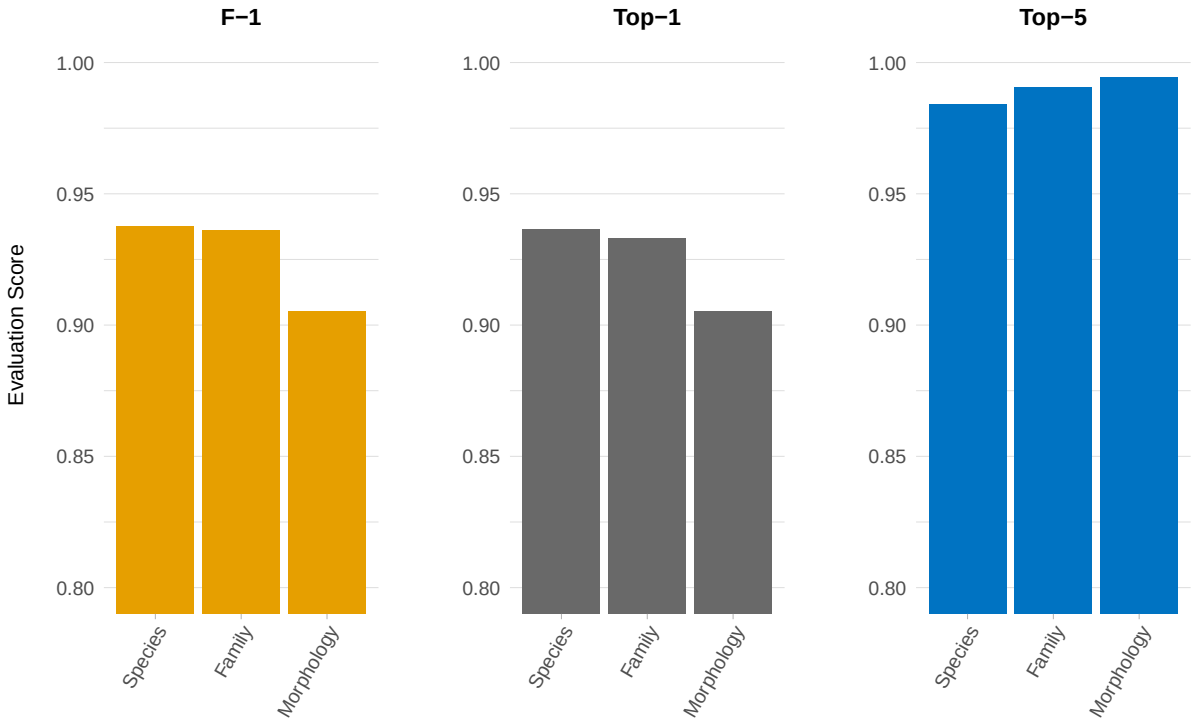

FIGURE 5: Performance of the VGG-16 architecture on three classification types. *Bars within a single graphs show the accuracy for each classification type for a single evaluation metric. Note, the y axis break, which was used to highlight differences in accuracy. No error bars are used, as models were only trained once. A slight decrease is observed for F-1 and Top-1 accuracies moving from species to morphology classification. Negligible increases were observed for the Top-5 accuracy.*

Across all classification types, the 'Empty' group had below average accuracy scores. All but the 'Bay cat' and the 'Pig-tailed maqacue' class exhibited above 80% accuracies scores. Other noticeable low performing groups across classifications were 'Even-toed Ungulates' on

the morphology level, and 'Deer', 'Pigs' and 'Old-world monkeys' on the family level. A full breakdown of the class accuracies, as well as confusion matrices that visualise the model performance, can be found in the supporting information (SI 6.2.1). It is noteworthy that there was no clear increase in per class accuracy for classes with increased number of images (see SI 6.2.3).

3.2 Background analysis

An additional 37,740 images were removed from the dataset, as no vegetation indices were available for some camera trap locations. This resulted in a total of 419,556 images, with 'Empty' images making up 50,060 of images. There were 10 family classes found across all vegetations, which were used as classification groups. The number of images within each vegetation were of comparable sizes (Table S7).

Results of the background analysis are visualised in Figure 6. Training the VGG-16 model on a mixture of vegetation types resulted in high accuracies, with performance decreasing by 10.8%, 10.4% and 1.00% for the F-1, Top-1 and Top-5 metrics, respectively, going from vegetation 1 (Open area) to vegetation 5 (Undisturbed forest). The Top-5 accuracy remained at an almost perfect score for all vegetations. Using images from a single vegetation type as training data and testing against images with the same vegetation resulted in slightly higher accuracies in all metrics, when compared to training the model with images from all vegetations. The difference in accuracy between training a model on a single and mixture of vegetations, and the changes in performances depending on the vegetation type used in the test set are fully illustrated in Figure S5.

Training the model with a single vegetation type, and testing the model on images from different vegetations heavily affected model performance negatively across all metrics. The F-1, Top-1 and Top-5 score dropped by 63.8%, 58.6% and 16.8% respectively on average (using the model performance when trained and tested on the same vegetation type as reference accuracies). These negative effects were dampened when the model was trained on vegetation type 3 or 5, with decreases of 54.1%, 47.8% and 13.7% for vegetation 3, and 56.7%, 51.7% and 13.4% for vegetation 5 in F-1, Top-1 and Top-5 accuracy, respectively.

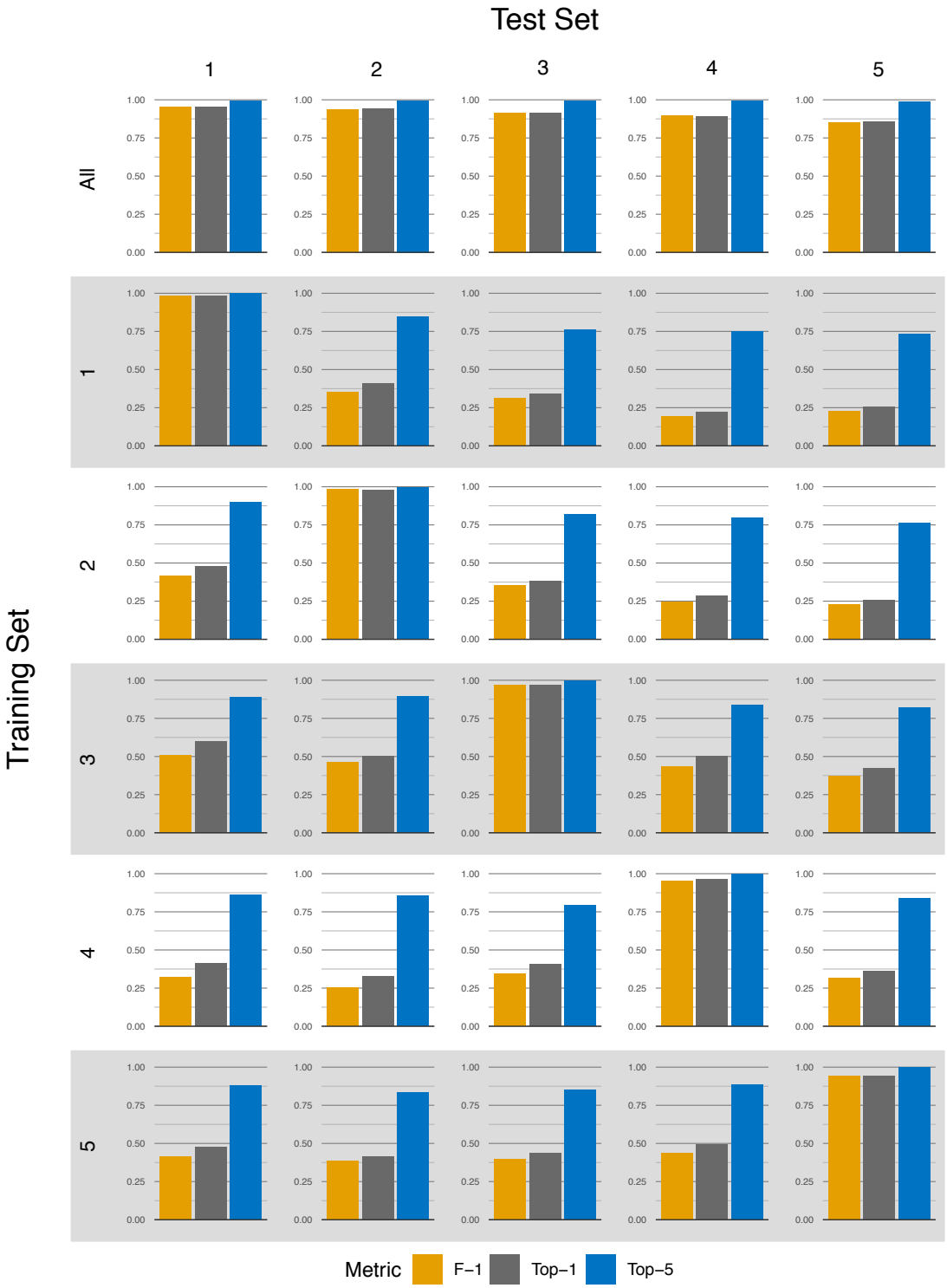

FIGURE 6: Background effects on VGG-16 identification ability. *Numberings indicated the vegetation type, going from open areas (1) to Undisturbed forest (5). Bars show the score obtained for the three evaluation metrics. Training on a mixture of backgrounds allows the model to accurately identify objects, as given by both the Top-1 and F-1 score. However, if trained on a single background, only test data on the same background allows for similar performance (diagonal). Testing ConvNets on vegetation types different from the training vegetation type, greatly reduces accuracies. Small differences in the Top-5 accuracy are still of large significance, since there are only 10 classes in total.*

The worst model performances were observed when the model was trained on images from vegetation 1 and tested on images from other vegetation types, with an average drop in accuracy of 72.2%, 68.5% and 22.6% for F-1, Top-1 and Top-5 score, respectively. However, vegetation 1 exhibited the lowest performance drops when used as a test set. With an exception of the Top-5 metric between vegetation 3 and vegetation 4 (Disturbed forest), the model performance dropped increasingly on average by 13.6%, 17.7% and 9.29% for F-1, Top-1 and Top-5, respectively, going from vegetation 1 to vegetation 5 as test sets, i.e. from open areas to undisturbed forest. A visualisation and breakdown of the percentage accuracy drops can be seen in Figure S4.

4 DISCUSSION

4.1 Findings and implications

This is one of the first studies investigating the underlying assumptions associated with using convolutional neural networks on ecological datasets. Importantly, the results reveal several limitation affecting automatic object identification methods in biomonitoring.

VGG-16 was the only model to fit the SAFE camera trap dataset with industry standard accuracies, while the Inception-v3 and ResNet-50 architectures exhibited model over-fitting. These findings are complemented by Nguyen et al. (2017), who observed reduced accuracies for the ResNet-50 architecture when applied to an ecological dataset utilising a transfer learning approach. Both Inception-v3 and ResNet-50 make use of so-called normalisation layers (Szegedy et al., 2015; He et al., 2015), which may explain the observed phenomenon. Normalisation layers attempt to optimise network performance by normalising layer inputs (Ioffe & Szegedy, 2015), but appear to function poorly in conjunction with transfer learning, possibly due to improper implementation in the machine learning library used here (Vryniotis, 2017). While this hindered an extensive comparison of ConvNet architectures as a consequence, it demonstrates that ConvNets need to be uniquely adjusted to datasets through the use of different hyper-parameters.

All grouping scenarios achieved accuracies comparable with those achieved by recent applications of ConvNets to ecological datasets (Nguyen et al., 2017; Norouzzadeh et al., 2018;

Tabak et al., 2018). Contrary to my expectations, species-level classification resulted in the best model performance with surprisingly high accuracies for rare classes. The transfer learning approach has shown to greatly facilitate ConvNet performance on small image datasets (Dabrowski & Michalik, 2017), so much that it allows for 93.0% identification accuracies on ecological camera trap dataset with less than 1000 labelled images for 20 classes (Schneider et al., 2018). Overall, these results show promise for ecological studies with small datasets.

Broader classification types exhibited lower, but very similar accuracies, suggesting that grouping breadth can be altered to suit the needs of different research. Classes containing deer and pigs were among the lowest performing groups and had a greater influence on model performance when broader classification were used, due to the decrease in total number of grouping categories. The effect of these classes is likely to have been exacerbated for the even-toed ungulates, due to high phenotypic variation within the group. Slight difference in performance between groupings may have also been observed because groups contained different images of the same species, as they were randomly selected and capped at 5000 images. Gomez et al. (2016) previously demonstrated above 90% accuracies distinguishing animals between birds, large- and small mammals using a ConvNet. Combined with my findings, broad taxonomic groups are a feasible, and possibly better option for integrating rare species into ConvNet identification workflows (e.g. grouping rare and poor performing bay cat with other cats increased F-1 score from 0.88 to 0.97, respectively). Broad classifications would thereby allow for the initial segmenting of data with high accuracies, and subsequent manual reviewing to select for focal species of interest.

The background analysis revealed high scene specificity. When training and testing on images with different background vegetation, the negative effects were reduced either by training the ConvNet on images containing undisturbed and heavily-disturbed forest habitats, or using images with open areas for the test set. Still, it is evident that in order to achieve industry standard accuracies requires training sets to include a large variety of background types, or images from the same environment as the test images. This likely explains the observed reduction of accuracy by Tabak et al. (2018) when applying their trained ConvNet on images containing the same species, but in different environments.

While the classes for the scenery analysis were standardised on the family level, some species are not found within all vegetation types, causing slight variation on the species level. This may have exacerbated the drop in accuracy, as the two vegetations that contained the most species within them performed slightly better when tested on images from other vegetations. Given that there were only 10 classes, however, the drop in Top-5 accuracy indicates that background is indeed having a notable effect. Furthermore, if the presence or absence of species heavily affected accuracies, we would expect the ConvNet to perform better when trained and tested on vegetations with more overlapping species, which is not the case here. Nevertheless, overfitting may have occurred during training to a small degree. During later epochs, model weights resulting in negligible accuracy improvements were often saved, possibly causing species to be identified in the context of their environments.

Regardless, scenery effects have been observed outside of ecological datasets as well (Zhu et al., 2014; Zhang et al., 2015) and increasing attempts have been made to develop 'background subtraction' methods that reduce the noise of background scenery (Giraldo-Zuluaga et al., 2017; Janzen et al., 2017; Yousif et al., 2017a,b). Given my findings, it is likely that such methods are to be a necessity for widely applicable automatic object detection tools in ecology. Publicly available tools, such as those made by Tabak et al. (2018) which do not incorporate background removal techniques, will only be applicable to areas of similar habitats.

Overall, my findings reveal that qualitative features of images within training sets heavily affect accuracies and may be more important than the actual size of datasets. With more research into such factors, and with increased efforts to make ConvNets more user friendly (Wäldchen et al., 2018), deep neural networks have tremendous potential to increase the scope of camera trap based research, mainly through the immense time savings they produce compared to manual classification (Norouzzadeh et al., 2018). Machine learning tools can even be installed on camera traps to immediately assess images taken (Elias et al., 2017), which would allow solely images of interest to be saved, and thereby greatly reducing the frequency at which data needs to be collected by field ecologists.

With the number of taxonomists and identification experts decreasing (Hopkins & Freckleton, 2002), automation techniques will become increasingly more important, particularly outside

of image based mammal studies (Wäldchen & Mäder, 2018). ConvNets models have been used as identification tools for insects (Cheng et al., 2017), marine animals (Qin et al., 2016; Salman et al., 2016) and plants (Rahnemoonfar & Sheppard, 2017; Sun et al., 2017), as well as for video footage (French et al., 2015) and audio data (Sprengel et al., 2016; Mac Aodha et al., 2018). However, the need for greater collaboration and data sharing in ecological research cannot be understated (Steenweg et al., 2017) and is highlighted here as well. Global datasets representative of most habitats and species will be required in order for ConvNets to become a versatile tool for biomonitoring.

4.2 Caveats and limitations

Computational intensity heavily limits the research of ConvNets (LeCun et al., 2015), but was largely overlooked by this study. Depending on the architecture, ConvNets can reach a magnitude of 10^8 parameters (Table S2), which need to be tuned. This makes training computationally expensive, and models were therefore only trained once, hindering further statistical analyses and inferences. The number of hyper-parameters tested was also restricted for the same reason, and may have resulted in sub-optimal model performances, particularly for the Inception-v3 and ResNet-50 architectures.

An unusual finding here, was the absence of accuracy improvement with increasing number of images (Figure S6), which is commonly the case for object identification problems within the deep learning community (Goodfellow et al., 2016). Smaller groups generally performed better than groups that were limited to 5000 images, which influenced the average accuracies achieved by the models trained. This is likely to be the case, because small classes have few trigger events, so that images will be very similar in the training and test set, and a ConvNet can more easily identify the animal. Future work should, therefore, use one single image from each single trigger event. This would probably cause the species level classification to drop in accuracy, which would further support grouping animals into broader classes.

The capping is likely to have influenced the grouping analysis. Since images were randomly allocated to classes, images of rare species, which have a smaller chance of being selected, may have not been included in broader groupings. Preferably, rare species should have been prioritised when assigning images to groups, or a different approach tackling the class imbal-

ance could have been used (He & Garcia, 2009). For example, oversampling is a method where images from a rare class are used multiple times during training, and has been shown to be effective in ecological dataset (Norouzzadeh et al., 2018).

4.3 Outlook and future work

The work presented here is a start at investigating the qualitative features of images in camera trap datasets that affect model performance. However, given the time constraint of this study, I was unable to further investigate whether animals were identified by themselves or in the context of their environments. To build on this study, it would be beneficial to visualise the features within images that prompted the network to assign an object to a specific class. Class activation maps are graphics that allow for this, by overlaying an image with a heatmap of the pixels involved in the classification, and should be generated in future work.

Further investigation in regards to different architectures, features and parameters is also needed. While the three ConvNet architectures studied here have been most commonly used for ecological datasets, many more architectures are publicly available. Additionally, a major extension to the grouping analysis of this work would be testing the ability of ConvNets trained on broad taxonomic classes to identify species it had not encountered before. Lastly, comparisons of model performance with respect to different hyper-parameters were not conducted for this investigation. Altering the optimisation algorithm, image sizes, learning rate dynamic and other parameters could greatly affect the ability of ConvNets to identify objects, particularly for architectures such as those that performed poorly on the dataset used here.

5 CONCLUSION

ConvNets have shown great potential for automatic object identification in image based ecological research, but still largely operate as a black box. The objective of this study was to assess the ability of ConvNets to identify wildlife in camera trap images with respect to the breadth of classification groups and background scenery. The results showed similar accuracies for all groupings with a slight decrease for broader classifications. Performance was heavily dependent on the type of background vegetation a ConvNet was parameterised and tested on. Based on these findings, I recommend that researchers with limited data, either train ConvNets

using a transfer learning approach, or utilise broadly defined classes. Additionally, camera trap studies should only utilise pre-trained ConvNets in their research, if they have been trained on images from similar habitats. Publicly made available ConvNets classifiers will need to be trained on a large variation of background vegetations in order for them to be widely utilised. Future work should continue investigating the qualitative features of camera trap images that affect ConvNet performance. Determining the underlying assumptions and limitations of ConvNets will contribute to the automation of biomonitoring, and thereby greatly focus and facilitate conservation efforts.

ACKNOWLEDGEMENTS

First and foremost, I extend my gratitude to Dr Phil Chapman and Dr Oliver Wearn for providing me with their camera trap data. I also thank Sarab Sethi for helping shape the topic of this thesis. Next, I would like to again express my gratitude to Dr Oliver Wearn for providing extensive ecological insight throughout the project. Last but not least, I wish to thank Dr James Rosindell for continued input and guidance.

References

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., Kudlur, M., Levenberg, J., Monga, R., Moore, S., Murray, D.G., Steiner, B., Tucker, P., Vasudevan, V., Warden, P., Wicke, M., Yu, Y. & Zheng, X. (2016) Tensorflow: A system for large-scale machine learning. *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)*, 265–283. http://arxiv.org/abs/1605.08695

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**, 675–685. https://doi.org/10.1111/1365-2664.12432

Chao, A. (1989) Estimating population size for sparse data in capture-recapture experiments. *Biometrics*, **45**, 427–438. https://doi.org/10.2307/2531487

Chen, G., Han, T.X., He, Z., Kays, R. & Forrester, T. (2014) Deep convolutional neural network based species recognition for wild animal monitoring. *IEEE International Conference on Image Processing (ICIP)*, 858–862. https://doi.org/10.1109/ICIP.2014.7025172

Cheng, X., Zhang, Y., Chen, Y., Wu, Y. & Yue, Y. (2017) Pest identification via deep residual learning in complex background. *Computers and Electronics in Agriculture*, **141**, 351–356. https://doi.org/10.1016/j.compag.2017.08.005

Chollet, F. et al. (2015) Keras. GitHub repository. https://github.com/fchollet/keras

Dabrowski, M. & Michalik, T. (2017) How effective is transfer learning method for image classification. *Federated Conference on Computer Science and Information Systems*, 3–9. https://doi.org/10.15439/2017F526

Dauphin, Y.N., Pascanu, R., Gulcehre, C., Cho, K., Ganguli, S. & Bengio, Y. (2014) Identifying and attacking the saddle point problem in high-dimensional non-convex optimization. *Advances in neural information processing systems*, 2933–2941. http://dl.acm.org/citation.cfm?id=2969033.2969154

Dickinson, J.L., Zuckerberg, B. & Bonter, D.N. (2010) Citizen science as an ecological

research tool: challenges and benefits. *Annual review of ecology, evolution, and systematics*, **41**, 149–172. https://doi.org/10.1146/annurev-ecolsys-102209-144636

Elias, A.R., Golubovic, N., Krintz, C. & Wolski, R. (2017) Where's the bear?-automating wildlife image processing using iot and edge cloud systems. *Proceedings of the Second International Conference on Internet-of-Things Design and Implementation*, 247–258. IEEE. https://doi.org/10.1145/3054977.3054986

Federhen, S. (2011) The ncbi taxonomy database. *Nucleic acids research*, **40**, D136–D143. https://doi.org/10.1093/nar/gkr1178

French, G., Fisher, M., Mackiewicz, M. & Needle, C. (2015) Convolutional neural networks for counting fish in fisheries surveillance video. *Proceedings of the Machine Vision of Animals and their Behaviour (MVAB)*, 7.1–7.10. https://dx.doi.org/10.5244/C.29.MVAB.7

Giraldo-Zuluaga, J.H., Salazar, A., Gomez, A. & Diaz-Pulido, A. (2017) Camera-trap images segmentation using multi-layer robust principal component analysis. *The Visual Computer*, 1–13. https://doi.org/10.1007/s00371-017-1463-9

Gomez, A., Diez, G., Salazar, A. & Diaz, A. (2016) Animal identification in low quality camera-trap images using very deep convolutional neural networks and confidence thresholds. *International Symposium on Visual Computing*, 747–756. https://doi.org/10.1007/978-3-319-50835-1_67

Goodfellow, I., Bengio, Y. & Courville, A. (2016) *Deep Learning*. MIT Press. http://www.deeplearningbook.org

He, H. & Garcia, E.A. (2009) Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering*, **21**, 1263–1284. https://doi.org/10.1109/TKDE.2008.239

He, K., Zhang, X., Ren, S. & Sun, J. (2015) Deep residual learning for image recognition. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778. https://doi.org/10.1109/CVPR.2016.90

Hinchliff, C.E., Smith, S.A., Allman, J.F., Burleigh, J.G., Chaudhary, R., Coghill, L.M., Crandall, K.A., Deng, J., Drew, B.T., Gazis, R., Gude, K., Hibbett, D.S., Katz, L.A., Laughinghouse,

549 H.D., McTavish, E.J., Midford, P.E., Owen, C.L., Ree, R.H., Rees, J.A., Soltis, D.E.,
550 Williams, T. & Cranston, K.A. (2015) Synthesis of phylogeny and taxonomy into a
551 comprehensive tree of life. *Proceedings of the National Academy of Sciences*, **112**,
552 12764–12769. https://doi.org/10.1073/pnas.1423041112

553 Hopkins, G. & Freckleton, R.P. (2002) Declines in the numbers of amateur and professional
554 taxonomists: implications for conservation. *Animal Conservation*, **5**, 245–249.
555 https://doi.org/10.1017/S1367943002002299

556 Hubel, D.H. & Wiesel, T.N. (1962) Receptive fields, binocular interaction and functional
557 architecture in the cat's visual cortex. *The Journal of physiology*, **160**, 106–154.
558 https://doi.org/10.1113/jphysiol.1962.sp006837

559 Ioffe, S. & Szegedy, C. (2015) Batch normalization: Accelerating deep network training by
560 reducing internal covariate shift. *arXiv preprint*. http://arxiv.org/abs/1502.03167

561 Janzen, M., Visser, K., Visscher, D., MacLeod, I., Vujnovic, D. & Vujnovic, K. (2017)
562 Semi-automated camera trap image processing for the detection of ungulate fence crossing
563 events. *Environmental monitoring and assessment*, **189**, 527.
564 https://doi.org/10.1007/s10661-017-6206-x

565 Kingma, D.P. & Ba, J. (2014) Adam: A method for stochastic optimization. *arXiv preprint*.
566 http://arxiv.org/abs/1412.6980

567 Krizhevsky, A., Sutskever, I. & Hinton, G.E. (2012) Imagenet classification with deep
568 convolutional neural networks. *Advances in neural information processing systems*,
569 1097–1105. http://papers.nips.cc/paper/
570 4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf

571 Kucera, T.E. & Barrett, R.H. (2011) A history of camera trapping. *Camera traps in animal*
572 *ecology*, 9–26. https://doi.org/10.1007/978-4-431-99495-4_2

573 LeCun, Y., Bengio, Y. & Hinton, G. (2015) Deep learning. *Nature*, **521**, 436.
574 https://doi.org/10.1038/nature14539

575 LeCun, Y., Bengio, Y. et al. (1995) Convolutional networks for images, speech, and time
576 series. *The handbook of brain theory and neural networks*, **3361**, 1995.
577 http://dl.acm.org/citation.cfm?id=303568.303704

578 Mac Aodha, O., Gibb, R., Barlow, K.E., Browning, E., Firman, M., Freeman, R., Harder, B.,
579 Kinsey, L., Mead, G.R., Newson, S.E., Pandourski, I., Parsons, S., Russ, J.,
580 Szodoray-Paradi, A., Szodoray-Paradi, F., Tilova, E., Girolami, M., Brostow, G. & Jones,
581 K.E. (2018) Bat detective—deep learning tools for bat acoustic signal detection. *PLoS*
582 *computational biology*, **14**, e1005995. https://doi.org/10.1371/journal.pcbi.1005995

583 Marcus Rowcliffe, J. (2017) Key frontiers in camera trapping research. *Remote Sensing in*
584 *Ecology and Conservation*, **3**, 107–108. https://doi.org/10.1002/rse2.65

585 McCallum, J. (2013) Changing use of camera traps in mammalian field research: habitats,
586 taxa and study types. *Mammal Review*, **43**, 196–206.
587 https://doi.org/10.1111/j.1365-2907.2012.00216.x

588 Nguyen, H., Maclagan, S.J., Nguyen, T.D., Nguyen, T., Flemons, P., Andrews, K., Ritchie, E.G.
589 & Phung, D. (2017) Animal recognition and identification with deep convolutional neural
590 networks for automated wildlife monitoring. *Data Science and Advanced Analytics (DSAA),*
591 *2017 IEEE International Conference on*, 40–49. https://doi.org/10.1109/DSAA.2017.31

592 Nichols, J.D. & Williams, B.K. (2006) Monitoring for conservation. *Trends in ecology &*
593 *evolution*, **21**, 668–673. https://doi.org/10.1016/j.tree.2006.08.007

594 Norouzzadeh, M.S., Nguyen, A., Kosmala, M., Swanson, A., Palmer, M.S., Packer, C. &
595 Clune, J. (2018) Automatically identifying, counting, and describing wild animals in
596 camera-trap images with deep learning. *Proceedings of the National Academy of Sciences*,
597 **115**, E5716–E5725. https://doi.org/10.1073/pnas.1719367115

598 Parr, C.S., Wilson, N., Leary, P., Schulz, K.S., Lans, K., Walley, L., Hammock, J.A., Goddard,
599 A., Rice, J., Studer, M., Holmes, J.T. & Corrigan Jr., R.J. (2014) The encyclopedia of life v2:
600 providing global access to knowledge about life on earth. *Biodiversity Data Journal*.
601 https://doi.org/10.3897/BDJ.2.e1079

602 Pimm, S.L., Alibhai, S., Bergl, R., Dehgan, A., Giri, C., Jewell, Z., Joppa, L., Kays, R. &
603 Loarie, S. (2015) Emerging technologies to conserve biodiversity. *Trends in ecology &*
604 *evolution*, **30**, 685–696. https://doi.org/10.1016/j.tree.2015.08.008

605 Qin, H., Li, X., Liang, J., Peng, Y. & Zhang, C. (2016) Deepfish: Accurate underwater live fish

recognition with a deep architecture. *Neurocomputing*, **187**, 49–58. https://doi.org/10.1016/j.neucom.2015.10.122

Rahnemoonfar, M. & Sheppard, C. (2017) Deep count: fruit counting based on deep simulated learning. *Sensors*, **17**, 905. https://doi.org/10.3390/s17040905

Rosindell, J. & Harmon, L.J. (2012) Onezoom: a fractal explorer for the tree of life. *PLoS biology*, **10**, e1001406. https://doi.org/10.1371/journal.pbio.1001406

Roskov, Y., Kunze, T., Paglinawan, L., Orrell, T., Nicolson, D., Culham, A., Bailly, N., Kirk, P., Bourgoin, T., Baillargeon, G. et al. (2013) Species 2000 & ITIS catalogue of life. *Catalogue of Life Annual Checklist*. http://www.catalogueoflife.org/annual-checklist/2013/

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C. & Fei-Fei, L. (2015) Imagenet large scale visual recognition challenge. *International Journal of Computer Vision (IJCV)*, **115**, 211–252. https://doi.org/10.1007/s11263-015-0816-y

Salman, A., Jalal, A., Shafait, F., Mian, A., Shortis, M., Seager, J. & Harvey, E. (2016) Fish species classification in unconstrained underwater environments based on deep learning. *Limnology and Oceanography: Methods*, **14**, 570–585. https://doi.org/10.1002/lom3.10113

Sauermann, H. & Franzoni, C. (2015) Crowd science user contribution patterns and their implications. *Proceedings of the National Academy of Sciences*, **112**, 679–684. https://doi.org/10.1073/pnas.1408907112

Schneider, S., Taylor, G.W. & Kremer, S.C. (2018) Deep learning object detection methods for ecological camera trap data. *arXiv preprint*. http://arxiv.org/abs/1803.10842

Simonyan, K. & Zisserman, A. (2014) Very deep convolutional networks for large-scale image recognition. *arXiv preprint*. http://arxiv.org/abs/1409.1556

Sokolova, M. & Lapalme, G. (2009) A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, **45**, 427–437. https://doi.org/10.1016/j.ipm.2009.03.002

633 Spampinato, C., Farinella, G.M., Boom, B., Mezaris, V., Betke, M. & Fisher, R.B. (2015)
634 Special issue on animal and insect behaviour understanding in image sequences.
635 *EURASIP Journal on Image and Video Processing*, **2015**, 1.
636 https://doi.org/10.1186/1687-5281-2015-1

637 Sprengel, E., Jaggi, M., Kilcher, Y. & Hofmann, T. (2016) Audio based bird species
638 identification using deep learning techniques. *Working Notes of CLEF (Cross Language*
639 *Evaluation Forum)*, 547–559

640 Steenweg, R., Hebblewhite, M., Kays, R., Ahumada, J., Fisher, J.T., Burton, C., Townsend,
641 S.E., Carbone, C., Rowcliffe, J.M., Whittington, J. et al. (2017) Scaling-up camera traps:
642 monitoring the planet's biodiversity with networks of remote sensors. *Frontiers in Ecology*
643 *and the Environment*, **15**, 26–34. https://doi.org/10.1002/fee.1448

644 Sun, Y., Liu, Y., Wang, G. & Zhang, H. (2017) Deep learning for plant identification in natural
645 environment. *Computational intelligence and neuroscience*, **2017**.
646 https://doi.org/10.1155/2017/7361042

647 Swann, D.E., Kawanishi, K. & Palmer, J. (2011) Evaluating types and features of camera traps
648 in ecological studies: A guide for researchers. *Camera traps in animal ecology*, 27–43.
649 Springer. https://doi.org/10.1007/978-4-431-99495-4_3

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

653 Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J. & Wojna, Z. (2015) Rethinking the inception
654 architecture for computer vision. *Proceedings of the IEEE conference on computer vision*
655 *and pattern recognition*, 2818–2826. http://arxiv.org/abs/1512.00567

656 Tabak, M.A., Norouzzadeh, M.S., Wolfson, D.W., Sweeney, S.J., VerCauteren, K.C., Snow,
657 N.P., Halseth, J.M., Di Salvo, P.A., Lewis, J.S., White, M.D., Teton, B., Beasley, J.C.,
658 Schlichting, P.E., Boughton, R.K., Wight, B., Newkirk, E.S., Ivan, J.S., Odell, E.A., Brook,
659 R.K., Lukacs, P.M., Moeller, A.K., Mandeville, E.G., Clune, J. & Miller, R.S. (2018) Machine
660 learning to classify animal species in camera trap images: applications in ecology. *bioRxiv*,
661 346809. https://doi.org/10.1101/346809

Villa, A.G., Salazar, A. & Vargas, F. (2017) Towards automatic wild animal monitoring: Identification of animal species in camera-trap images using very deep convolutional neural networks. *Ecological Informatics*, **41**, 24–32. https://doi.org/10.1016/j.ecoinf.2017.07.004

Vryniotis, V. (2017) Change BN layer to use moving mean/var if frozen. GitHub repository. https://github.com/keras-team/keras/pull/9965

Wäldchen, J. & Mäder, P. (2018) Machine learning for image based species identification. *Methods in Ecology and Evolution*, **0**. https://doi.org/10.1111/2041-210X.13075

Wäldchen, J., Rzanny, M., Seeland, M. & Mäder, P. (2018) Automated plant species identification—Trends and future directions. *PLoS computational biology*, **14**, e1005993. https://doi.org/10.1371/journal.pcbi.1005993

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**, 162–171. https://doi.org/10.1016/j.biocon.2017.05.007

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

Yousif, H., He, Z. & Kays, R. (2017a) Object segmentation in the deep neural network feature domain from highly cluttered natural scenes. *IEEE International Conference on Image Processing (ICIP)*, 3095–3099. https://doi.org/10.1109/ICIP.2017.8296852

Yousif, H., Yuan, J., Kays, R. & He, Z. (2017b) Fast human-animal detection from highly cluttered camera-trap images using joint background modeling and deep learning classification. *IEEE International Symposium on Circuits and Systems (ISCAS)*, 1–4. https://doi.org/10.1109/ISCAS.2017.8050762

Zhang, C., Li, H., Wang, X. & Yang, X. (2015) Cross-scene crowd counting via deep convolutional neural networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 833–841. https://doi.org/10.1109/CVPR.2015.7298684

Zhu, W., Liang, S., Wei, Y. & Sun, J. (2014) Saliency optimization from robust background

detection. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2814–2821. https://doi.org/10.1109/CVPR.2014.360

6 Supporting Information

6.1 Methodology

6.1.1 Data cleanup

The SAFE project camera trap dataset included several non-species-specific animal tags, which had to be excluded from the analysis. While they could have been used in the analysis of higher morphological classes, I decided to discard these tags for simplicity, as they only made up a fraction of the images ($\sim$0.01%). If a tag was non-species specific, but was the lowest taxonomic rank for the taxa in the dataset and contained more than 20 images, I did not remove the group (for example Maxomys). This was to allow for enough images to capture phenotypic variation within the group. The only non-species-specific group that had more than 20 images but was still removed was ants as they are too small to identify by the model. Images labelled with the following non-species-specific tags were removed:

- Bat
- Bird sp.
- Cat
- Civet
- Dragonfly
- Eagle
- Frog
- Mousedeer sp.
- Muntjac sp.
- Owl
- Pill Millipede
- Pitta
- Pitta sp.
- Porcupine sp.
- Rat sp.
- Skink
- Spiny Rat
- Squirrel
- Treeshrew sp.
- Unidentified Treeshrew

I additionally removed any any reptilian species, to only investigate birds and mammals. Reptiles only had a total of 90 images and included three species (Asian Tortoise, Reticulated Python, and Roughneck Monitor Lizard). Some species had multiple tags associated with them. I therefore standardised the following duplicate tags:

- Banded palm civet to Banded civet
- Coucal to Greater coucal
- Dog to Domestic dog
- Malay civet to Malayan civet
- Malay porcupine to Malayan porcupine
- Magpie robin to Oriental magpie robin
- Orangutan to Bornean orangutan
- Tembedau to Banteng

6.1.2 Training and test set split

Datasets are usually split into 80-20% or 75-25% training and test set, respectively. Here, I followed the advice of the ZSL researchers to create custom splits, as summarised in Table S1. Test sets were kept imbalanced, as they represent a true sample of images you would obtain from a camera trap survey.

TABLE S1: Training and test set split of total dataset. *Four different types of splits were performed, depending on the total number of images within a class.*

Total Class Size (N)	Training Set Size	Test Set Size
$N \geq 5200$	5000	200
$5200 > N \geq 800$	$N - 200$	200
$800 > N \geq 66$	$N \times 0.75$	$N \times 0.25$
$66 > N \geq 60$	50	$N - 50$

6.1.3 Animal Grouping

Phylogenetic relatedness was higher within groups than between groups, except for the rodents class. The moon rat (*Echinosorex gymnura*) is most closely related to tree-shrews, however, due to its morphological features I decided to class it with rodents. Humans were used as a separate class, instead of grouping them with other primates. Birds are typically not identified in camera-trapping studies (often focussed on mammals), and some bird species can be difficult to identify from images alone. They were therefore classed by order or class.

6.1.4 Vegetation Score

The vegetation indices were taken from Wearn et al. (2017), who assessed the vegetation types at the SAFE study site. The *ad verbum* descriptions are as follows:

1. **Open area**: Dominated by grasses and small shrubs (<1 m in height). Typically on logging roads or old log landing areas.

2. **Herbaceous scrub**: Dominated by herbs (typically Zingiberaceae), vines and shrubs, with no trees >3 m in height (except oil palm Elaeis guineensis). Typically representing secondary re-growth from clear-felling, or large gaps due to landslides.

3. **Heavily-disturbed forest**: High scrub or dense understorey layer (typically with vines and Dinochloa climbing bamboo species), with a low, heavily-broken canopy layer (<20 m). Possibly some large isolated trees (>20 m). Intensively-logged area or large gap disturbance.

4. **Disturbed forest**: Mostly pioneer tree species (typically Macaranga species), but some old-growth dipterocarp species may be present. Discontinuous canopy. Lower intensity of logging or natural disturbance.

5. **Undisturbed forest**: Dominated by old-growth dipterocarps. High, continuous canopy with sparsely-vegetated understorey. Unlogged, with little recent disturbance evident.

6.1.5 Model Architectures

TABLE S2: Description of model architectures used. *Architecture varies in number of layers, the type of layers and the order of structure, which changes the number of parameters needed to be tuned. All of the models are placed highly in the ImageNet competition, and are considered to be top performing architectures.*

Model	No. of Parameters	Description
Inception-v3	23,851,784	Makes use of batch normalisation layers, and a combination of 1×1, 3×3 and 5×5 kernels.
ResNet-50	25,636,712	Outputs bypass the inputs of the next layers and are fed back into the model two consecutive layers. Also makes use of batch normalisation layers.
VGG-16	138,357,544	Each convolutional layers uses a small 3×3 convolutional filters.

6.1.6 Computing Languages

TABLE S3: Overview of computing software and packages used. *Each language and package use is shown with the version under which the analysis was run. A short description is given to programs and packages used, detailing the tasks for which they were used for.*

Computing Language	Programs used
Bash v.3.2.57	• macOS 'sips' v.10.12.6: Resize images • exiftool v.11.08: Extract image metadata
Python v.3.6.5	• pandas v.0.23.0: Data manipulation and wrangling • Keras v.2.1.6 (Chollet et al., 2015): High-level neural network library • TensorFlow v.1.8.0 (Abadi et al., 2016): Open-source machine learning framework
R v.3.4.3	• taxize v.0.9.3: Obtain taxonomic information • dplyr v.0.7.4: Data manipulation and wrangling • ggplot v.2.2.1: Creation of plots

6.1.7 Code Repository

The code used to run this analysis can be accessed on Box at:

`https://imperialcollegelondon.box.com/v/phb13-msc-thesis-17-18`

6.2 Results

6.2.1 Grouping Analysis

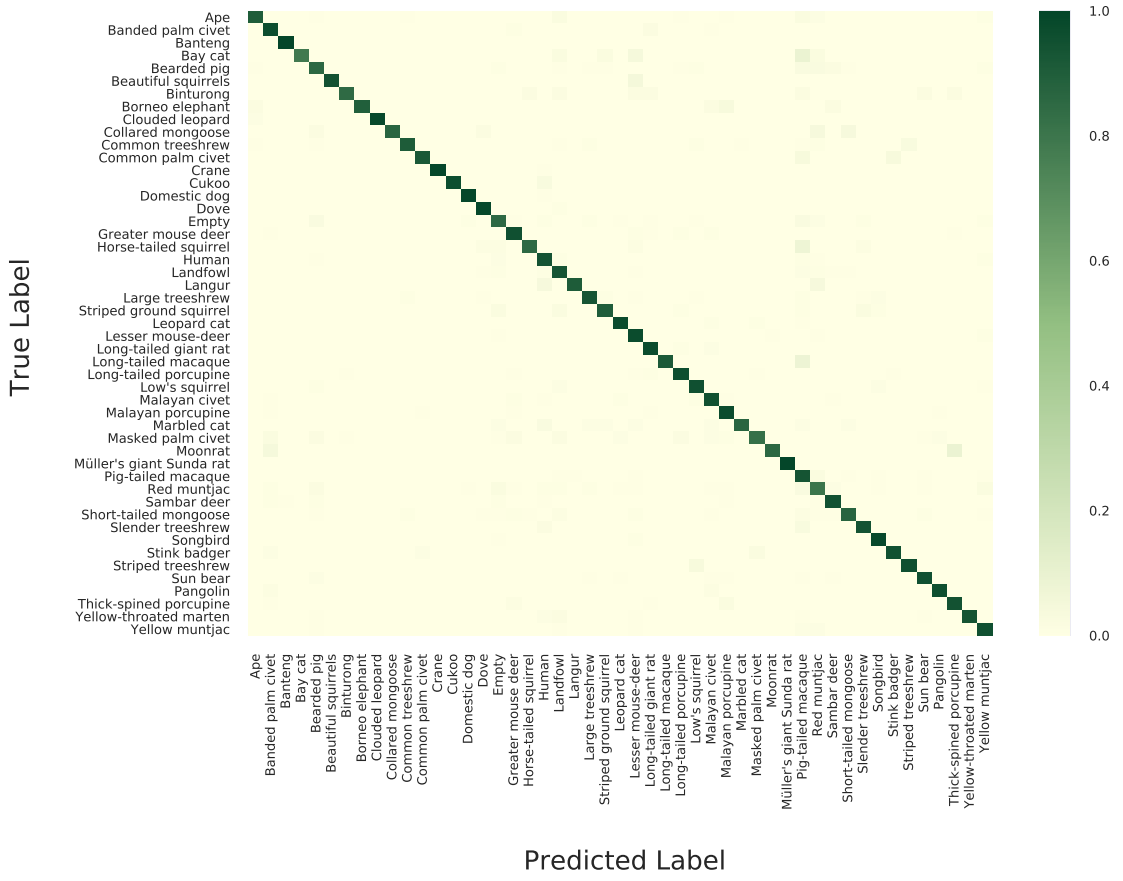

FIGURE S1: Confusion matrix of species grouping. *The confusion matrix is a visualisation of the accuracy of the convolutional neural network. The colour intensity indicates the proportion of classifications made for each respective class. The diagonal shows correct predictions. In general, the ConvNet exhibited a high accuracy, with most of the area being of yellow colour. Pig-tailed macaque, however, was often misclassified for example.*

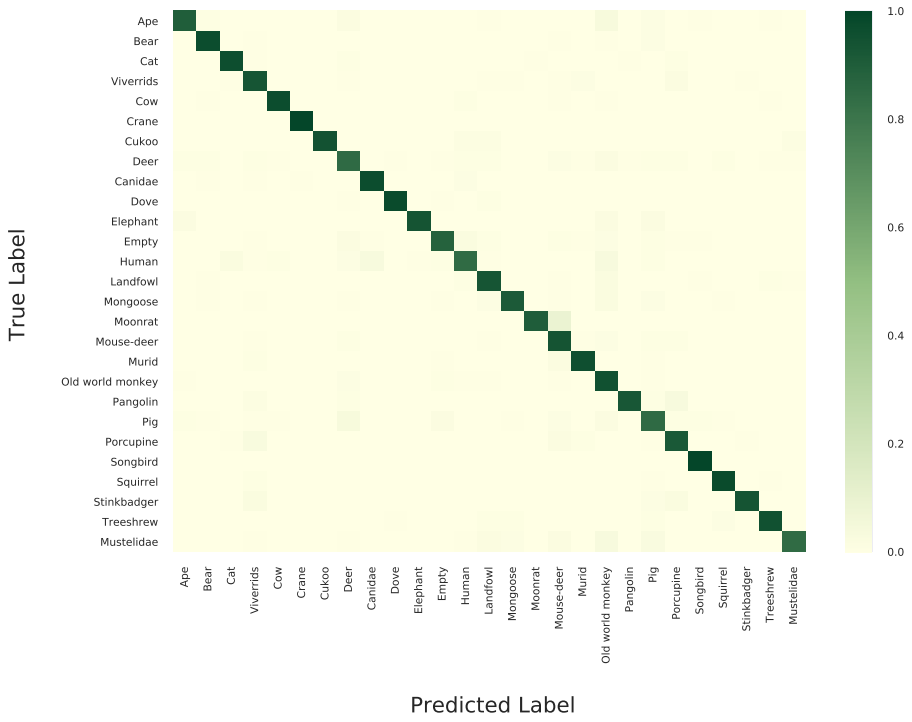

FIGURE S2: Confusion matrix of family grouping. *A visualisation of the accuracy of the ConvNet for each class, as given by the colouring intensity. The diagonal shows correct predictions. Old-world monkeys were often misclassified for example.*

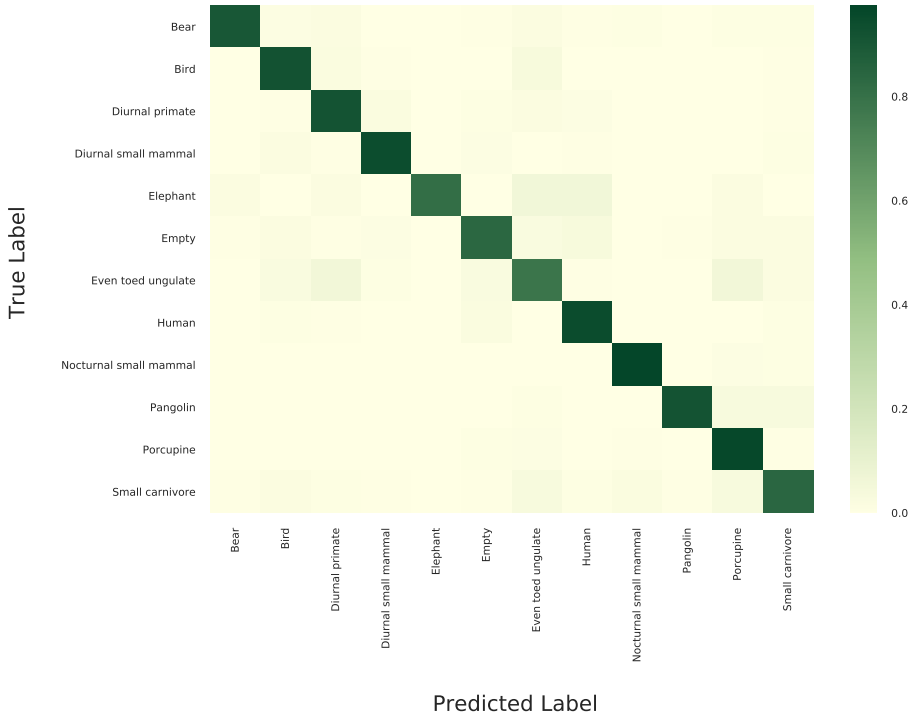

FIGURE S3: Confusion matrix of morphology grouping. *A visualisation of the accuracy of the ConvNet for each class, as given by the colouring intensity. The diagonal shows correct predictions. Even-toed ungulates are often misclassified for example.*

TABLE S4: Species grouping breakdown. *The species in each group, and the per class accuracy metrics. A transfer error caused slight loss of images for capped groups.*

Group Name	Scientific Name	Training Set Size	Test Set Size	F1-score	Precision	Recall
Ape	*Pongo pygmaeus*	869	200	0.94	0.98	0.91
	Cephalopachus bancanus					
	Hylobates muelleri					
Banded palm civet	*Hemigalus derbyanus*	4,207	200	0.94	0.92	0.97
	Diplogale hosei					
Banteng	*Bos javanicus*	964	200	1.00	1.00	1.00
Bay cat	*Pardofelis badia*	126	42	0.88	1.00	0.79
Bearded pig	*Sus barbatus*	4,599	197	0.86	0.86	0.86
Beautiful squirrels	*Callosciurus adamsi*	56	19	0.97	1.00	0.95
	Callosciurus notatus					
	Callosciurus prevostii					
Binturong	*Arctictis binturong*	141	47	0.90	0.95	0.85
Borneo elephant	*Elephas maximus*	146	48	0.95	1.00	0.90
Clouded leopard	*Neofelis diardi*	238	80	0.99	1.00	0.99
Collared mongoose	*Herpestes semitorquatus*	138	46	0.93	1.00	0.87
Common palm civet	*Paradoxurus hermaphroditus*	74	25	0.92	0.92	0.92
Common treeshrew	*Tupaia longipes*	705	200	0.94	0.97	0.91
	Ptilocercus lowii					
Crane	*Rallina fasciata*	416	138	1.00	1.00	0.99
	Amaurornis phoenicurus					
	Ardea alba					
Cukoo	*Centropus sinensis*	172	57	0.98	1.00	0.96
	Carpococcyx radiceus					
	Spilornis cheela					
Domestic dog	*Canis lupus familiaris*	1,465	200	0.99	0.99	1.00
Dove	*Chalcophaps indica*	1,018	200	0.98	0.98	0.99
	Columbiformes					
Empty	-	4,766	199	0.87	0.88	0.85
Greater mouse deer	*Tragulus napu*	4,987	200	0.94	0.92	0.96
Horse-tailed squirrel	*Sundasciurus hippurus*	183	61	0.90	0.96	0.85
Human	*Homo sapiens*	4,985	200	0.93	0.91	0.95
Landfowl	*Lophura ignita*	4,906	200	0.89	0.85	0.94
	Argusianus argus					
	Lophura bulweri					
	Arborophila charltonii					
	Excalfactoria chinensis					
	Gallus gallus					
	Rollulus rouloul					
Langur	*Presbytis rubicunda*	65	22	0.93	0.95	0.91
	Presbytis hosei					
Large treeshrew	*Tupaia tana*	776	198	0.94	0.95	0.93
Leopard cat	*Prionailurus bengalensis*	1,778	200	0.97	0.96	0.98
Lesser mouse-deer	*Tragulus kanchil*	4,901	200	0.94	0.91	0.98
Long-tailed giant rat	*Leopoldamys sabanus*	1,087	200	0.96	0.95	0.98
Long-tailed macaque	*Macaca fascicularis*	72	24	0.96	1.00	0.92
Long-tailed porcupine	*Trichys fasciculata*	3,009	200	0.96	0.95	0.97
Low's squirrel	*Sundasciurus lowi*	731	200	0.95	0.95	0.96
	Sundasciurus tenuis					
Malayan civet	*Viverra tangalunga*	4,967	200	0.93	0.91	0.95
	Prionodon linsang					
Malayan porcupine	*Hystrix brachyura*	4,958	200	0.96	0.94	0.98
Marbled cat	*Pardofelis marmorata*	204	68	0.92	0.98	0.87
Masked palm civet	*Paguma larvata*	622	200	0.89	0.97	0.83
Moonrat	*Echinosorex gymnura*	62	21	0.90	0.95	0.86
Mueller's giant Sunda rat	*Maxomys*	148	50	1.00	1.00	1.00
	Sundamys muelleri					
	Rattus rattus					
Pangolin	*Manis javanica*	338	113	0.96	0.96	0.96
Pig-tailed macaque	*Macaca nemestrina*	4,962	200	0.84	0.76	0.94
Red muntjac	*Muntiacus muntjak*	4,924	200	0.82	0.83	0.81
Sambar deer	*Rusa unicolor*	4,990	200	0.95	0.94	0.95
Short-tailed mongoose	*Herpestes brachyurus*	525	175	0.92	0.96	0.87
Slender treeshrew	*Tupaia gracilis*	268	90	0.93	0.92	0.94
Songbird	*Pellorneum capistratum*	1,201	200	0.97	0.96	0.99
	Arachnothera longirostra					
	Pittidae					
	Copsychus stricklandii					
	Malacocincla malaccensis					
	Enicurus leschenaulti					
	Copsychus saularis					
	Pycnonotus goiavier					
	Lonchura atricapilla					
	Malacopteron magnirostre					
	Ptilocichla leucogrammica					
Stink badger	*Mydaus javanensis*	242	81	0.96	0.97	0.95
Striped ground squirrel	*Lariscus hosei*	374	124	0.93	0.96	0.90
	Exilisciurus exilis					
	Rheithrosciurus macrotis					
	Aeromys thomasi					
Striped treeshrew	*Tupaia dorsalis*	226	76	0.94	0.91	0.96
Sun bear	*Helarctos malayanus*	1,749	199	0.96	0.96	0.95
Thick-spined porcupine	*Hystrix crassispinis*	497	166	0.96	0.98	0.95
Yellow muntjac	*Muntiacus atherodes*	4,995	152	0.97	1.00	0.94
Yellow-throated marten	*Aonyx cinereus*	449	200	0.93	0.91	0.96
	Martes flavigula					
	Mustela nudipes					
Total / Average	-	84,281	6,918	0.94	0.94	0.94

TABLE S5: Family grouping breakdown. *The taxonomic groups in each class, and the per class accuracy metrics. A transfer error caused slight loss of images for capped groups.*

Group Name	Scientific Name	Training Set Size	Test Set Size	F1-score	Precision	Recall
Ape	Hominidae Hylobatidae Tarsiidae	819	200	0.93	0.97	0.90
Bear	Ursidae	1,670	200	0.96	0.96	0.97
Bovidae	Bovidae	964	200	0.97	0.98	0.97
Canidae	Canidae	1,465	200	0.97	0.96	0.97
Cat	Felidae	2,450	200	0.97	0.97	0.97
Crane	Ardeidae Rallidae	411	137	1.00	0.99	1.00
Cukoo	Accipitridae Cuculidae	172	57	0.97	1.00	0.95
Deer	Cervidae	4,992	200	0.85	0.85	0.85
Dove	Columbidae Streptopelia	1,018	200	0.98	0.99	0.98
Elephant	Elephantidae	146	48	0.96	0.98	0.94
Empty	-	4,810	198	0.91	0.94	0.88
Human	Homo sapiens	5,000	199	0.88	0.91	0.84
Landfowl	Phasianidae	4,996	200	0.92	0.90	0.94
Mongoose	Herpestidae	666	200	0.94	0.97	0.92
Moonrat	Erinaceidae	62	21	0.93	0.95	0.90
Mouse-deer	Tragulidae	5,000	200	0.91	0.87	0.94
Murid	Muridae	1,253	200	0.96	0.96	0.96
Mustelidae	Mustelidae	392	130	0.91	0.97	0.85
Old world monkey	Cercopithecidae	4,990	200	0.87	0.80	0.95
Pangolin	Manidae	333	111	0.95	0.98	0.93
Pig	Suidae	4,938	200	0.85	0.85	0.85
Porcupine	Hystricidae	4,998	200	0.91	0.91	0.92
Songbird	Pellorneidae Nectariniidae Pittidae Muscicapidae Pycnonotidae Estrildidae	1,157	200	0.99	0.98	1.00
Squirrel	Sciuridae	1,400	200	0.97	0.97	0.98
Stinkbadger	Mephitidae	242	81	0.96	0.97	0.94
Treeshrew	Ptilocercidae Tupaiidae	2,070	200	0.96	0.97	0.95
Viverrids	Viverridae	5,000	200	0.91	0.88	0.94
Total / Average	-	61,414	4,582	0.93	0.93	0.93

TABLE S6: Morphology grouping breakdown. *The taxonomic groups in each class, and the per class accuracy metrics. A transfer error caused slight loss of images for capped groups.*

Group Name	Scientific Name	Training Set Size	Test Set Size	F1-score	Precision	Recall
Bear	Ursidae	1,748	198	0.94	0.98	0.90
Bird	Aves	4,906	200	0.90	0.89	0.93
Diurnal primate	Cercopithecidae Hominidae Hylobatidae	4,948	200	0.90	0.87	0.92
Diurnal small mammal	Ptilocercidae Sciuridae Tupaiidae	3,964	200	0.94	0.94	0.95
Elephant	Elephantidae	146	48	0.90	1.00	0.81
Empty	-	4,791	200	0.87	0.89	0.84
Even toed ungulate	Bovidae Cervidae Suidae Tragulidae	4,876	200	0.80	0.81	0.79
Human	*Homo sapiens*	4,983	200	0.93	0.91	0.95
Nocturnal small mammal	Erinaceidae Muridae Tarsiidae	1,409	200	0.97	0.96	0.98
Pangolin	Manidae	338	113	0.95	0.98	0.92
Porcupine	Hystricidae	4,975	199	0.91	0.85	0.96
Small carnivore	Canidae Felidae Herpestidae Mephitidae Mustelidae Viverridae	4,989	200	0.86	0.88	0.85
Total / Average	-	42,073	2,158	0.90	0.91	0.90

6.2.2 Background Analysis

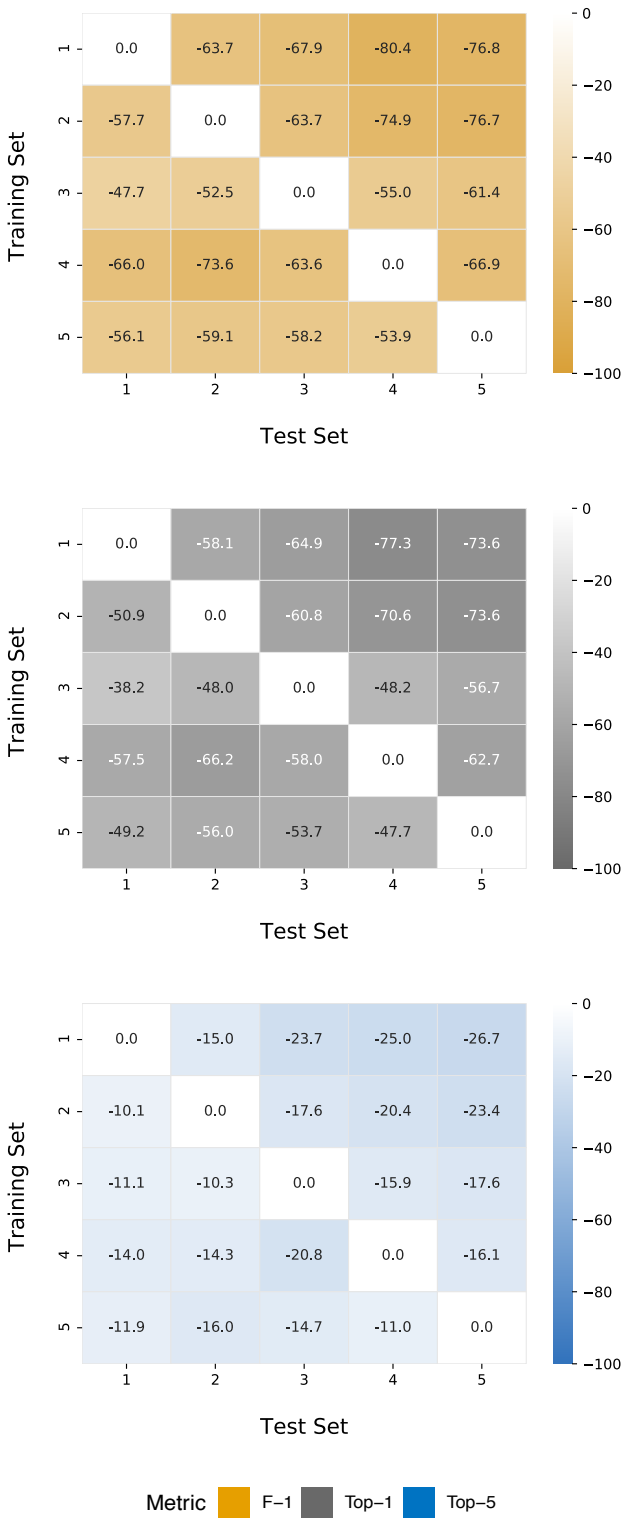

FIGURE S4: Heatmap of the accuracy difference for the VGG-16 architecture when trained and tested on different vegetation types. *The magnitude of the difference is shown by the colour intensity. The accuracy obtained from using the same vegetation type for the training and test set were taken as maximum reference.*

TABLE S7: Training and test set sizes of each class, for each vegetation type. *The superscript numbers indicate in which vegetation the species were found. If no superscript is included they were found in all vegetation types. Training on all vegetations, was the total of each individual group, capped at 5000 for the training and 200 for the test.*

Group Name	Scientific Name	Vegetation 1 Training Set Size	Vegetation 1 Test Set Size	Vegetation 2 Training Set Size	Vegetation 2 Test Set Size	Vegetation 3 Training Set Size	Vegetation 3 Test Set Size	Vegetation 4 Training Set Size	Vegetation 4 Test Set Size	Vegetation 5 Training Set Size	Vegetation 5 Test Set Size
Cat	*Pardofelis badia* [3, 4] *Pardofelis marmorata* [2, 3, 4, 5] *Prionailurus bengalensis* *Neofelis diardi* [1, 3, 4, 5]	303	101	863	200	410	137	113	38	364	121
Deer	*Muntiacus atherodes* [1, 3, 4, 5] *Muntiacus muntjak* *Rusa unicolor*	5,000	200	5,000	200	4,989	200	5,000	200	4,997	200
Empty	-	4,999	200	5,000	200	4,765	199	1379	200	1,692	200
Human	*Homo sapiens*	5,000	200	4,999	200	5,000	200	2,735	200	5,000	200
Landfowl	*Arborophila charltonii* [3, 4, 5] *Argusianus argus* [1, 3, 4, 5] *Excalfactoria chinensis* [2] *Gallus gallus* [3] *Lophura bulweri* [3, 4, 5] *Lophura ignita* *Rollulus rouloul* [2, 3, 4, 5]	755	200	1,114	200	4,985	200	4,381	200	4,450	200
Mouse-deer	*Tragulus kanchil* *Tragulus napu* [1, 3, 4, 5]	146	49	190	63	2,594	200	1,585	200	5,000	200
Old world monkey	*Macaca fascicularis* [1, 2, 5] *Macaca nemestrina* *Presbytis hosei* [5] *Presbytis rubicunda* [3, 4, 5]	2,232	200	302	100	4,981	200	3,070	200	5,000	200
Pig	Sus barbatus	4,972	200	4,941	200	4,901	200	5,000	200	4,980	200
Porcupine	*Hystrix brachyura* *Hystrix crassispinis* [3, 5] *Trichys fasciculata*	439	146	1,320	200	3,898	200	2,342	200	3,659	200
Viverrids	*Arctictis binturong* [3, 4, 5] *Diplogale hosei* [3, 5] *Hemigalus derbyanus* [2, 3, 4, 5] *Paguma larvata* *Paradoxurus hermaphroditus* [1, 2, 3] *Prionodon linsang* [3, 4, 5] *Viverra tangalunga*	1,611	200	2,615	200	4,427	200	1,209	200	3,231	200
Total / Average	-	25,457	1,696	26,344	1,763	40,950	1,936	26,814	1,838	38,373	1,921

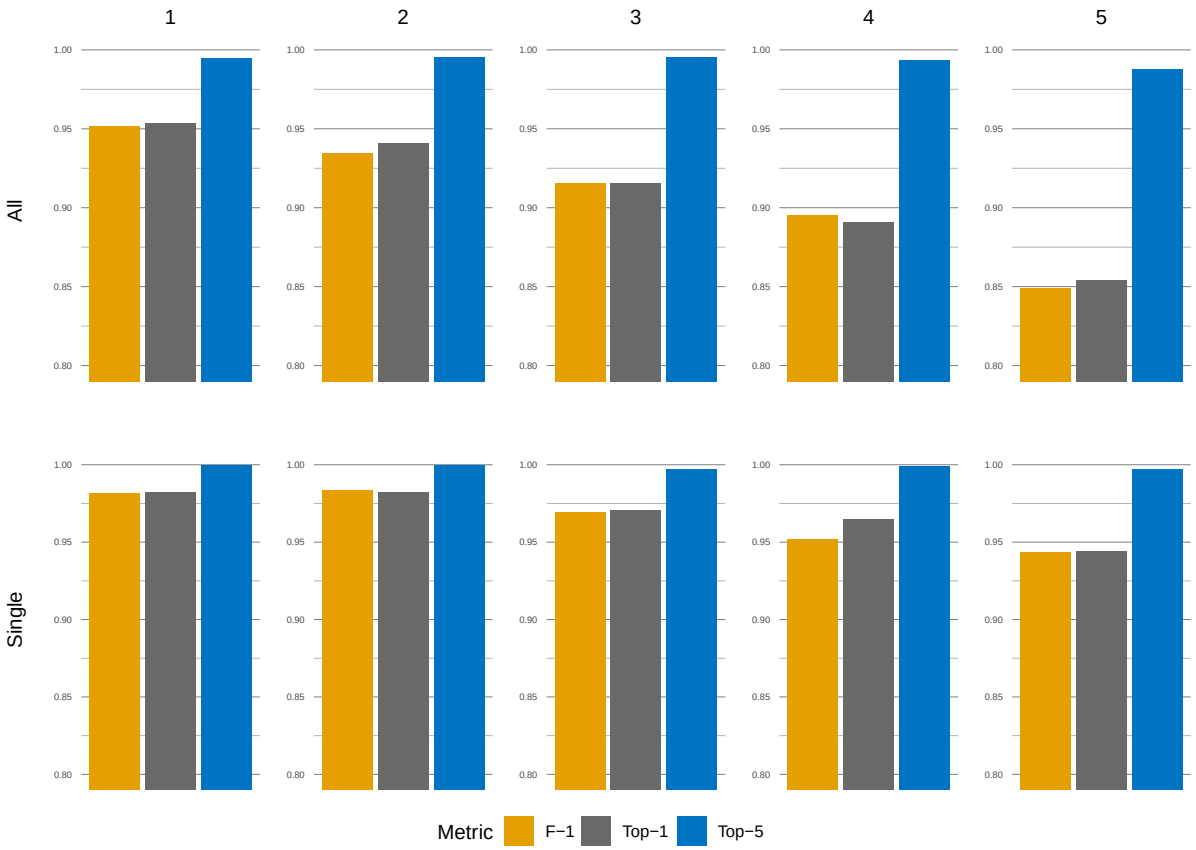

FIGURE S5: Model performance when trained on all or a single vegetation type, and tested on single vegetations *Numbers represent the vegetation type of the test set, with 1 being open area and 5 being undisturbed forest). Note the y-axis break. Training on a single vegetation type and testing on the same vegetation as the training, resulted in higher accuracies than when a mixture of vegetations were used. In both instances there was a general decrease in model performance, as you moved from vegetation 1 to vegetation 5.*

6.2.3 Number of images effect

Contrary to Tabak et al. (2018) and Norouzzadeh et al. (2018), class accuracy did not increase with increased number of images. It is likely that this occurred because the images used in the training and test set for small classes come from the same trigger events. Larger classes are likely to have images in the test set from trigger events, which are not in the training set, which makes classification more difficult. While this may have resulted in poorer performance of the VGG-16 model, it is probable that the larger groups exhibit accuracies which are more representative of the models performance.

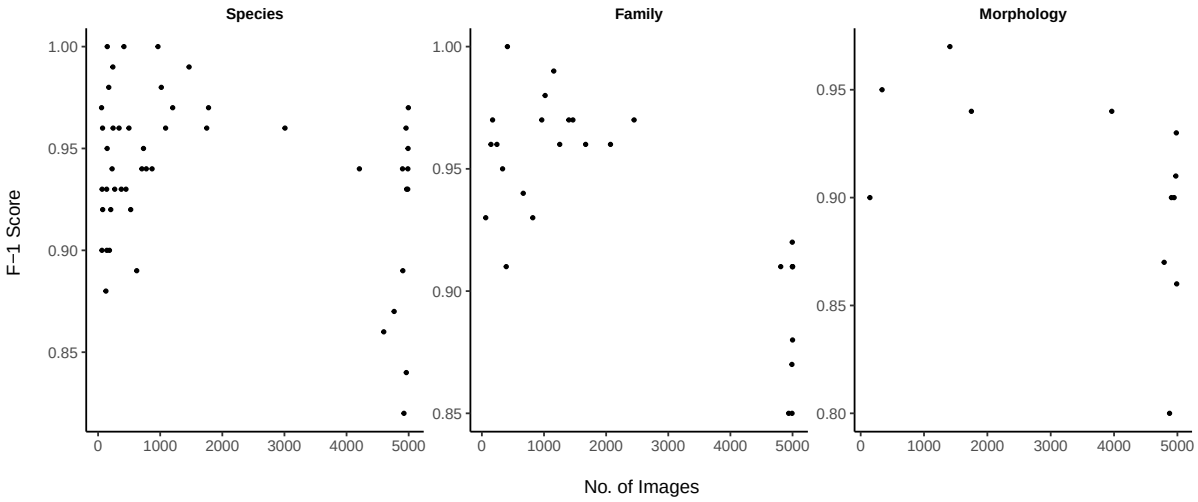

FIGURE S6: The correlation between number of images per class and accuracy of that class. *Each graphs show the per class accuracies for a single grouping type. There was no clear increase in accuracy as the number of images of a class increased. A positive relationship appears to occur for classes with image sizes below 4000 images. However, classes with highly extensive images, which were usually the classes that were capped, exhibited no notably higher accuracies, in fact sometimes even lower.*