Automated Location Invariant Animal Detection In Camera Trap Images Using Publicly Available Data Sources

Andrew Shepley[1] | Greg Falzon[2] | Paul Meek[3,4] | Paul Kwan[5]

[1]School of Science and Technology, University of New England; Armidale, NSW, Australia
[2]College of Science and Engineering , Flinders University; Adelaide, SA, Australia
[3]Vertebrate Pest Research Unit, NSW Department of Primary Industries, PO Box 530, Coffs Harbour, NSW, Australia
[4]School of Environmental and Rural Science, University of New England, Armidale, NSW, Australia
[5]School of IT and Engineering, Melbourne Institute of Technology, Australia

Correspondence: Andrew Shepley: asheple2@une.edu.au

Keywords: *animal identification, artificial intelligence, camera trapping, camera trap images, deep convolutional neural networks, deep learning, wildlife ecology, wildlife monitoring, location invariance, infusion*

Abstract

1. A time-consuming challenge faced by camera trap practitioners is the extraction of meaningful data from images to inform ecological management. An increasingly popular solution is automated image classification software. However, most software solutions are not sufficiently robust to be deployed on a large scale due to lack of location invariance

when transferring models between sites. This prevents optimal use of ecological data and results in significant expenditure of time and resources to annotate and retrain deep learning models.

2. In this study, we aimed to (a) assess the value of publicly available image datasets in the training of deep learning models for camera trap object detection focusing on images obtained from FlickR and iNaturalist (FiN), (b) develop a method to be used by ecologists to train location invariant image processing object detection models and (c) explore the use of small subsets of camera trap images in the optimization of FiN training.
3. We collected and annotated 3 datasets of images of the following classes; striped hyena, rhinoceros and pig, from the image sharing websites, and used transfer learning to train 3 object detection models in the task of animal detection. We compared the performance of these models to the performance of 3 models trained on the Wildlife Conservation Society and Camera CATalogue datasets, when tested on out of sample Snapshot Serengeti datasets. Furthermore, we explored optimization of the FiN trained models via infusion of small subsets of camera trap images to increase robustness for challenging detection cases.
4. In all experiments, the mean Average Precision (mAP) of the FiN trained models was significantly higher (82.33-88.59%) than that achieved by the models trained only on camera trap datasets (38.5-66.74%). The infusion of camera trap images into FiN training further improved mAP, with increases ranging from 1.78-32.08%.
5. Ecology researchers can use FiN images for training deep learning object detection solutions for camera trap image processing to develop location invariant, robust, out-of-the-box software. This would allow AI technologies to be deployed on a large scale in ecological applications. Datasets and code related to this study are open source and available on this repository: https://github.com/ashep29/infusion

1. Introduction

Automated survey methods such as camera trapping and passive acoustic monitoring are widely used in ecological research (Rovero and Zimmermann 2016, Sugai, Silva et al. 2018, Gibb, Browning et al. 2019). These methods provide invaluable insight into a plethora of ecological information including species occurrence, activity patterns and behavior (O'Connell, Nichols et al. 2011). However, they often result in the collection of large quantities of data, which must be processed, requiring a significant commitment of time and resources for manual or supervised classification (Swinnen, Reijniers et al. 2014, Young, Rode-Margono et al. 2018). Reducing the processing time and resources necessary for traditional data analysis such as manual analysis and citizen science (Swanson, Kosmala et al. 2015, Nguyen, Maclagan et al. 2017) has prompted increasing research into the adoption of Artificial Intelligence (AI) software in automated data classification (Falzon, Meek et al. 2014, Norouzzadeh, Nguyen et al. 2018, Willi, Pitman et al. 2018).

Object detector and image classifier software (models) have already been adopted to some extent in the processing of camera trap images (Yu, Jiangping et al. 2013, Gomez Villa, Salazar et al. 2016, Norouzzadeh, Nguyen et al. 2018, Willi, Pitman et al. 2018, Tabak, Norouzzadeh et al. 2019, Falzon, Lawson et al. 2020). These tools rely on data-driven deep learning to identify complex patterns which can be used for classification without feature engineering as described by (Miao, Gaynor et al. 2019). However, most solutions presented thus far have shown limited transferability to image data outside the domain of the training data (Beery, Van Horn et al. 2018, Willi, Pitman et al. 2018). This results in the need to develop models specific to each domain, however this process is time and resource intensive, requiring repeated collection and manual annotation of camera trap data, and computationally expensive training of deep neural networks (Falzon, Lawson et al. 2020). Thus, there is a clear need to develop location invariant object detectors, which are deep learning models that can be transferred from one location to another, achieving acceptable results without having to be retrained. Such out-of-the-box solutions are attractive due to their potential for extensive application, particularly in circumstances where the development of domain or study-specific models is prohibitively expensive or otherwise unattainable.

Achieving location invariance requires training data to be characterized by high intra-dataset variability. This is because neural networks learn patterns in data, meaning low intra-dataset variability can result in learning of domain specific features such as camera angle, lighting, and vegetation, reducing location invariance (Torralba and Sinha 2003, Miao, Gaynor et al. 2019, Singh, Lindshield et al. 2020). Therefore, camera trap images must be obtained from many sources to be able to train effective object detectors and classifiers. However, the process of collecting camera trap images from an extensive network of cameras from many domains is time and resource intensive and may be unfeasible for smaller scale studies or those focusing on rare or elusive species. Even when researchers have access to camera trap network, collecting enough images for training object detectors can prove difficult. (Maurice 2019) deployed 15 cameras for 2 months resulting in the collection of only 41 images of the pangolin (the target species), a number which would be insufficient for effective neural network training (Shahinfar, Meek et al. 2020). Other factors which limit the accessibility and availability of camera trap images include the reticence of researchers to share existing camera trap data, or lack of data for novel species studies.

These limitations in data accessibility and availability limit the adoption of automated AI solutions in ecological camera trap image processing (Schneider, Taylor et al. 2018). Thus, alternative data sources must be identified and evaluated to assist in the development of object detectors capable of being deployed in any domain, at any location, achieving acceptable results regardless of camera trap image availability. Possible solutions include publicly available sources of animal imagery, such as FlickR (flickr.com) and iNaturalist (inaturalist.org). FlickR is a consumer photo sharing website, hosting approximately 10 billion images, shared by over 90 million monthly users. It is characterized by high intra-dataset variability, high accessibility and a wide range of species types in highly varying contexts, with minimal unintentional bias, as images are not collected for a specific purpose (Everingham, Van Gool et al. 2010). It is arguably the most extensively used source of image data in object detection benchmark datasets, including ImageNet (Deng, Dong et al. 2009), MS COCO (Lin, Maire et al. 2014), the Open Images Dataset (Kuznetsova, Rom et al. 2020) and PASCAL VOC (Everingham, Van Gool et al. 2010). iNaturalist contains over 45 million observations of biodiversity data including both flora and fauna. Labelling of images on iNaturalist may be more accurate than FlickR due to its purpose as a biodiversity data sharing website and it does contain more camera trap images than FlickR. Other potential image sources include Pinterest (www.pinterest.com), Imgur (www.imgur.com), pixabay (www.pixabay.com) and 500px (www.web.500px.com). These image sources are highly beneficial in training general, location invariant neural networks as they exhibit an extensive range of contextual features, not necessarily present in camera trap imagery.

Despite their benefits as out-of-the-box solutions, universal or general object detectors usually fail to achieve the high accuracy attainable by domain-specific object detectors (Rebuffi, Bilen et al. 2017, Wang, Cai et al. 2019). Due to the need to achieve high accuracy object detection and classification in ecological research, it may therefore be necessary to optimize location invariant models for domain-specific studies. This is particularly relevant when processing camera trap imagery characterized by features which differ strongly from non-camera trap data, including infrared imagery, poor quality illumination and blurry images.

Therefore the aims of this study are twofold:

i) To evaluate the use of publicly available image sources, in the development of location invariant camera trap object detectors.

ii) To develop an optimization strategy dubbed 'infusion' to improve the performance of location invariant object detectors in domain-specific applications.

In this study, we will demonstrate our proposed approach on three single class applications. The rare species Striped Hyena (*Hyaena hyaena*) was chosen due to the sparsity of camera trap training data, and the difficulty in discriminating between the striped hyena and the more common spotted hyena. Furthermore, other studies have highlighted it as a species of particular interest due to the difficulty they faced in detecting its presence in camera trap images, for example, (Willi, Pitman et al. 2018) failed to detect any of the 27 striped hyenas present in their test dataset. Next, the iconic and critically endangered Rhinoceros (*Rhinocerotidae*) was also chosen, due to the high research interest in monitoring its prevalence and changes in populations. Finally, the pest family *Suidae* (pigs, boars and hogs) was included due to the significant role it plays across global ecosystems and its host status for a range of diseases such as Swine Fever, which are a major threat to agricultural industries.

2. Related Work

a. Traditional Methods: Manual Analysis and Citizen Science

The majority of camera trap image processing is achieved by manual analysis conducted by ecologists, or via citizen science. Manual analysis involves the use of software programs to manually tag animals in images/capture events. Each image sequence or capture event is treated as a detection, and the ecologist must manually select a tag reflecting the identity of the animal. Once tagging is complete, a verification process is undertaken to identify and correct mistaken classifications. These tagged images can then be interrogated according to the purpose of the study, using tools such as R scripts, or specially developed GUI programs. Manual analysis of images is a significant resource demand on ecologists and research teams, requiring large expenditures in time and resources, hindering effective biodiversity management.

This time-consuming task may also be undertaken by citizen scientists, who are volunteers that contribute to scientific enquiry by collecting or processing image data (Nguyen, Maclagan et al. 2017). Large citizen science-based programs such as Zooniverse (www.zooniverse.org) enable the effective classification of millions of camera trap images (Jones, Allen et al. 2018). Citizen science projects have many benefits for researchers including customization of projects and annotation requirements in accordance with the aims of projects. However, the effectiveness of citizen science in rapidly processing large volumes of image data with sufficient accuracy is limited (Meek and Zimmerman 2016), causing large delays between the data collection and interpretation stages, which may be detrimental to ecological management (Fox, Bourn et al. 2019). Furthermore, the need to upload significant amounts of data onto publicly accessible websites may pose privacy risks (Sagarra, Gutiérrez-Roig et al. 2015) or poaching concerns and undermine the protection of rare or endangered species by revealing their geographical location and behavioral habits to poachers (Falzon, Lawson et al. 2020).

b. Automated Image Processing Using Deep Learning

Due to the shortcomings of traditional methods, research has centered primarily on integration of automated image processing within camera trap research (Meek, Fleming et al. 2014, Meek, Ballard et al. 2015, Fegraus and MacCarthy 2016, Willi, Pitman et al. 2018, Young, Rode-Margono et al. 2018). To achieve this, neural networks such as Deep Convolutional Neural Networks (DCNNs) are trained on large amounts of annotated image data (thousands to millions of images) to recognize discriminative features belonging to target classes (Zhao, Zheng et al. 2019). Handcrafted features specified by researchers are not used, instead the features are 'learned' via updating of weights during training. When the DCNN is confident in the presence of an object in an image, it maps bounding boxes, segmentation masks, or classification labels to the image or object (Ren, He et al. 2015). If a DCNN is very deep, consisting of many layers, it will have many trainable parameters (usually millions) which gives rise to the need for large annotated image datasets used in training these parameters from scratch. This is necessary for the network to learn complex features (Samala, Chan et al. 2016). Although DCNNs can be used to classify data with high accuracy, their usability can be limited by insufficient training data which may lead to overfitting (memorization of training data), and consequently, inability of the model to generalize to new data (Zhao 2017).

Early attempts at automated camera trap classification and object detection tasks using neural networks were dependent on significant amounts of pre-processing (Yu, Jiangping et al. 2013) and resulted in relatively poor accuracy (Swinnen, Reijniers et al. 2014, Chen, Han et al. 2015). However, most modern solutions use minimal pre-processing, or automate pre-processing (Giraldo Zuluaga, Salazar et al. 2017). Accuracy and recall attained by deep learning solutions is also increasing significantly, as large annotated datasets become available and progress is achieved in training methods, such as the adoption of transfer learning (Gomez Villa, Salazar et al. 2016, Willi, Pitman et al. 2018). Transfer learning involves the repurposing of learned features for another task (Yosinski, Clune et al. 2014). This allows general features learned on a large, highly varied dataset such as ImageNet (Deng, Dong et al. 2009) which contains 3.2 million images, or Snapshot Serengeti (Swanson, Kosmala et al. 2015), which contains 7.3 million images to be transferred to a smaller, similar dataset containing only hundreds to thousands of images. Transfer learning has been shown to improve accuracy and the ability to generalize as well as reducing training time and the quantity of data needed (Khan, Hon et al. 2019). Its effectiveness in ecological camera trap applications has been established by (Norouzzadeh, Nguyen et al. 2017) and (Willi, Pitman et al. 2018).

c. Image Classification vs. Object Detection

The majority of camera trap image processing solutions achieve image classification rather than object detection (Gomez Villa, Salazar et al. 2016, Nguyen, Maclagan et al. 2017, Norouzzadeh, Nguyen et al. 2017, Willi, Pitman et al. 2018, Miao, Gaynor et al. 2019, Tabak, Norouzzadeh et al. 2019). Image classification is a process by which a whole image is labeled as containing a given object, for example, if a pig is featured in an image, it will be labelled ‘pig. However, image classification is limited in situations where an image contains more than one species, e.g. a pig and a wildebeest (Schneider, Taylor et al. 2018). Object localization and counting is also not effectively achieved by image classification and models tend to struggle to distinguish between an empty frame and a small background object (Yousif, Yuan et al. 2019). In contrast, object detection is the process of locating and identifying one or more objects in an image. The model plots bounding boxes of varying classification confidence and association class labels, around each object in an image (see Figure 1 for comparison). It is more useful than image classification because it allows more information to be extracted from the images, such as the number of objects in an image, as well as information about reproduction, distribution, quantification and comparison of behavior across individual animals within a species group based on factors such as age and gender (Schneider, Taylor et al. 2018).

Another major benefit of object detection is the reduced impact of background and environmental features on object classification. Unlike image classifiers, which learn patterns in the entire image, object detectors only learn patterns within the constraints of the bounding boxes, and actively negative sample on the image background (area not included in the bounding boxes) (Wang, Hu et al. 2019, Zhao, Zheng et al. 2019). This enables object detectors to better generalize to new domains, thus facilitating location invariance. Despite these benefits, object detection necessitates a significantly higher expenditure of time and resources, due to the need to annotate all training images with bounding boxes and labels.

Consequently, most studies achieve image classification rather than object detection. In contrast, due to the major benefits provided by object detectors for automated camera trap image processing , this study focuses on object detection rather than image classification. For a more detailed overview of available image classification methods, refer to Appendix S1.

Several studies have achieved object detection in the context of camera trap image processing, however none have achieved location invariance, with testing using restricted to in-sample datasets. (Yousif, Yuan et al. 2019) employed sequence-level background subtraction using handcrafted Histogram of Oriented Gradient (HOG) (Dalal and Triggs 2005) features to localize moving objects in camera trap images. This study did not aim to identify animal species, instead simply distinguished between humans and animals, and eliminated empty frames. Although it achieved high accuracy in this task, its application was not extended beyond eastern North America.

A novel ecological image processing software solution for use on a laptop by field ecologists and wildlife managers was developed by (Falzon, Lawson et al. 2020). It provides object detection and localization as well as species classification and object counting capabilities via training of YOLOv2 DarkNet-19 (Redmon and Farhadi 2016) Deep Convolutional Neural Networks (DCNN) on both daytime and infrared imagery. It boasts fast processing speeds and acceptable accuracy, achieved on a local machine, within a dedicated on-demand application. Tailored models can be applied to trap sites in Australia, New Zealand, North America, Serengeti and the USA. However, optimal performance is only achieved when models are trained and developed for a specific environment, camera trap imaging configuration and species cohort. Thus, it suffers from lack of location invariance and robustness, as its accuracy and recall decrease significantly when it is used outside the scope of the environments on which it was trained.

(Schneider, Taylor et al. 2018) addressed the problem of object detection in camera trap images, with the aim of identifying, quantifying and localizing animal species. They used transfer learning to train a YOLOv2 model, achieving recall of 93% and accuracy of 80.4% on the Reconyx (www.reconyx.com) and Snapshot Serengeti (Swanson, Kosmala et al. 2015) datasets. The Reconyx dataset contained 946 images of 20 species, while the Snapshot Serengeti dataset contained 4,097 images of 48 species. They also trained a Faster R-CNN model (Ren, He et al. 2015) achieving 76.7% recall and 72.2% accuracy. They used a model pretrained on the MS COCO dataset (Lin, Maire et al. 2014) to initialize transfer learning. However, the robustness of the model was not evaluated on out of sample images, which is camera trap imagery obtained from traps and geographical locations not included in the training data. It also suffered from class imbalance with lower accuracy and recall for classes with fewer instances. Our research indicates this limitation can be overcome by sourcing images from publicly available data sources.

d. Improving Location Invariance via Dataset Construction

The suboptimal performance and inability of neural networks to generalize to contexts beyond the domain of the training data is a strong area of research interest. As early as 2008, studies in contextual object detection examined the consequences of 'unintentional regularities' in datasets resulting in object detectors learning associations between objects and their backgrounds, inhibiting their ability to detect objects out of context (Hoiem, Efros et al. 2008, Sudderth, Torralba et al. 2008). (Everingham, Van Gool et al. 2010) noted that classifiers tend to learn the context of an object rather than model the appearance of the object. Thus, when the object is dissociated with its context, the classifier fails to detect it due to extensive use of image composition and context, resulting in a significant drop in performance. These findings were confirmed by (Miao, Gaynor et al. 2019) in an ecological context via the use of GRAD-CAM technology applied to models trained solely on camera trap images, illustrating the tendency of neural networks to learn background features as elements of an object if image background and context is not highly varied. It is therefore essential to broaden the context of animal imagery to extend beyond a restricted range of camera traps to ensure robustness and location and context invariance.

This phenomena of contextual association was also found by (Everingham, Van Gool et al. 2010) to be particularly prevalent in neural networks trained on images taken by researchers for a specific purpose. Consistencies within datasets, such as camera trap images collected within the context of a specific project, create an inner dataset bias, which results in the development of models less capable of generalization to other camera trap contexts. On this basis, we postulate that collection of camera trap images for neural network training mimics collection of images under laboratory or controlled conditions, whereby features such as lighting, camera angle, distance of objects from the camera, and background features are consistent across many images, thus encouraging contextual association. This is supported by (Willi, Pitman et al. 2018) who noted that their models, trained on camera trap images, would need to be retrained for use out of sample in other camera traps which did not form part of the training set. In contrast, networks trained on data sourced from consumer photo sharing websites such as FlickR are more capable of generalization (Torralba and Efros 2011) due to the inherently high intra-dataset variability and reduced likelihood of inner dataset bias.

3. Datasets and Annotation

The datasets used in this study were collated using images from FlickR and iNaturalist. We also used camera trap image datasets obtained from www.lila.science including Snapshot Serengeti (SS), Wildlife Conservation Society (WCS) Camera Traps, as well as other sites specified in more detail below. All datasets, annotations, and the algorithms used for dataset collection and processing, as well as auto-annotation of images are available here: https://github.com/ashep29/infusion.

a. FlickR and iNaturalist

We developed and used a Python script to download images from FlickR using the FlickR API. This allowed us to download images with multiple keywords at once. The keywords used are shown in Table 1. We downloaded a maximum of 200 images per keyword, to maximize the variety of search results. Our datasets were restricted to Creative Commons images. We also developed a Python script to download images from iNaturalist using a csv file containing URLs of relevant observations downloaded from inaturalist.org.

Table 1: *Keyword searches used to download images from FlickR and iNaturalist. Scientific names tended to return more accurately labelled images.*

Rhinocerotidae	**Hyaena hyaena**	**Suidae**
diceros AND bicornis *ceratotherium AND simum* *dicerorhinus AND sumatrensis* *white AND rhinoceros* *rhinoceros*	*striped AND hyena* *Hyaena AND hyaena*	*Phacochoerus AND africanus* *Sus AND scrofa* *sanglier* *warthog OR warthogs* *wild AND pig OR boar OR hog* *feral AND pig OR boar OR hog*

Duplicates and near duplicates were removed using a Structural Similarity Index (SSIM) (Zhou, Bovik et al. 2004) clustering algorithm we developed (see Appendix S4). We deleted all images with a similarity score above 0.8, where a score of 1.0 represents a 100% similarity between 2 images. Near duplicates are images with strong visual similarity, containing only small distortions, slight variations and occlusions (Everingham, Van Gool et al. 2010). Interestingly, the datasets downloaded from FlickR and iNaturalist were mutually exclusive, with not one image present on one site, being also present on the other. Although this does not mean that images obtained from FlickR will not be available via iNaturalist, it does suggest that users of FlickR may often not be users of iNaturalist. Details about the final datasets are shown in Table 2. Subsamples of the final datasets are illustrated by Figure 1.

Table 2: *Final number of images obtained from FlickR and iNaturalist for both the single class and multi-class experiments, after duplicate removal and cleaning. Datasets are referred to hereon according to their source, abbreviated as FiN (FlickR-iNaturalist) and class name.*

Dataset Name	***Class***	***FlickR***	***iNaturalist***	***Total Images***
FiN_rhino	*Rhino*	784	881	1665
FiN_striped_hyena	*Striped hyena*	401	71	472
FiN_pig	*Pig*	606	0	606

Figure 1: *Subsamples of the FiN datasets. Top to bottom: striped hyena, rhinoceros, and pig. Images of were highly varied, and included both color/daytime and infrared images, as well as a large range of contexts and distances from the camera.*

b. Camera Trap Datasets

We obtained all camera trap data of rhinoceros and striped hyena from lila.science using a Python script we developed, which we have made available on our GitHub repository. We scoured all images of striped hyena and rhinoceros from both WCS Camera Traps (*WCS_striped_hyena* and *WCS_rhino*) and Snapshot Serengeti (*SS_striped_hyena* and *SS_rhino*) datasets (Swanson, Kosmala et al. 2015). We used the same script to obtain our *EU_pig* and *NA_pig* datasets from the Missouri Camera Traps (Zhang, He et al. 2016) and North American Camera Trap Images (Tabak, Norouzzadeh et al. 2018) datasets respectively, also from lila.science. A summary of all camera trap datasets is provided in Table 3.

Table 3: *Summary of the characteristics of the camera trap datasets used in this study. The term 'quality' refers to characteristics such as blurriness, pixilation, illumination etc. A poor-quality dataset will contain many images that are over or underexposed, blurriness caused by poor focus, or other features which make it harder to distinguish the identity of a target class and distort or damage key features. A visual subsample of these datasets is provided (see Figure 2).*

Dataset	Source	Location	Size	Characteristics
WCS_striped_hyena	Wildlife Conservation Society	Multiple	582	Moderate quality Night and day
SS_striped_hyena	Snapshot Serengeti	Tanzania	478	Moderate quality Infrared and day Includes partials
WCS_rhino	Wildlife Conservation Society	Multiple	333	Low quality Mostly infrared Many partials
SS_rhino	Snapshot Serengeti	Tanzania	153	Moderate quality Daytime Many partials
AU_pig	Custom	NSW, Australia	589	Low quality Mostly infrared High occlusion High density
SS_pig	Snapshot Serengeti	Tanzania	574	Moderate quality Mostly daytime

CC_pig	Camera CATalogue	South Africa	559	Moderate quality Partials Low density
NA_pig	North America Camera Trap Images	United States	514	High quality
EU_pig	Missouri Camera Traps	Europe	501	Difficult High occlusion

The *SS_pig* dataset is a subset of the Snapshot Serengeti dataset, and *CC_pig* is a subset of the Camera CATalogue project conducted by Panthera (www.panthera.org). Both are available from the Data Repository for the University of Minnesota, used by (Willi, Pitman et al. 2018) and released under a CC0 1.0 Universal Public Domain Dedication license. The Australian pig dataset (AU_pig) is a custom dataset, obtained during feral pig trapping and control operations. More information about each dataset is provided in Table 3, and a subset is shown in Figure 2.

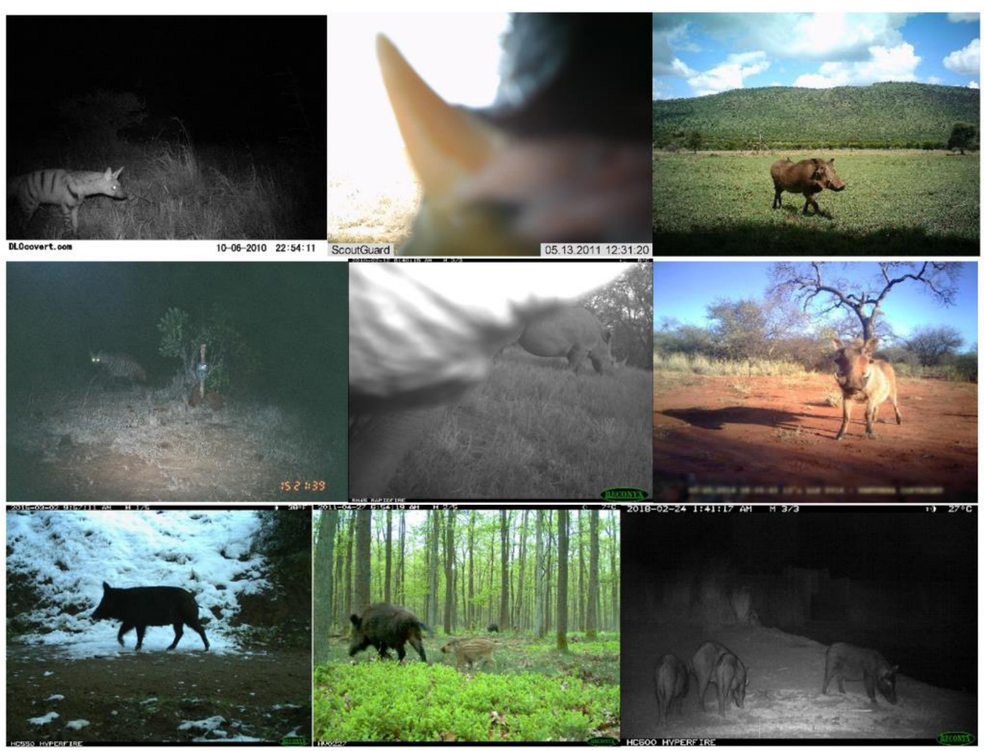


Figure 2: *Subsamples of the camera trap datasets. Top row: SS datasets, left to right; striped hyena, rhino, and pig. Middle row: WCS datasets, left to right; striped hyena, rhino, and pig. Bottom row: left; pig from NA_pig, middle; pigs from EU_pig and right; pigs from AU_pig.*

Each image in the final datasets were annotated with bounding boxes and corresponding class labels. Bounding box annotation involves the positioning of an axis aligned box surrounding an object. We used an auto-annotator tool we developed to roughly annotate all the images. We then edited any suboptimal bounding boxes using the graphical annotation tool labelImg (Tzutalin 2015)[i] to ensure all objects were correctly annotated. Annotations were saved in PASCAL VOC format.

4. Training and Evaluation Methodology

In this study, we conducted two major experiments. Firstly, we compared the performance of models trained on FlickR-iNaturalist (FiN) datasets only to those trained only on camera trap data using evaluation on out of sample test sets. Next, we optimized the FiN models by infusing small subsets of camera trap imagery into the FiN training set, evaluating performance on out of sample test sets. Details about the model architecture and training parameters are provided in Appendix S3. Additional information on transfer learning is also provided. The experiments outlined in this section were also verified on a multi-class application documented in Appendix S5.

a. Comparison between FiN and Camera Trap Data in Developing Location Invariant Object Detectors

To evaluate the potential for publicly available data from FlickR and iNaturalist to be used in the development of location invariant object detectors for camera trap image processing, we trained Keras-RetinaNet (Lin, Goyal et al. 2018) models on FiN datasets, and compared their performance to that of RetinaNet models trained on camera trap data when tested on out of sample camera trap images.

We trained 3 single-class RetinaNet models on FiN datasets. These models are referred to as *FiN_Classname*, e.g. *FiN_rhino* refers to a rhino detector trained on FiN data. We also trained 2 single class (rhino and striped hyena) RetinaNet models using the *WCS_striped_hyena* and *WCS_rhino* datasets, as well as 4 pig detectors, on the *AU_pig*, *CC_pig*, *NA_pig* and *EU_pig* datasets. All models are named based on the source of their training data. Note, we were able to train 4 pig models due to greater availability of data when compared with rare species such as rhino and striped hyena.

The datasets were randomly split into training and validation sets, with 90% of images reserved for training, and 10% used for validation. Each training set was supplemented with 800 explicit negative samples to improve discrimination between target species and non-target species or background. A detailed breakdown of the training and validation splits as well as the out of sample test set is provided in Table 4.

Table 4: *Data distribution for models trained on datasets obtained from FlickR/iNaturalist, abbreviated as FiN (FlickR-iNaturalist), and models trained using camera trap images alone abbreviated as follows; WCS (Wildlife Conservation Society), AU (Australia), NA (North America), CC (Camera CATalogue) and EU*

(Europe). All models were tested on out of sample images obtained from Snapshot Serengeti.

Models	Training set (90%)	Validation set (10%)	Out of Sample Test set (SS)
FiN_striped_hyena	425	47	478
WCS_striped_hyena	524	58	
FiN_rhino	1499	166	153
WCS_rhino	300	33	
FiN_pig	545	61	574
AU_pig	530	59	
CC_pig	503	56	
NA_pig	463	51	
EU_pig	451	50	

All models were tested using out of sample images from the Snapshot Serengeti (SS) datasets, i.e. *SS_striped_hyena*, *SS_rhino* and *SS_pig*. Each test set was supplemented with 200 negative samples to prevent biased evaluation of false positives. These negative samples were derived from the Snapshot Serenget, and consisted of empty images, or images of non-target species. For more information relating to the negative sampling data collection process, refer to Appendix S2.

b. Infusion: Optimization of Location Invariant Models Using Camera Trap Imagery

Next, we conducted experiments to evaluate an optimization process that would allow ecologists to improve object detection performance with minimal infusion of camera trap images into the FiN training set. Infusion is the process of supplementing the training set with a small subset of camera trap images, to improve robustness to the particularities of camera trap data, such as infrared, high occlusion, blurriness etc. Infusion was conducted both out of sample and in-sample. Out of sample results are presented in this manuscript. For in-sample results, refer to Appendix S6.

Due to the large number of highly similar images present within camera trap datasets, the infusion subsets were not randomly selected. Instead, our SSIM algorithm was used to retain only images with low SSIM scores, with the aim of maximizing intra-dataset variability. The SSIM algorithm allowed us to randomly select one frame from each cluster of images (usually one capture event, or different capture events with very similar properties).

Figure 3: *Graphical illustration of image clustering using an SSIM algorithm. The test image represented by 1.0 is compared with every other image. Highly dissimilar images have low SSIM scores (<0.4).*

Our research indicates that image pairs with an SSIM value above 0.4 have sufficiently high similarity to be clustered. For example, Figure 3 illustrates the output of the SSIM algorithm graphically, clearly showing the three clusters formed by visually similar images, the image denoted by the arrow (the test image) is compared to each other image, with values closest to 1 indicating high similarity with the test image, This method allows researchers to compile highly varied datasets automatically, minimizing the need for extensive time-consuming image sorting and annotation.

Out of sample infusion was conducted by training 4 additional models for each species, with incremental infusion of the SSIM sorted camera trap images from the WCS and CC datasets into the FiN training data. These images were added in increments of 5% from 5-20%, as shown by Table 5. For example, the *FiN_rhino* dataset comprised of 1665 images. To achieve 5% infusion, 83 images from the *WCS_rhino* dataset were added to the *FiN_rhino* dataset. 90% of these images were retained for training, with 10% reserved for monitoring training via the validation set. This process was repeated for all percentiles and species shown in Table 5.

Table 5: *Incremental infusion of camera trap images into FiN training. An additional 800 negative samples were included in the training set. Models are named according to the class name and infusion percentile. Note the infusion images are trap images. The infusion training set is made up of FiN + infusion images. The validation set is FiN validation + infusion images.*

Class	Model name	Infusion Source	№ infusion images	Infusion training set	Infusion Validation set
Hyaena	hyaena_inf_05	WCS_hyena	24	446	50
	hyaena_inf_10		47	467	52
	hyaena_inf_15		71	489	54
	hyaena_inf_20		94	509	57
Rhino	rhino_inf_05	WCS_rhino	83	1573	175
	rhino_inf_10		167	1649	183
	rhino_inf_15		250	1723	192
	rhino_inf_20		333	1798	200
Pig	pig_inf_05	CC_pig	30	572	64
	pig_inf_10		61	600	67
	pig_inf_15		91	627	70
	pig_inf_20		121	654	73

The models were then tested on the out of sample Snapshot Serengeti test sets presented in Section 4(a). Both the training and test sets were supplemented with negative samples as described in Section 4(a).

c. Model Evaluation

To evaluate the performance of our models, mean Average Precision (mAP) results will be provided. mAP is calculated as documented in the PASCAL VOC benchmark (Everingham, Van Gool et al. 2010). A high mAP indicates that the model is detecting the majority of objects with high accuracy, and minimal retention of false positives. Accuracy is measured using Intersection over Union (IoU), which is a measure of the overlap between the detection box and the ground truth bounding box.

5. Results

a. Comparison between FiN and Camera Trap Data in Developing Location Invariant Object Detectors

The results of training on FiN data compared with training on camera trap data are presented in Figure 4. All results were collected on the out of sample Snapshot Serengeti test sets. The models trained on FiN datasets achieved mAP results ranging between 82.33% and 88.59%, while the models trained on camera trap data achieved mAP results ranging from 38.5% to 66.74%. In all cases, the FiN models outperformed the models trained on camera trap images.

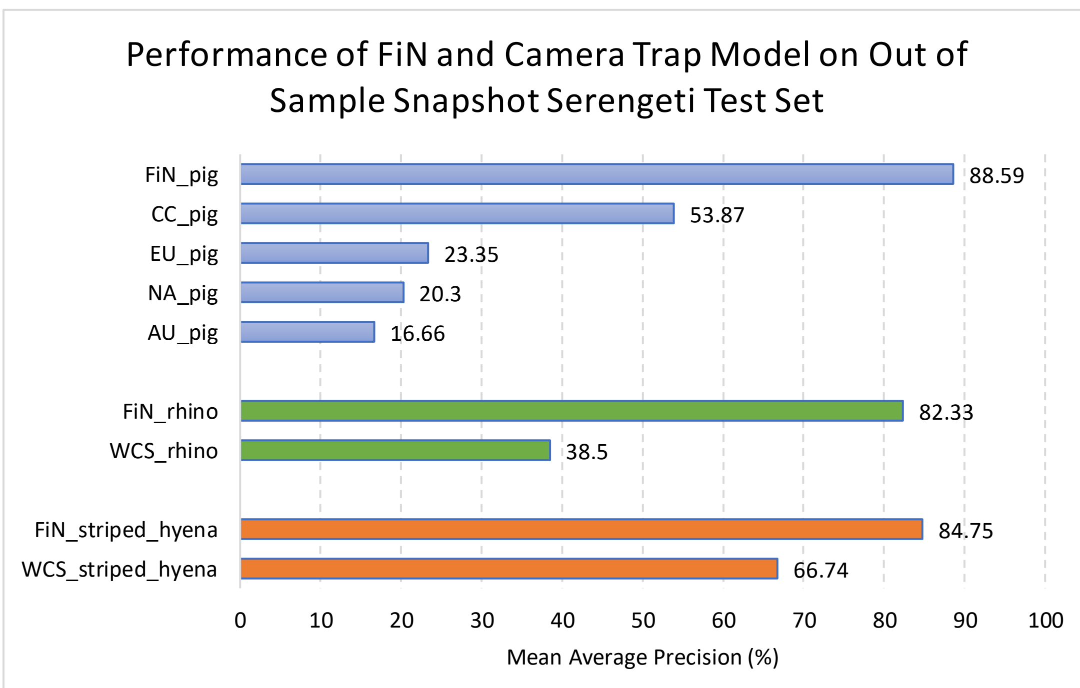


Figure 4: *Comparison of the mAP results achieved by the models trained on FiN data, and those trained on camera trap datasets. In all cases, the FiN models outperformed*

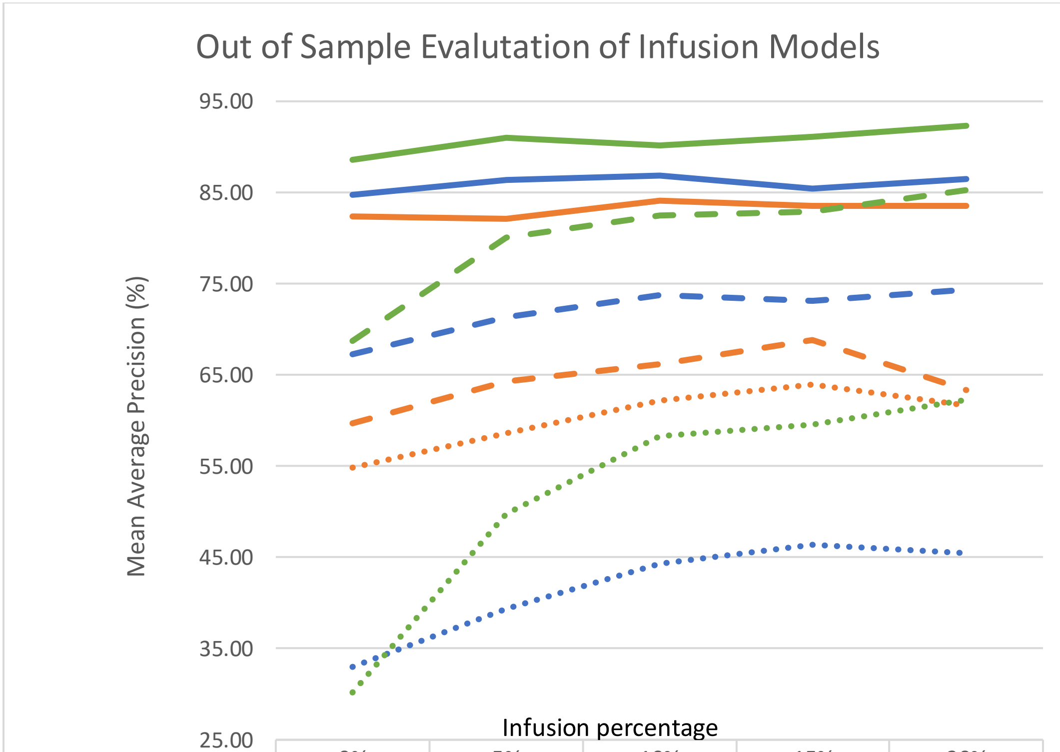
the camera trap models.

The *FiN_pig* model achieved a mAP of 88.59% when tested on the out of sample *SS_pig* dataset. This was far superior to the *CC_pig* model, which was trained on camera trap images of warthogs from the Camera CATalogue (CC) dataset, achieving a mAP of only 53.87%. Although both the *CC_pig* dataset and the *SS_pig* dataset contained the same subspecies (*Phacochoerus africanus*), the *CC_pig* model did not generalize well to the *SS_pig* test set. This may be because the *SS_pig* dataset was characterized by more variation in background, greater variation in the distance of pigs from the camera and greater contrast. Notably, the worst performing pig model was trained on data from Australia (*AU_pig*). This is very likely due to the large number of low quality infrared images present in the training data, which encouraged the model to return a high rate of false positives, and the large disparity between contextual features such as vegetation and species type (the Australia subspecies was *Sus scrofa*, while the SS subspecies was *Phacochoerus africanus*).

In comparison, the significantly greater intra-dataset variability present in the FiN datasets allowed for better model generalization when compared to the models trained only on single location camera trap data. This trend was observed across all classes, with the *FiN_striped_hyena* and *FiN_rhino* models significantly outperforming the *WCS_striped_hyena* and *WCS_rhino* models.

b. Infusion: Optimization of Location Invariant Models Using Camera Trap Imagery

The results presented in the previous section indicate that the models trained on FiN datasets can be used to effectively process images collected at any camera trap site with an acceptable level of location invariance. However, camera trap images possess particular characteristics which differentiate them from FiN images. In difficult cases, the mAP achieved by FiN models may not be sufficiently high for practical purposes, particularly when higher confidence thresholds are used, for example, for a given study, the confidence threshold may be set to 50%, meaning all detections with a classification score lower than 50% would be ignored. Thus, we present the results of our infusion optimization experiments, illustrated by Figure 5. In all cases, infusion resulted in an increase in mAP when evaluated on out of sample images.



Threshold	Model	0%	5%	10%	15%	20%
conf@5%	Striped hyena	84.75	86.39	86.86	85.43	86.52
conf@5%	Rhinoceros	82.33	82.12	84.11	83.59	83.53
conf@5%	Pig	88.59	91.00	90.14	91.10	92.32
conf@30%	Striped hyena	59.70	64.32	66.18	68.83	63.39
conf@30%	Rhinoceros	67.27	71.38	73.75	73.16	74.35
conf@30%	Pig	68.73	80.05	82.45	82.90	85.27
conf@50%	Striped hyena	32.99	39.37	44.34	46.38	45.47
conf@50%	Rhinoceros	54.84	58.56	62.18	63.95	61.68
conf@50%	Pig	30.19	49.81	58.27	59.54	62.27

Figure 5: *Results of the infusion experiments on the out of sample SS test set. Infusion resulted in improvement across all models, particularly when evaluated at higher confidence thresholds. Infusion of 5% significantly improves performance, however optimum performance occurs at 10-15%, with the mAP results plateauing beyond 15%.*

At a confidence threshold of 5% (the standard threshold for mAP measurement (Lin, Goyal et al. 2018)), out of sample infusion did not result in a pronounced improvement, with gains in mAP results ranging from 1.78-3.73%. However, in practical deployment, a confidence threshold of 5% would rarely be used, with ecologists favoring higher thresholds to ensure confident classification of species. It is at these higher thresholds that the benefits of infusion are best demonstrated. For example, at a confidence threshold of 30%, the mAP improved by 7.08-16.54%, while at a confidence threshold of 50% it improved by 9.11- 32.08%. It is well established that increasing the confidence threshold decreases recall (the number of true positives retained in the final output), and consequently decreases mAP (Willi, Pitman et al. 2018). Note, we did not conduct evaluations of the models at confidence thresholds above 50% because almost all detections with scores above 50% were true positives, which meant increasing the threshold simply removed true positives. Selecting a confidence threshold for a given application is highly dependent on the quality of training data, extent of negative sampling and the model used. The supplementation of FiN training with out of sample camera trap imagery is therefore highly beneficial as it allows more true positives to be retained, because the overall confidence of correctly detected objects is improved. This is a result of the improved robustness to the particularities of camera trap imagery.

The results presented in Figure 5 indicate that the addition of a small percentage of camera trap images into the FiN training dataset can significantly improve performance. In most cases, the greatest improvement occurred with infusion of 5%, with performance continuing to improve as infusion was increased to 15%. As infusion was increased beyond 15%, performance plateaued, or decreased, with only 4 out of 9 results improving beyond 15%.

6. Discussion

We investigated the use of FiN images as an alternative to camera trap images in the task of DCNN training for location invariant camera trap image processing tasks, on three case studies, namely striped hyena, rhinoceros and pig. Specifically, we established the greater transferability of the FiN trained models when compared to models trained on camera trap datasets, and their high usability as location invariant object detectors. We then demonstrated how such models can be optimized via out of sample infusion, which was shown to increase the confidence of detections, allowing more true positives to be retained at higher confidence thresholds.

Our results show that FiN training significantly improves model robustness and location invariance. Particularly, it provides ecologists with a practical, cost effective, out of the box solution, capable of detecting animals even in the most challenging camera trap environments. We not only established that FiN data alone can be used to achieve good results, but these models can be improved with minimal infusion of camera trap data to improve robustness to the particularities of camera trap imagery. This suggests that ecologists can train object detectors using FiN imagery, and if camera trap data is available for their target species, use it to infuse the FiN training data. This model can then be used to process out of sample images from any camera trap, achieving a sufficiently high mAP to be deployed in most applications.

Furthermore, in circumstances where model performance is still considered suboptimal, they may then infuse the model with in-sample camera trap images, for further optimization. Although in-sample infusion makes the model more location variant, it does provide a means by which ecologists can train powerful models capable of achieving results in the 90th percentile, with very few training images, as demonstrated by the results of in-sample infusion presented in Appendix S6. As demonstrated by various studies in automated camera trap image processing, achieving robust object detectors via training solely on camera trap images usually requires thousands to millions of images (Norouzzadeh, Nguyen et al. 2017, Willi, Pitman et al. 2018, Tabak, Norouzzadeh et al. 2019). In-sample infusion overcomes this requirement by leveraging off the robustness of the FiN model, and the strong availability of FiN imagery to allow ecologists to train high accuracy optimized deep leaning models with very few camera trap images, significantly reducing the time and resources necessary to develop automated deep leaning object detectors.

In light of the growing number of camera trap based projects undertaken by ecologists, this research provides an invaluable method by which researchers can process extensive image data regardless of the location from which the images were obtained, and the particularities of the camera trap site or species. This method has been proven on several species, including rare species, for which camera trap data for training models is often sparse. As illustrated by (Willi, Pitman et al. 2018), the lack of camera trap data for rare species poses significant problems when training multi-class object detectors, as the large class imbalance between common species and rare species causes object detectors to misclassify species, by over enthusiastically classifying species based on how common they are in the dataset rather than via their features. This was observed by (Willi, Pitman et al. 2018) who noted that insufficient images of the rare striped hyena in their dataset resulted in their model achieving a mAP of 0% on this class. We have specifically addressed this problem by proposing the use of FiN images of striped hyena to rectify limitations in data availability.

The use of FlickR as the principal training data also rectifies another major problem faced by researchers. Studies have indicated that deep learning models have a tendency to return overly confident predictions (Willi, Pitman et al. 2018) when trained on camera trap data and deployed in-sample. This is due to the high consistency in image quality, lighting, camera angle and geographical and vegetation features in camera trap data. Furthermore, many trap images feature obscured or poor quality imagery of animals which if used in the training set, may cause the network to make unrealistically optimistic predictions, by attributing 100% confidence to visual features which may not display sufficiently distinct characteristics present solely in the target class. In contrast, the higher resolution of FiN images and large variations between images forces the model to reduce the confidence attributed to poor quality or obscured animals. Their greater robustness allows them to be deployed out of sample, further minimizing this problem.

One potential benefit in using FiN imagery for training image processing models is the high availability of already annotated animal images. Because FlickR is a major source of images used in datasets such as ImageNet and MS COCO, many animal classes have already been annotated with bounding boxes, which are freely available for downloading. Using the method proposed in the paper would therefore significantly reduce the time and resource expenditure necessary for model development, by leveraging off the work already completed by the broader object detection community. We were unable to use annotated FlickR images from ImageNet as it was under maintenance, however it may prove to be a valuable resource in the development of future models. This study was limited to the evaluation of FlickR and iNaturalist images, and did not evaluate alternative images sources mentioned in Section 1.

This research did not investigate the application of the FiN and infusion training method using alternative object detectors such as YOLO (Redmon and Farhadi 2016), and Faster R-CNN (Ren, He et al. 2015). Applying the findings of this study to these architectures may be beneficial. YOLO is a faster, more efficient object detector, which may be more suited to video processing, while Faster RCNN generally achieves higher accuracies, but is slower. RetinaNet was chosen as it achieves a good balance between the computational efficiency of YOLO and the accuracy of Faster-RCNN, which made it an appropriate choice for the difficult task of camera trap image processing. In this study, we have only demonstrated location invariance using RetinaNet. Although it goes beyond the scope of this study, it would be interesting to ascertain whether changes in model architecture would influence the robustness of location invariance models. Another possible area of research could be the application of this method to object segmentation-based image processing. Object segmentation builds upon the benefits of object detection by excluding background features. This limits the influence of contextual features on model performance, thus improving model accuracy and overall performance, however it is likely that they would encounter the same modelling bias faced by bounding box-based object detection models.

One limitation of this study is that it only evaluates the models in terms of the Snapshot Serengeti dataset. We could only evaluate on one dataset for the classes 'striped hyena' and 'rhinoceros' due to lack of data availability. To maintain consistency, we also only presented results for the class 'pig' on Snapshot Serengeti in this manuscript. However, to verify the usability of this method at any location and for any dataset, we present more extensive results in Appendix S7 for the class pig, for which we had more data available, thus showing location invariance across 4 extra test locations.

Finally, the proposed method may be extended to other image modalities. For example, it could be extended to drone imagery (Kellenberger, Volpi et al. 2017, Xu, Wang et al. 2020). Drone images are often captured from an aerial perspective, meaning they would contain quite different features to those present available on FlickR. Applying our findings to object detection in the context of drone imagery would be interesting, particularly with infusion of a small subset of drone images to boost performance and allow better generalization to the particularities of drone imagery. This would determine how transferable FiN images are to new modalities. It could also be extended to other applications such as underwater animal imagery (Dawkins, Sherrill et al. 2017, Christensen, Mogensen et al. 2018), surveillance footage (Raghunandan, Mohana et al. 2018), and thermal camera imagery (Rodin, Lima et al. 2018, Bondi, Jain et al. 2020). This may present opportunities to rectify image shortages, or problems with low intra-dataset variability, particularly in novel studies.

7. Conclusion

This study successfully demonstrated the use of FiN datasets in training location invariant deep learning object detection models in the task of camera trap image processing. It also evaluated an optimization process dubbed infusion, to improve robustness to the particularities of camera trap imagery. Results presented across three single class models on out of sample test sets indicate the aims of this study have been achieved. However, our approach is limited by its inability to achieve high precision out of sample object detection, which is still best achieved via in-sample training or infusion. Furthermore, this method was not evaluated on alternative object detection frameworks and did not provide findings on an extensive multi-class dataset. Nevertheless, this study provides a promising pathway to develop robust, location invariant models using publicly accessible data sources.

Furthermore, development of these models will facilitate the widespread deployment of AI in ecological management. The findings of this study could also be extended beyond camera trapping to other object detection tasks and image modalities such as drone imagery. Furthermore, the methodology of using transfer learning and publicly available datasets characterized by high intra-dataset variability and minimal unintentional bias to train location and context invariant AI-based data processing software could be extended beyond images to other forms of data.

8. Acknowledgements

Andrew Shepley is supported by an Australian Postgraduate Award. We would like to thank the Australian Department of Agriculture and Water Resources, the Centre for Invasive Species Solutions, NSW Environmental Trust, University of New England and the NSW Department of Primary Industries for supporting this project. We appreciate the Creative Commons Images provided through FlickR; Australian camera trap images provided to us by Mark Lamb and Jason Wishart; the Snapshot Serengeti, University of Missouri Camera Traps and North American Camera Traps datasets through the Labeled Information Library of Alexandria: Biology and Conservation and the Camera CATalogue dataset provided through the Data Repository of the University of Minnesota.

9. Author Contributions

Andrew Shepley, Greg Falzon and Paul Kwan conceived the ideas and designed methodology; Andrew Shepley, Paul Meek and Greg Falzon collected the data; Andrew Shepley and Greg Falzon analysed the data; Andrew Shepley led the writing of the manuscript. All authors contributed critically to the drafts and gave final approval for publication.

10. Data Accessibility

Image and Annotation datasets: All image datasets and corresponding annotations will be made available via Dryad upon acceptance.

Code and scripts: All code and scripts will be made available via Dryad upon acceptance.

11. Reference List

Beery, S., G. Van Horn and P. Perona (2018). Recognition in Terra Incognita, Cham, Springer International Publishing.

Bondi, E., R. Jain, P. Aggrawal, S. Anand, R. Hannaford, A. Kapoor, J. Piavis, S. Shah, L. Joppa, B. Dilkina and M. Tambe (2020). BIRDSAI: A Dataset for Detection and Tracking in Aerial Thermal Infrared Videos. 2020 IEEE Winter Conference on Applications of Computer Vision (WACV).

Chen, G., T. Han, Z. He, R. Kays and T. Forrester (2015). "Deep convolutional neural network based species recognition for wild animal monitoring." 2014 IEEE International Conference on Image Processing, ICIP 2014: 858-862.

Christensen, J. H., L. V. Mogensen, R. Galeazzi and J. C. Andersen (2018). Detection, Localization and Classification of Fish and Fish Species in Poor Conditions using Convolutional Neural Networks. 2018 IEEE/OES Autonomous Underwater Vehicle Workshop (AUV).

Dalal, N. and B. Triggs (2005). "Histograms of Oriented Gradients for Human Detection." IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2005) **2**.

Dawkins, M., L. Sherrill, K. Fieldhouse, A. Hoogs, B. Richards, D. Zhang, L. Prasad, K. Williams, N. Lauffenburger and G. Wang (2017). An Open-Source Platform for Underwater Image and Video Analytics. 2017 IEEE Winter Conference on Applications of Computer Vision (WACV).

Deng, J., W. Dong, R. Socher, L.-J. Li, K. Li and F. F. Li (2009). "ImageNet: a Large-Scale Hierarchical Image Database." IEEE Conference on Computer Vision and Pattern Recognition: 248-255.

Everingham, M., L. Van Gool, C. Williams, J. Winn and A. Zisserman (2010). "The Pascal Visual Object Classes (VOC) challenge." International Journal of Computer Vision **88**: 303-338.

Falzon, G., C. Lawson, K.-W. Cheung, K. Vernes, G. A. Ballard, P. J. S. Fleming, A. S. Glen, H. Milne, A. Mather-Zardain and P. D. Meek (2020). "ClassifyMe: A Field-Scouting Software for the Identification of Wildlife in Camera Trap Images." Animals **10**(1): 58.

Falzon, G., P. D. Meek and K. Vernes (2014). Computer Assisted identification of small Australian mammals in camera trap imagery. Camera Trapping: Wildlife Management and Research. Paul Meek, Peter Fleming, Guy Ballard et al. Melbourne, Australia, CSIRO Publishing: 299-306.

Fegraus, E. H. and J. MacCarthy (2016). Camera Trap Data Management and Interoperability. Camera Trapping for Wildlife Research. F. R. a. F. Zimmerman. Exeter UK, Pelagic Publishing: 33-42.

Fox, R., N. Bourn, E. Dennis, R. Heafield, I. Maclean and R. Wilson (2019). "Opinions of citizen scientists on open access to UK butterfly and moth occurrence data." Biodiversity and Conservation.

Gibb, R., E. Browning, P. Glover-Kapfer and K. E. Jones (2019). "Emerging opportunities and challenges for passive acoustics in ecological assessment and monitoring." Methods in Ecology and Evolution **10**(2): 169-185.

Giraldo Zuluaga, J., A. Salazar, A. Gomez Villa and A. Diaz-Pulido (2017). "Automatic Recognition of Mammal Genera on Camera-Trap Images using Multi-Layer Robust Principal Component Analysis and Mixture Neural Networks."

Gomez Villa, A., A. Salazar and J. Vargas-Bonilla (2016). "Towards Automatic Wild Animal Monitoring: Identification of Animal Species in Camera-trap Images using Very Deep Convolutional Neural Networks." Ecological Informatics **41**.

Hoiem, D., A. Efros and M. Hebert (2008). "Putting Objects in Perspective." International Journal of Computer Vision **80**: 3-15.

Jones, F., C. Allen, C. Arteta, J. Arthur, C. Black, L. Emmerson, R. Freeman, G. Hines, C. Lintott, Z. Macháčková, G. Miller, R. Simpson, C. Southwell, H. Torsey, A. Zisserman and T. Hart (2018). "Time-lapse imagery and volunteer classifications from the Zooniverse Penguin Watch project." Scientific Data **5**: 180124.

Kellenberger, B., M. Volpi and D. Tuia (2017). Fast animal detection in UAV images using convolutional neural networks. 2017 IEEE International Geoscience and Remote Sensing Symposium (IGARSS).

Khan, N., M. Hon and N. Abraham (2019). "Transfer Learning with intelligent training data selection for prediction of Alzheimer's Disease."

Kuznetsova, A., H. Rom, N. Alldrin, J. Uijlings, I. Krasin, J. Pont-Tuset, S. Kamali, S. Popov, M. Malloci, A. Kolesnikov, T. Duerig and V. Ferrari (2020). "The Open Images Dataset V4." International Journal of Computer Vision **128**(7): 1956-1981.

Lin, T.-Y., P. Goyal, R. Girshick, K. He and P. Dollar (2018). "Focal Loss for Dense Object Detection." IEEE Transactions on Pattern Analysis and Machine Intelligence **PP**: 1-1.

Lin, T.-Y., M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár and C. Zitnick (2014). "Microsoft COCO: Common Objects in Context." **8693**.

Maurice, M. E. (2019). A Survey on the Status of Pangolins By Camera Trapping in Deng-Deng National Park, Eastern Region, Cameroon, Ommega Internationals.

Meek, P. D., G.-A. Ballard and P. J. S. Fleming (2015). "The pitfalls of wildlife camera trapping as a survey tool in Australia." Australian Mammalogy **37**(1): 13-22.

Meek, P. D., P. Fleming, A. G. Ballard, P. B. Banks, A. W. Claridge, S. McMahon, J. Sanderson and D. E. Swann (2014). Putting contemporary camera trapping in focus. Camera Trapping in Wildlife Research and Management. B. PD Meek, A. G., Banks, P. B., Claridge, A. W., Fleming, P. J. S., Sanderson, J. G., and Swann, D. Melbourne, Victoria, CSIRP Publishing**:** 349-356.

Meek, P. D. and F. Zimmerman (2016). Camera Traps and Public Engagement. Camera Trapping for Wildlife Research. F. a. Z. Rovero, F. Exeter, Pelagic Publishing UK**:** 219-231.

Miao, Z., K. Gaynor, J. Wang, Z. Liu, O. Muellerklein, M. S. Norouzzadeh, A. McInturff, R. Bowie, R. Nathan, S. Yu and W. Getz (2019). "Insights and approaches using deep learning to classify wildlife." Scientific Reports **9**.

Nguyen, H., S. Maclagan, T. Nguyen, T. Nguyen, P. Flemons, K. Andrews, E. Ritchie and D. Phung (2017). "Animal Recognition and Identification with Deep Convolutional Neural Networks for Automated Wildlife Monitoring."

Norouzzadeh, M. S., A. Nguyen, M. Kosmala, A. Swanson, C. Packer and J. Clune (2017). "Automatically identifying wild animals in camera trap images with deep learning." Proceedings of the National Academy of Sciences **115**.

Norouzzadeh, M. S., A. Nguyen, M. Kosmala, A. Swanson, M. S. Palmer, C. Packer and J. Clune (2018). "Automatically identifying, counting, and describing wild animals in camera-trap images with deep learning." Proceedings of the National Academy of Sciences **115**(25): E5716.

O'Connell, A., J. D. Nichols and K. U. Karanth (2011). Camera traps in animal ecology: Methods and analyses.

Raghunandan, A., Mohana, P. Raghav and H. V. R. Aradhya (2018). Object Detection Algorithms for Video Surveillance Applications. 2018 International Conference on Communication and Signal Processing (ICCSP).

Rebuffi, S.-A., H. Bilen and A. Vedaldi (2017). "Learning multiple visual domains with residual adapters." 506--516.

Redmon, J. and A. Farhadi (2016). "YOLO9000: Better, Faster, Stronger."

Ren, S., K. He, R. Girshick and J. Sun (2015). "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks." IEEE Transactions on Pattern Analysis and Machine Intelligence **39**.

Ren, S., K. He, R. Girshick, X. Zhang and J. Sun (2015). "Object Detection Networks on Convolutional Feature Maps." IEEE Transactions on Pattern Analysis and Machine Intelligence **39**.

Rodin, C. D., L. N. d. Lima, F. A. d. A. Andrade, D. B. Haddad, T. A. Johansen and R. Storvold (2018). Object Classification in Thermal Images using Convolutional Neural Networks for Search and Rescue Missions with Unmanned Aerial Systems. 2018 International Joint Conference on Neural Networks (IJCNN).

Rovero, F. and F. Zimmermann (2016). Camera Trapping for Wildlife Research, Pelagic Publishing.

Sagarra, O., M. Gutiérrez-Roig, I. Bonhoure and J. Perelló (2015). "Citizen Science Practices for Computational Social Science Research: The Conceptualization of Pop-Up Experiments." Frontiers in Physics **3**.

Samala, R., H.-P. Chan, L. Hadjiiski, M. Helvie, J. Wei and K. Cha (2016). "Mass detection in digital breast tomosynthesis: Deep convolutional neural network with transfer learning from mammography." Medical Physics **43**: 6654.

Schneider, S., G. Taylor and S. Kremer (2018). "Deep Learning Object Detection Methods for Ecological Camera Trap Data." 321-328.

Shahinfar, S., P. Meek and G. Falzon (2020). " “How many images do I need?” Understanding how sample size per class affects deep learning model performance metrics for balanced designs in autonomous wildlife monitoring." Ecological Informatics **57**: 101085.

Singh, P., S. M. Lindshield, F. Zhu and A. R. Reibman (2020). Animal Localization in Camera-Trap Images with Complex Backgrounds. 2020 IEEE Southwest Symposium on Image Analysis and Interpretation (SSIAI).

Sudderth, E., A. Torralba, W. Freeman and A. Willsky (2008). "Describing Visual Scenes Using Transformed Objects and Parts." International Journal of Computer Vision **77**: 291-330.

Sugai, L., T. Silva, J. Ribeiro Jr and D. Llusia (2018). "Terrestrial Passive Acoustic Monitoring: Review and Perspectives." BioScience **69**.

Swanson, A., M. Kosmala, C. Lintott, R. Simpson, A. Smith and C. Packer (2015). "Snapshot Serengeti, high-frequency annotated camera trap images of 40 mammalian species in an African savanna." Scientific Data **2**: 150026.

Swinnen, K., J. Reijniers, M. Breno and H. Leirs (2014). "A Novel Method to Reduce Time Investment When Processing Videos from Camera Trap Studies." PloS one **9**: e98881.

Tabak, M., M. S. Norouzzadeh, S. Sweeney, K. Vercauteren, N. Snow, J. Halseth, P. Salvo, J. Lewis, M. White, B. Teton, R. Boughton, B. Wight, E. Newkirk, E. Odell, R. Brook, A. Moeller, E. Mandeville, J. Clune, R. Miller and P. Schlichting (2019). "Machine learning to classify animal species in camera trap images: Applications in ecology." Methods in Ecology and Evolution **10**: 585-590.

Tabak, M., M. S. Norouzzadeh, D. Wolfson, S. Sweeney, K. Vercauteren, N. Snow, J. Halseth, P. Salvo, J. Lewis, M. White, B. Teton, J. Beasley, P. Schlichting, R. Boughton, B. Wight, E. Newkirk, J. Ivan, E. Odell, R. Brook and R. Miller (2018). "Machine learning to classify animal species in camera trap images: applications in ecology."

Torralba, A. and A. Efros (2011). "Unbiased look at dataset bias." Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition: 1521 - 1528.

Torralba, A. and P. Sinha (2003). "Contextual Priming for Object Detection." International Journal of Computer Vision **53**.

Wang, X., Z. Cai, D. Gao and N. Vasconcelos (2019). Towards Universal Object Detection by Domain Attention. 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).

Wang, X., X. Hu, C. Chen, Z. Fan and S. Peng (2019). Improving Object Detection with Consistent Negative Sample Mining. Neural Information Processing, Cham, Springer International Publishing.

Willi, M., R. Pitman, A. Cardoso, C. Locke, A. Swanson, A. Boyer, M. Veldthuis and L. Fortson (2018). "Identifying Animal Species in Camera Trap Images using Deep Learning and Citizen Science." Methods in Ecology and Evolution **10**.

Xu, B., W. Wang, G. Falzon, P. Kwan, L. Guo, Z. Sun and C. Li (2020). "Livestock classification and counting in quadcopter aerial images using Mask R-CNN." International Journal of Remote Sensing **41**(21): 8121-8142.

Yosinski, J., J. Clune, Y. Bengio and H. Lipson (2014). "How transferable are features in deep neural networks?": 3320-3328.

Young, S., J. Rode-Margono and R. Amin (2018). "Software to facilitate and streamline camera trap data management: A review." Ecology and Evolution **8**(19): 9947-9957.

Yousif, H., J. Yuan, R. Kays and Z. He (2019). "Animal Scanner: Software for classifying humans, animals, and empty frames in camera trap images." Ecology and Evolution **9**.

Yu, X., W. Jiangping, R. Kays, P. Jansen, T. Wang and T. Huang (2013). "Automated identification of animal species in camera trap images." EURASIP Journal on Image and Video Processing **1**.

Zhang, Z., Z. He, G. Cao and C. Wenming (2016). "Animal Detection From Highly Cluttered Natural Scenes Using Spatiotemporal Object Region Proposals and Patch Verification." IEEE Transactions on Multimedia **18**: 1-1.

Zhao, W. (2017). "Research on the deep learning of the small sample data based on transfer learning." AIP Conference Proceedings **1864**: 020018.

Zhao, Z.-Q., P. Zheng, S.-T. Xu and X. Wu (2019). "Object Detection With Deep Learning: A Review." IEEE Transactions on Neural Networks and Learning Systems **PP**: 1-21.

Zhou, W., A. C. Bovik, H. R. Sheikh and E. P. Simoncelli (2004). "Image quality assessment: from error visibility to structural similarity." IEEE Transactions on Image Processing **13**(4): 600-612.

[i] https://github.com/tzutalin/labelImg