Automating Conspecific Identification of *Saimiri boliviensis* Squirrel Monkeys Using Machine Learning

Jovi Luo
School of Computer Science
The University of Auckland
Auckland, New Zealand
jluo019@aucklanduni.ac.nz

Adam Sinclair
School of Computer Science
The University of Auckland
Auckland, New Zealand
asin473@aucklanduni.ac.nz

Ian Chen
School of Computer Science
The University of Auckland
Auckland, New Zealand
iche195@aucklanduni.ac.nz

Addison Liu
School of Computer Science
The University of Auckland
Auckland, New Zealand
aliu349@aucklanduni.ac.nz

Daria Solovyeva
School of Computer Science
The University of Auckland
Auckland, New Zealand
dsol083@aucklanduni.ac.nz

***Abstract*—Detection and identification of individuals of a particular animal species from photo or video footage is imperative for monitoring and tracking of animals within enclosures, and for animal conservation efforts in general. During the last several decades the use of Deep Learning for computer vision problems has been getting increasingly popular, with many different state-of-the-art models available today. Convolutional Neural Networks are one of the options - widely used and readily accessible, they only recently started to be employed to tackle automation of the tedious task of manual animal observation and identification.**

This paper explores whether *Given photos of individual squirrel monkeys (S. boliviensis), can we come up with a pipeline to automatically identify them?* To achieve this, we introduce an extension to existing fully automated pipelines for conspecific (individual) identification. Previous works employ Mask R-CNN for segmentation and VGG base architecture for identification, but we introduce Triplet Loss as an extension for identification. Additionally, we novelly apply our research in the domain of Squirrel Monkeys. We test our methodology using Repeated Stratified K-Fold Cross Validation. Our findings show that an end to end pipeline for conspecific identification of this species is indeed possible, with our selected methodology achieving a closed-set accuracy of 86.4% and open-set accuracy of 70%.

Our work reflects the adaptability of this methodology to new domains, with the exciting possibility of extension into mobile identification for easy use by conservationists, zoologists, and general enthusiasts alike.

***Index Terms*—animal detection, individual classification, deep learning, CNN, squirrel monkey, metric learning**

I. Introduction

Wildlife conservation around the world involves a large degree of manpower, time and effort. For a particular species, conservationists can monitor various different aspects; for example, habitats, social behaviours, feeding habits, mating cycles and nesting. Species that are endangered are often in low numbers due to habitat loss [1], introduced species preying on them [2] and poor adaptability, among other reasons.

Something that is not specific to endangered species is that animals are often found injured e.g. from getting caught in a trap laid by humans. The changing health of animals will cause behavioural differences to occur. These individuals are rescued, treated, healed, rehabilitated and are often released back into the wild. Before they are released, a method of tracking these individuals is devised and implemented first, as monitoring their progress in the wild can be considered a later stage of rehabilitation. Often this is done by tagging (a procedure that is often used by farmers to identify their livestock by individual [3]), as well as the implantation of a microchip that can then be used to electronically track the position of the chipped individual [4], [5]. The method of tracking is different for each species.

Identification between different species of animals is quite advanced and is well documented [3], [6]. Unfortunately, among a group of individuals of the same species, identification of individuals is more difficult by comparison. Using tags is reasonable, since each tag will have a unique ID on it that identifies a unique individual. However, reading the tags requires a clear view of the ID on the tag, which would likely require physically moving close enough to see the tag to read its details. Using implanted microchips is also reasonable, as microchips are unique between individuals, but in some cases it is inconvenient. For instance, Auckland Zoo's black-capped squirrel monkeys (*Saimiri boliviensis*) all look quite similar, and it is quite difficult to ascertain the identities of each monkey visually. At present, a very small number of

the zookeepers maintaining the squirrel monkeys are able to visually identify them individually. For the other staff and volunteers, even with a written guide, knowing which monkey is which is a difficult task. Currently, Auckland Zoo's main method of identifying the monkeys with 100% accuracy is by using a scanner on the monkey in the location where the microchip has been implanted. This, however, requires the scanner to be very close to the location of the microchip, and the scan requires a small duration to complete. Combined with the nature of the squirrel monkeys to be energetic, active and arboreal, scanning their chips is quite time-consuming.

The identification of conspecific individuals is an area where application of machine learning algorithms can greatly improve on current methods. These methods, like the tracking methods, will vary between species, but are, for the most part, still reliant on either older technology or having an expert on-hand with knowledge of the individuals and how to identify them. For some species, it is still reliant on manual study of images. With older technology, it does not necessarily provide an engine that can perform automatic identification [7]. Having an expert on-hand is an inconvenient requirement that is monetarily and temporally expensive. As such, we propose a method for automating the process of capturing an image of an animal and identifying that individual apart from other conspecific individuals to make the process more efficient.

We reached out to Auckland Zoo [8] and decided that their black-capped squirrel monkeys were a good candidate for an automated identification process. Therefore, our ultimate research question asks: "Given photos of individual squirrel monkeys (*S. boliviensis*), can we come up with a pipeline to automatically identify them?"

We answer this by using *S. boliviensis* data collected and provided to us by Auckland Zoo's primate team with a convolutional neural network (CNN) to segment the squirrel monkeys, such that a particular identifying visual feature is the main focus. This output is then given to another CNN-based model that can utilise one of two approaches to identification: 1) a VGG16-based classification approach [9], and 2) a metric learning approach [10]. Within the metric learning approach, two loss functions are tested: contrastive loss and triplet loss.

This report provides greater support for the use of machine learning models in conspecific animal identification, specifically for *S. boliviensis* individuals. Individual identification of them is difficult, and little work on integrating modern machine learning technologies into this has been done so far.

II. Related Work

Individual detection and identification of animals within a species generally requires trained professionals to carry out the job and involves time constraints, human error, and difficulties in data centralization. [11]–[13]

Deep Learning presents a sensible alternative with varying degrees of automation of the tedious process of animal detection and identification, making it possible to apply the same algorithms or pipelines to different datasets of the same individuals. Successful implementations remove the need for a specialist to be present at all times and allow organisations like Zoos to install camera feeds to track animal behavior outside of normal work hours of the personnel [14].

CNNs are very well suited to visual identification and learn to recognize and identify objects based on similar features, given enough situational examples over many iterations of training. Some of the most popular architectural choices for CNNs include Faster R-CNN, VGG, ResNet, Mask R-CNN and YOLO. CNNs are a natural choice for our project, but many options exist and selecting the best one is difficult. The size of the animal, how active it is, and what changes its features can go through are all relevant factors to consider, as well as whether the final application is intended to work in real-time. While fur matting is not a problem for our identification like it was for Crouse et al. [11] we discovered fur density to be a factor in how difficult it is for Mask R-CNN to segment out older, fluffier individuals, as face borders were less defined.

Mask R-CNN [15] is an extension of Faster R-CNN, with additional masking capabilities which allow segmentation of complex objects from the scene. It tends to be slower than YOLO [16], but has an advantage in terms of accuracy [17].

Use of faces as the most identifying feature for squirrel monkeys was supported by all the primate-focused individual identification papers [11], [13], [14], [18], [19]. Our inclusion of varying face angles, as well as the much smaller face sizes of squirrel monkeys compared to chimpanzees or gorillas were the main reasons for choosing Mask R-CNN as the detection model. In order to avoid including irrelevant information, the faces are cropped with Mask R-CNN before passing them to VGG16 for identification.

Automatic individual detection and identification pipelines range from simple implementations of one main model to complex stacks of them, each with a specific purpose. Brookes et al. [14] implement a YOLOv3 powered pipeline for both detection and identification of 7 gorillas and achieve 92% mAP on single frames. One of their main applications is real-time speed, allowing them to plan installations of gorilla surveillance cameras at their zoo. No data augmentation is done. Meanwhile Chen et al. [20] have 4 steps in their pipeline for panda identification, namely detection, segmentation, alignment and id prediction, all based on modified versions of ResNet-50. They augment their data with rotations and a greyscale filter. Ferreira et al. [12] automate data collection and labelling. Their pipeline consists of Mask R-CNN for detection and segmentation of individual birds and VGG19 for identification, at 93.4% overall accuracy. Mask R-CNN serves a vital purpose of getting rid of the background which takes up more space than the birds themselves, and is re-trained for this purpose using manually segmented images. They use the VGG Image Annotator, which is our choice for this project also due to its simplicity.

Guo et al [18] carry out identification over a massive amount of individuals from 41 primate and 4 carnivore species, using Faster R-CNN for detection and a tri-attention network for

identification with overall accuracy of 94.11%, at speeds of 50 images per second. While their tri-attention network improves on CNN models using 3 channels of attention, it is not freely available at the moment and was not considered a viable option for this project.

Individual identification of chimpanzees by Schofield et al [13] combines a Single Shot VGG16-based Detector and face tracking for detection and a VGG-M-based model for identification, using 50hrs of footage from 23 individuals. Precision goes from 81% on many face angles to 91% if only front views are used. Schofield et al. demonstrate the problems of identification on unconstrained data, where lighting, face angles and obstructions can vary greatly. They do not implement any data augmentation. We notice similar problems in our own dataset, where the Mask R-CNN model runs into problems with multiple individuals present in the scene, or has problems with overexposed frames. Despite that [13] demonstrates accuracy on real-world, complicated data, without simplifying the problem by getting rid of frames with multiple individuals like other papers do [12]. Additionally, VGG16 is simpler than VGG19 which should make our eventual mobile application lighter, possibly at the expense of some accuracy.

In general, while some papers used data augmentation for cleaner inputs and more data [12], [20] many others forwent this entirely [13], [14], [18]. Some common problems included misclassification of smaller individuals far away from the camera [14] or detection of faces at bad angles (profile or top-down views in our project). Occlusions by environment objects are another common occurrence [13], [18] which is evident in our dataset during videos of monkey feeding sessions, or simply because of foliage and how actively they move around. This project does not limit the data to only front-on faces as that severely limits the amount of faces picked up by the model, and limits the usability of the app we eventually hope to release for the Auckland Zoo.

III. Methodology[1]

A. Data

We identified that, based on previous work done on conspecifically identifying other animals [11], [12], [20], [21], including other primates [13], [14], [18], [19], a visual feature to focus on is the face. We expected that the data we would need would include many frames that would have an individual monkey in focus with its face unobscured. We suggested to Auckland Zoo that, for simplicity, shooting short videos of individual monkeys with a high-resolution camera, with a vocal identifier for the monkey at the beginning of the video, would be usable. The videos would not be expected to maintain focus on a squirrel monkey's face for every frame. Instead, the videos would have a high chance of containing frames that matched our criteria, and we could sample these frames from the videos to use for training, testing and evaluating our models.

The resulting data acquired from Auckland Zoo were six .zip archives (each corresponding to a unique squirrel monkey) containing 18-30 short .avi videos each. These were filmed by a squirrel monkey zookeeper at a resolution of 1080x1980 at 30 frames per second on an iPhone 11 Pro. Each video began with the zookeeper saying the name of the squirrel monkey being focused on.

A problem we have with the data is that it is in video form, and for the purposes of training the segmentation and identification models, the data needs to have a face in focus. Here, frame sampling is used to obtain the ideal images to train on. We decided that performing manual frame sampling on the data would net us the ideal set of frames for training with. Frames that were considered usable for training were typically focused on a monkey's face with high fidelity, regardless of the clarity of other components of the monkey or other objects in frame. These frames also had both eyes and a significant portion of the rest of the face visible. This excludes profile shots and top-down shots. The filter is subjective, due to the selection of frames being mediated by a human.

Conversion of the videos into still frames was done with the software DaVinci Resolve [22]. This renders the video files into a collection of frames spanning the length of the video (possibly with slight editing to cut the videos down to contain the frames that can be observed to fit our criteria at first glance). Frames were sampled by two team members at 30 frames per second (the native frame rate of the videos) and 16 frames per second respectively. 16 frames per second was chosen to limit the hard drive space the exported frames would occupy, as .tiff files are lossless [23] and the large number of frames quickly fills up a large portion of the drive's available capacity.

The resulting frames were then manually viewed and picked based on visibility of eyes and the rest of the face, focus and fidelity. Example frames for each of the 6 monkeys that were supplied to us by the Auckland Zoo can be found in Figure 1.

B. Segmentation

The first stage in the identification pipeline requires segmenting out faces from the frame and cropping to just the individual face of the monkey most prominent in the frame. To achieve this, in a similar method to previous works [11], [12], we employed a Mask R-CNN. This stage of the pipeline will output a cropped image exclusively containing the individual's face over a black background, for the next stage of our pipeline to utilize. To train Mask R-CNN to identify monkey faces as a class, a training set of 438 (60-80 per monkey) manually annotated and segmented images was used, created with the VGG Image Annotator [24]. Results were tested using a test set of 63 annotated and segmented images.

To implement the Mask R-CNN in Tensorflow, we used the popular Matterport implementation[2]. The Mask R-CNN architecture possesses a number of unique hyperparameters specific to the model. These include the choice of backbone,

[1]Our code is available on GitHub: Segmentation, Identification

[2]Matterport GitHub

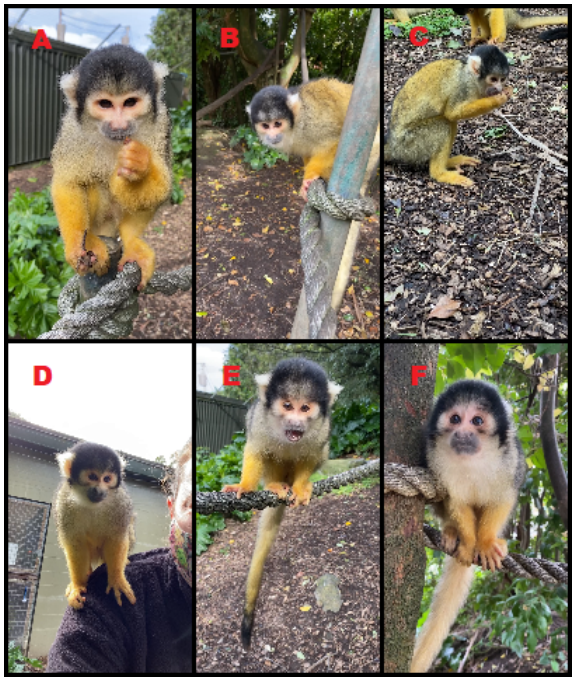

Fig. 1: Examples of sampled frames for the *S. boliviensis* squirrel monkeys: A) Arani, B) Inti, C) Ocuri, D) Poco, E) Rattaplan, F) Romy.

max instances, minimum detection confidence, image dimensions, and loss weights. As with other models, learning rate and epochs are also able to be tuned. Common choices for backbone include ResNet50, ResNet101, and ResNext101. We chose the ResNet101 backbone, pre-trained on the COCO (Common Objects in COntext)[25] dataset. Despite the COCO dataset not having a monkey class, the benefits of transfer learning have been apparent in similar implementations [11], [12]. The more complex architecture provided in the ResNet101 backbone, compared to ResNet50, creates greater discriminatory power at the cost of increased training time. The maximum number of instances was set to be 1 for our particular use case; the aim is for the model to detect and return only the most prominent individual in the frame.

The image dimension was fixed to 512x512, in contrast to the original paper [15], as this achieved the best balance between performance and efficiency.

To tune the learning rate, 3-fold cross validation was performed. Five epochs were used to train the final layers, and fine tune the entire model respectively for each fold. Image size was held at a reduced 256x256. This was done to increase training speed, and assumes a uniform decrease in accuracy within the search space. The mean average precision (mAP) for each fold was then calculated, and averaged to provide the performance evaluation for the learning rate. Minimum detection confidence and loss weights were all held at the default proposed in the original paper [15].

To evaluate the performance of the Mask R-CNN model, mean average precision was used. This metric is defined as

$$mAP = \sum_{q=1}^{Q} avgP(q)$$

where Q is the number of queries in the set, $avgP(q)$ is the average precision of the given query, q. Precision is defined as

$$Precision = \frac{TruePositive}{TruePositive + FalseNegative}$$

where a sample is classified as a true positive if the intersect over union (IoU) between the predicted and ground truth masks is greater than the set threshold. The intersect over union defines the pixelwise area of intersect between the ground truth and prediction bounding boxes, divided by the area of union, as illustrated in Figure 2.

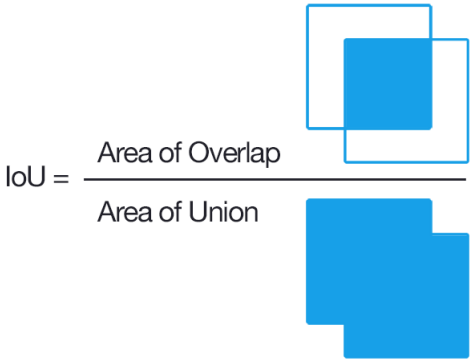

Fig. 2: Intersect over Union (Jaccard Index) [26]

Further to the baseline training of the model, two primary methodologies for improvements were identified: increasing the accuracy of true positives, and decreasing the false positive rate. To increase the accuracy of true positives, we take a boosting approach, retraining on those instances that were in fact instances of the monkey face class, but had IoU less than the chosen threshold, meaning a negative result. We progressively retrain on false negatives in the training set till accuracy in the validation set plateaus. To decrease the false positive rate, we explored adding images of the background class, especially frames of more problematic angles, such as the back of a monkey's head. Improving the model in this way proved challenging. Similar works confirm training on the background class appear to result in a decrease rather than increase in performance, seemingly as a quirk of the model's underlying Tensorflow implementation [27]. A possible amelioration would be averaging results between frames, or dropping low confidence frames at the identification phase, averaging or dropping the erroneous frame respectively.

C. Individual Identification

The survey [10] found that the metric learning approach outperformed the classification approach in identifying tiger individuals. The authors also found that metric learning with *triplet loss* had better performance than the *contrastive loss* in

recent studies of a diverse group of species (humans, chimpanzees, humpback whales, fruit flies, and Siberian tigers). This paper implemented the metric learning approach using *triplet loss with k-Nearest-Neighbours* (kNN) to identify six individual monkeys from Auckland Zoo and test the authors' claims. We discuss more detail in the section of results. In addition, we introduce a novel way of open set recognition to identify new individuals.

1) Dataset: In this paper, we have six monkey individuals from Auckland Zoo. After we use Mask R-CNN to segment these six monkeys' faces from the videos (provided by Auckland Zoo), we have $1,837$ images of monkey faces (*Fig.* 3), respectively, 240 images of Arani, 411 images of Inti, 530 images of Ocuri, 285 images of Poco, 157 images of Rattaplan and 214 images of Romy as the dataset is imbalanced with an unequal class distribution. We downsample the number of images from Arani, Inti, Ocuri, Poco and Romy to 157 to make the dataset balanced.

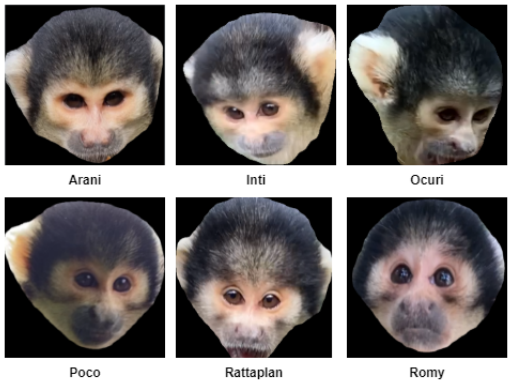

Fig. 3: Segmented monkey faces images for six monkey individuals

2) Implementation of Triplet Loss with kNN: We propose the metric learning approach using *triplet loss* with the *k-nearest neighbours classifier* with the monkey dataset. [28] introduced *triplet loss* with FaceNet, which preserves the similarity and dissimilarity of the images in a lower-dimensional space. This approach is promising in identifying monkey individuals by their faces. First, we created the triplets containing images of anchor individuals, positive individuals and negative individuals by *online triplet mining*. We aim to minimise the Euclidean distance between the positive pairs (i.e. anchor image and positive image) and maximise the distance between the negative pair (i.e. anchor image and negative image) simultaneously (*Fig.* 4), and optimise the weights and biases through backpropagation by minimising the *triplet loss* (*Eq.* 1).

$$L_t = \sum_{i=1}^{N} \left[\left\| f(x_i^a) - f(x_i^p) \right\|_2^2 - \left\| f(x_i^a) - f(x_i^n) \right\|_2^2 + \alpha \right]_+ \tag{1}$$

where $f(x_i) : \mathbb{R}^D \longrightarrow \mathbb{R}^d$, where $d \ll D$, which is an embedding function. x^a is an anchor input image, x^p is an input image from the same class of the anchor image, and x^n is an input image from a different class. α is a margin, to satisfy the following inequality (*Eq.* 2) to separate positive pairs from negative pairs,

$$\left\| f(x_i^a) - f(x_i^p) \right\|_2^2 + \alpha < \left\| f(x_i^a) - f(x_i^n) \right\|_2^2 \tag{2}$$

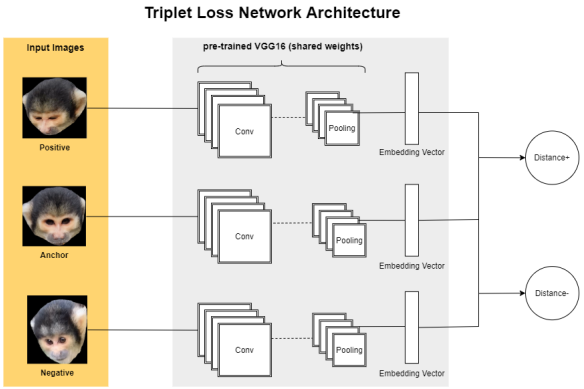

Fig. 4: Triplet Loss Network Architecture

Once we learn an embedding space by optimising the loss, we use the *k-Nearest Neighbours classifier* to classify the query image (*Fig.* 5).

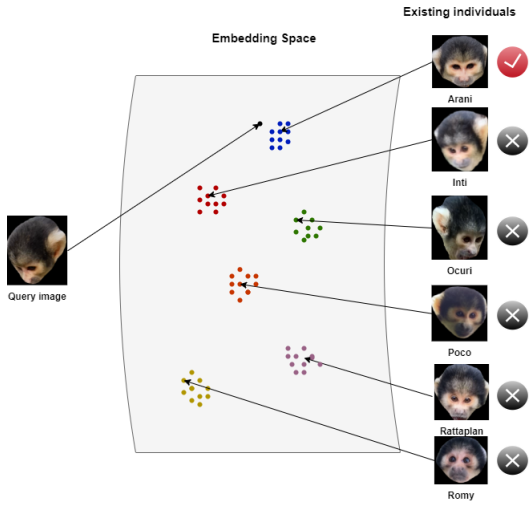

Fig. 5: New query image given (existing individual) in embedding space

3) Open Set Recognition: This paper also considers the identification of new individuals as an extension of our approach. We set up a distance threshold to find similarities and dissimilarities between query images and images of each existing individual, i.e. they are dissimilar if the distance between a query image and image of an existing individual is larger than the threshold or vice versa. We compute the distances between a query image and the center points of each existing individual in the embedding space. If all distances are larger than the threshold, we classify the query image

as a new individual (*Fig.* 6). Selecting a proper threshold is crucial, and we found the distance threshold is a trade-off between accuracy of classifying existing individuals and new individuals, i.e. the larger value of the threshold, the less accuracy of the new individual and the higher accuracy of the existing individual or vice versa. We discuss more detail in the Results section.

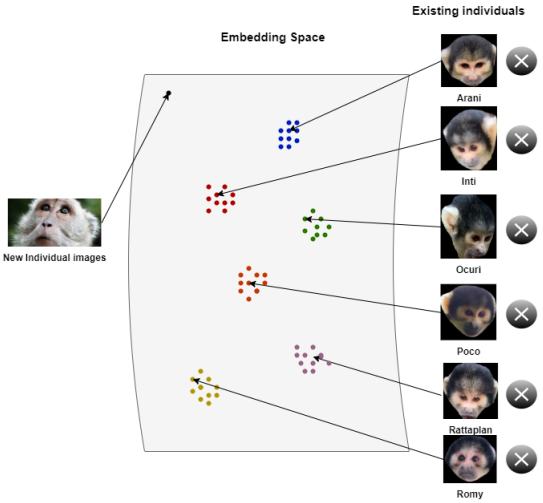

Fig. 6: New query image (new individual/unknown) given in embedding space

IV. Results

A. Segmentation

Before adding further training and improvements to the Mask R-CNN model, our first evaluation was on the base hyperparameters. We performed 3-fold cross validation to determine the best learning rate, from the search space of [0.01, 0.001, 0.0001]. The models were trained for 5 epochs for both the head layers and the whole model, on a 50 image dataset. The results for this are shown in Table I. Despite the indications of these findings, the learning rate of 0.001 was selected as the learning rate for further training. This is the default learning rate for MaskR-CNN, with its mediocre performance in testing likely due to insufficient testing epochs. This small number of folds, epochs, and images were a limitation of training capacity; the large model was prohibitive in both its training lengths and memory consumption, often leading to out of memory errors on the available GPUs[3].

Learning Rate	mAP @ 0.9 IoU
0.01	0.57
0.001	0.23
0.0001	0.0

TABLE I: Table showing mAP values from 3-fold cross validation

Using a base Mask R-CNN model pre-trained on COCO, our model was re-trained and assessed incrementally after a various number of epochs, with the aim of finding the best model being mindful of the risk of overfitting. Training consisted of two stages, during the first stage only the randomly initialised weights of the head layer are trained with all backbone layers frozen, then secondly all layers are trained and fine-tuned using a decreased learning rate (reduced by a factor of 10). Each stage ran on the same number of epochs as the other. The accuracy of the model was determined using mAP with a positive result identified as being above an IoU threshold of 0.9. This value was set rather high so that only full masks, without areas of the face missing, are to be accepted. Measurements were taken for both the validation set and testing set.

Epoch	mAP (val. set)	mAP (test set)
5	0.717	0.587
10	0.767	0.683
15	0.767	0.730
20	0.8	0.667
25	0.8	0.730
30	0.833	0.714
35	0.8	0.714
40	0.817	0.698
50	0.833	0.714

TABLE II: Table showing mAP values (at an IoU threshold of 0.9) calculated on validation and test sets for each incremental epoch run

Table II shows the results of training over various increments. After training for at least 25 epochs, the model was able to predict a highly accurate mask for 80% of images in the validation set and 70% of images in the test set.

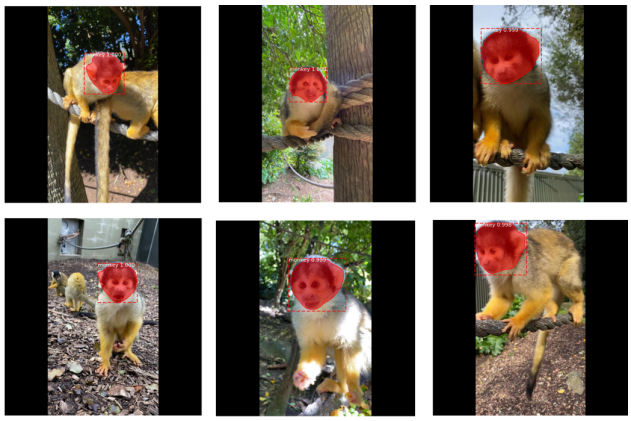

Fig. 7: Examples of predicted masks for the test set, output by our Mack R-CNN model

Figure 7 shows a visualisation of the predictions in the test set. It is able to correctly detect without fail a squirrel monkey face in our test set images, with a very high degree of confidence (at least 0.99 out of 1) and predict a fairly suitable mask of the face.

Additionally, as our test set consisted of carefully selected pictures where there is always a face visible, we have also

[3]NVIDIA GTX 1070

Fig. 8: Examples of problematic masks showing false positives on sub-optimal images

decided to test the model on sub-optimal images where the monkey face is partially visible or not visible at all, such as in frames of an unedited video for a more realistic representation of our intended use case. On these images, the model's predictions do contain several false positives, where the back of the head and portions of the torso are occasionally identified as faces, but exclusively for instances where the image does not clearly show a face. Figure 8 shows some examples. Our model did not output any false negatives, i.e. where a face exists but is not detected. Therefore, as long as an image contains a face, our model is able to confidently and accurately identify the face and create a acceptable segmentation.

B. Individual Identification

We have two manifolds in the experiments. *i) closed set*, all individuals appear in the test set, all of which exist in the training set. *ii) open set*, some individuals are in the test set, which does not appear in the training set. This setting is more akin to a real-world situation, where there are individuals unknown to the model.

1) Closed Set: In the closed set setting, we used a pre-trained *VGG16 classification model* with freezing layers (i.e. not trainable weights and biases) and *contrastive loss with kNN* as our baseline models, compared to our proposed model: *triplet loss with kNN*. We split the dataset into training and test set first, and the split ratio is 80:20. *ADAM* is our optimisation function with 10 epochs and batch size of 64. The input images are embedded into 128-dimensional space for metric learning using *contrastive loss* and *triplet loss*. We also tune hyperparameters of the learning rate (*LR*) for these three models and the k value of *kNN* for both *contrastive loss* and *triplet loss* by 5-fold cross-validation (*Fig.* 9). We select the search space of *LR*, respectively, 0.00001, 0.0001, 0.001, 0.01, and 0.1. The search space of the k value of *kNN* is [5, 11, 21, 31].

Once we fine-tuned these hyperparameters (*Table.* III), we found that the *VGG16 classifier* ($LR = 0.0001$) achieves $79.82\,\% \pm 3.91\,\%$, *contrastive loss with kNN* ($LR = 0.0001, k = 11$) achieves $79.6\,\% \pm 2.2\,\%$, and *triplet loss with kNN* ($LR = 0.001, k = 5$) achieves $86.4\,\% \pm 2.7\,\%$. (*Fig.* 10). Based on the post-hoc Turkey HSD statistical test at 95% confidence, there is no significant difference between the accuracy of the *VGG16 classifier* and the accuracy of the *contrastive loss* with *kNN*. But *triplet loss with kNN* has the highest accuracy. This evidence supports the best solution of those proposed in this paper.

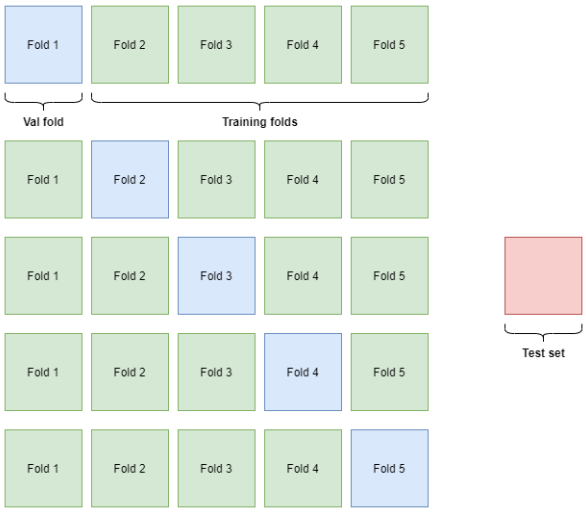

Fig. 9: k-fold Cross Validation $k = 5$

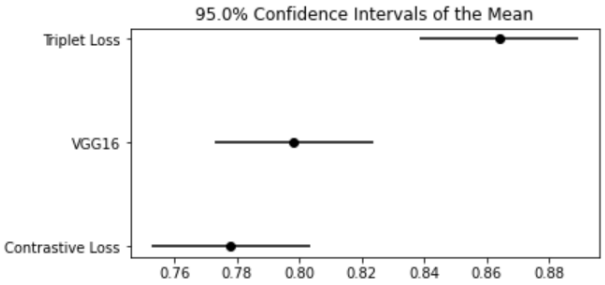

Fig. 10: Plot of accuracy for each model by 5-fold Cross Validation

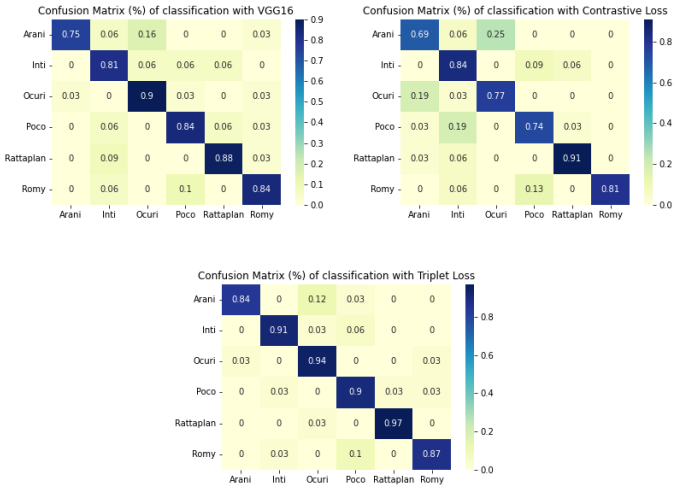

Fig. 11: Confusion Matrices: accuracy for each monkey individual for each monkey individual on the test set

	VGG16	Contrastive Loss	Triplet Loss
Accuracy	$79.82\,\% \pm 3.91\,\%$	$79.6\,\% \pm 2.2\,\%$	$86.4\,\% \pm 2.7\,\%$

TABLE III: Table of mean and standard deviation of accuracy for each model by 5-fold Cross Validation. *VGG16* ($LR = 0.0001$), *Contrastive Loss* ($LR = 0.0001, k = 11$), *Triplet Loss* ($LR = 0.001, k = 5$)

In addition, we evaluate the models on the test set and create confusion matrices to find which monkey individual is difficult to identify. (*Fig.* 11). We found Arani is most difficult to identify over three different models, respectively $75\,\%$ (*VGG16*), $69\,\%$ (*contrastive loss*), and $84\,\%$ (*triplet loss*) compared to the others. The majority of mis-classified cases is identified as Ocuri. Again, we found *triplet loss* outperformed the other approaches for classifying each individual.

2) Open Set: We introduce our proposed model, *triplet loss* with *kNN*, to open-set tasks, excluding the *VGG16 classification model*. *VGG16 classifier* does not have the capability to classify new individuals since the number of outputs of the *VGG16* classifier is fixed, therefore the *VGG16 classifier* is not tested on the open set.

We inherit the fine-tuned *triplet loss with kNN* ($LR = 0.001, k = 5$) from the closed set, and introduce a new hyperparameter distance threshold (*d_t*), which defines how far a distance is, if two images are dissimilar. If all distances between a query image and center points for each existing individual is above the threshold, we classify the query image as a new individual.

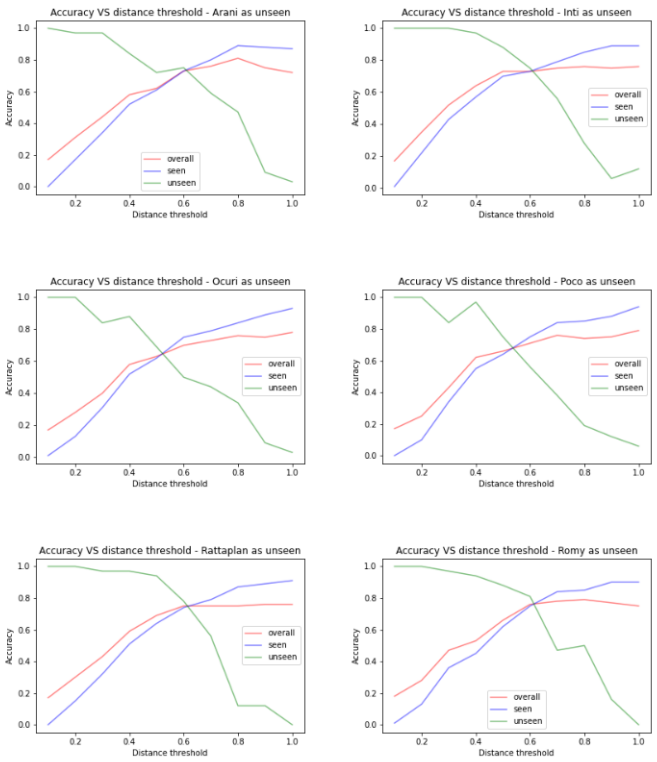

Fig. 12: Plot of accuracy of overall, *seen* and *unseen* over different distance threshold

The search space of *d_t* is [0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1], and we plug each *d_t* to evaluate the models on the test set by accuracy of an *overall*, *seen* (existing individuals), and *unseen* (new individual). We repeat the step six times, and each time we remove a different individual from the training set. (*Fig.* 12). We can see the accuracy of *unseen* (new individual) decrease and accuracy of *seen* (existing individual) increase when we increase the threshold. Intersections of the accuracy of *unseen* and *seen* are of interest to us, at 0.5 or 0.6 of the threshold, around $70\,\%$ accuracy. Therefore, we claim 0.5 or 0.6 is an ideal value of the threshold to keep a good balance between *seen* and *unseen* accuracy.

V. Conclusion

We have explored the possibility of coming up with a pipeline to automatically identify individual squirrel monkeys (*S. boliviensis*) given a photograph, showing it to be possible. We investigated various configurations of the Mask R-CNN segmentation model, as well as variations on the VGG16 classification model. To achieve this, we trained and evaluated a Mask R-CNN using multiple different hyperparameter configurations, and provided further specialised domain-specific training to increase performance. Similarly, we performed stratified k-fold cross validation on a VGG16 base architecture, extended with contrastive and triplet loss to improve its accuracy. Ultimately, we achieved a mAP@[0.9] of 0.730 for the Mask R-CNN and a closed-set accuracy of 86.4% using triplet loss.

Accuracy of the identification model is reasonably good, but there is room for improvement. Data augmentation was ultimately not applied here due to time constraints and frequent out-of-memory issues. Adding a data augmentation layer to the identification model would automatically apply augmentation to the images of the monkey faces. Furthermore, choosing optimal parameters for the Mask R-CNN model proved difficult, considering the immense training cost of training the model from scratch. Developing an efficient testing methodology would facilitate further tuning of these parameters. Collectively, this would likely increase the accuracy of the identification model. Having stronger hardware available would make the training process much quicker and less likely to encounter out-of-memory errors. The open set recognition is another challenge that we need to address carefully for further improvement.

Further work could also explore the temporal element of identification. Our current pipeline identifies frames in isolation. However, when viewing an individual, consecutive frames are likely to be of the same individual at a high enough frame rate. Accordingly, a sliding window average of predictions could be incorporated. This would reduce the effect of any inaccurate segmentations by the MaskR-CNN, or misclassifications by the Triplet Loss VGG model. Other approaches could be to use a face-tracking system such as that proposed by Schofield et. al. [13]. Further to the aforementioned accuracy improvements, this approach would additionally allow for identification of multiple individuals per

frame, rather than our approach which is limited to a single individual per frame.

Finally, conspecific identification in other species is a natural progression from our research, both in other primates and different genera entirely. Of particular interest is the Kororā (Little Blue Penguin), whose nesting habits is presently difficult to track in remote colonies in New Zealand, or other penguin species such as those in Antarctica[29].

To be truly beneficial, our findings must be incorporated into a user-friendly and portable application. Often in the wild where conservation work is conducted, accessing a computer is difficult if not infeasible. Accordingly, we propose further work in converting this pipeline into a lighter version, able to be run on a mobile device for easy identification. This would utilise the mobile device's camera to capture video footage of a *S. boliviensis* squirrel monkey individual in real-time, running through our proposed pipeline to get from the video (as raw input) to identifying the individual by name.

Contributions

Adam	• Abstract • Mask R-CNN Implementation (Research and Code) • Mask R-CNN Methodologies for Report • Conclusion • Report Proofing
Addison	• Experiments and results — Segmentation • Mask R-CNN Implementation (Research and Code) • Report Proofing
Daria	• Related Work • Manual Segmentation of Monkey Frames • Labelling and Annotation of frames for Mask R-CNN (VGG Image Annotator) • Report Proofing
Ian	• Introduction • Methodology — Data • Conclusion • Manual Segmentation of Monkey Frames • Report Proofing
Jovi	• Methodology — Individual Identification • Experiments and results — Individual Identification (data analysis and data visualization) • Metric Learning implementation using Contrastive Loss and Triplet Loss (Research and Code)

References

[1] T. M. Brooks, R. A. Mittermeier, C. G. Mittermeier, *et al.*, "Habitat loss and extinction in the hotspots of biodiversity," *Conservation Biology*, vol. 16, no. 4, pp. 909–923, 2002. DOI: 10.1046/j.1523-1739.2002.00530.x.

[2] P. M. Vitousek, C. M. D'Antonio, L. L. Loope, M. Rejmánek, and R. Westbrooks, "Introduced species: A significant component of human-caused global change," *New Zealand Journal of Ecology*, vol. 21, no. 1, pp. 1–16, 1997, ISSN: 01106465, 11777788. [Online]. Available: http://www.jstor.org/stable/24054520.

[3] A. S. Voulodimos, C. Z. Patrikakis, A. B. Sideridis, V. A. Ntafis, and E. M. Xylouri, "A complete farm management system based on animal identification using rfid technology," *Computers and Electronics in Agriculture*, vol. 70, no. 2, pp. 380–388, 2010, Special issue on Information and Communication Technologies in Bio and Earth Sciences, ISSN: 0168-1699. DOI: https://doi.org/10.1016/j.compag.2009.07.009. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0168169909001392.

[4] M. Mrozek, R. Fischer, M. Trendelenburg, and U. Zillmann, "Microchip implant system used for animal identification in laboratory rabbits, guineapigs, woodchucks and in amphibians," *Laboratory Animals*, vol. 29, no. 3, pp. 339–344, 1995. DOI: 10.1258/002367795781088298.

[5] T. Caceci, S. A. Smith, T. E. Toth, R. B. Duncan, and S. C. Walker, "Identification of individual prawns with implanted microchip transponders," *Aquaculture*, vol. 180, no. 1-2, pp. 41–51, 1999. DOI: 10.1016/s0044-8486(99)00144-1.

[6] W. Rossing, "Animal identification: Introduction and history," *Computers and Electronics in Agriculture*, vol. 24, no. 1, pp. 1–4, 1999, ISSN: 0168-1699. DOI: https://doi.org/10.1016/S0168-1699(99)00033-2. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0168169999000332.

[7] A.-M. Kitchen-Wheeler, "Visual identification of individual manta ray (manta alfredi) in the maldives islands, western indian ocean," *Marine Biology Research*, vol. 6, no. 4, pp. 351–363, 2010. DOI: 10.1080/17451000903233763.

[8] A. Robbins, *Re: Research Opportunity with Auckland Zoo*, Aug. 2021.

[9] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," *arXiv:1409.1556 [cs]*, Apr. 2015, arXiv: 1409.1556 version: 6. [Online]. Available: http://arxiv.org/abs/1409.1556 (visited on 09/16/2021).

[10] M. Vidal, N. Wolf, B. Rosenberg, B. P. Harris, and A. Mathis, "Perspectives on individual animal identification from biology and computer vision," *arXiv:2103.00560 [cs, q-bio]*, Feb. 2021, arXiv: 2103.00560. [Online]. Available: http://arxiv.org/abs/2103.00560.

[11] D. Crouse, R. L. Jacobs, Z. Richardson, *et al.*, "LemurFaceID: A face recognition system to facilitate individual identification of lemurs," *BMC Zoology*, vol. 2, no. 1, p. 2, Dec. 2017, ISSN: 2056-3132. DOI: 10.1186/s40850-016-0011-9.

[12] A. C. Ferreira, L. R. Silva, F. Renna, *et al.*, "Deep learning-based methods for individual recognition in small birds," *Methods in Ecology and Evolution*, vol. 11, no. 9, E. Codling, Ed., pp. 1072–1085, Sep. 2020, ISSN: 2041-210X, 2041-210X. DOI: 10.1111/2041-210X.13436.

[13] D. Schofield, A. Nagrani, A. Zisserman, *et al.*, "Chimpanzee face recognition from videos in the wild using deep learning," *Science Advances*, vol. 5, no. 9, eaaw0736, Sep. 2019, ISSN: 2375-2548. DOI: 10.1126/sciadv.aaw0736. [Online]. Available: https://advances.sciencemag.org/lookup/doi/10.1126/sciadv.aaw0736 (visited on 09/17/2021).

[14] O. Brookes and T. Burghardt, "A Dataset and Application for Facial Recognition of Individual Gorillas in Zoo Environments," *arXiv:2012.04689 [cs]*, Mar. 2021, arXiv: 2012.04689. [Online]. Available: http://arxiv.org/abs/2012.04689 (visited on 09/17/2021).

[15] K. He, G. Gkioxari, P. Dollár, and R. Girshick, "Mask R-CNN," *arXiv:1703.06870 [cs]*, Jan. 2018. [Online]. Available: http://arxiv.org/abs/1703.06870.

[16] J. Redmon and A. Farhadi, "YOLOv3: An Incremental Improvement," *arXiv:1804.02767 [cs]*, Apr. 2018, arXiv: 1804.02767. [Online]. Available: http://arxiv.org/abs/1804.02767 (visited on 09/17/2021).

[17] J. Huang, V. Rathod, C. Sun, *et al.*, "Speed/accuracy trade-offs for modern convolutional object detectors," *arXiv:1611.10012 [cs]*, Apr. 2017, arXiv: 1611.10012. [Online]. Available: http://arxiv.org/abs/1611.10012 (visited on 09/17/2021).

[18] S. Guo, P. Xu, Q. Miao, *et al.*, "Automatic Identification of Individual Primates with Deep Learning Techniques," en, *iScience*, vol. 23, no. 8, p. 101 412, Aug. 2020, ISSN: 25890042. DOI: 10.1016/j.isci.2020.101412. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S2589004220306027 (visited on 09/17/2021).

[19] C. L. Witham, "Automated face recognition of rhesus macaques," en, *Journal of Neuroscience Methods*, vol. 300, pp. 157–165, Apr. 2018, ISSN: 01650270. DOI: 10.1016/j.jneumeth.2017.07.020. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0165027017302637 (visited on 09/17/2021).

[20] P. Chen, P. Swarup, W. M. Matkowski, *et al.*, "A study on giant panda recognition based on images of a large proportion of captive pandas," en, *Ecology and Evolution*, vol. 10, no. 7, pp. 3561–3573, Apr. 2020, ISSN: 2045-7758, 2045-7758. DOI: 10.1002/ece3.6152. [Online]. Available: https://onlinelibrary.wiley.com/doi/10.1002/ece3.6152 (visited on 09/17/2021).

[21] M. Clapham, E. Miller, M. Nguyen, and C. T. Darimont, "Automated facial recognition for wildlife that lack unique markings: A deep learning approach for brown bears," *Ecology and Evolution*, vol. 10, no. 23, pp. 12 883–12 892, Dec. 2020, ISSN: 2045-7758, 2045-7758. DOI: 10.1002/ece3.6840. [Online]. Available: https://onlinelibrary.wiley.com/doi/10.1002/ece3.6840 (visited on 09/17/2021).

[22] *Davinci resolve 17*. [Online]. Available: https://www.blackmagicdesign.com/products/davinciresolve/.

[23] *TIFF: Revision 6.0*. Adobe Developers Association, 1995. [Online]. Available: https://www.adobe.io/content/dam/udp/en/open/standards/tiff/TIFF6.pdf.

[24] A. Dutta and A. Zisserman, "The VIA annotation software for images, audio and video," in *Proceedings of the 27th ACM International Conference on Multimedia*, ser. MM '19, Nice, France: ACM, 2019, ISBN: 978-1-4503-6889-6/19/10. DOI: 10.1145/3343031.3350535. [Online]. Available: https://doi.org/10.1145/3343031.3350535.

[25] T.-Y. Lin, M. Maire, S. Belongie, *et al.*, "Microsoft COCO: Common Objects in Context," *arXiv:1405.0312 [cs]*, Feb. 2015, arXiv: 1405.0312. [Online]. Available: http://arxiv.org/abs/1405.0312 (visited on 10/22/2021).

[26] *Intersection over Union (Iou) for object detection*, en-US, Nov. 2016. [Online]. Available: https://www.pyimagesearch.com/2016/11/07/intersection-over-union-iou-for-object-detection/ (visited on 10/21/2021).

[27] *Handle zero pad & Allow bg to train RPN by keineahnung2345 · Pull Request #1088 · matterport/Mask_rcnn*, en. [Online]. Available: https://github.com/matterport/Mask_RCNN/pull/1088 (visited on 10/22/2021).

[28] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: A unified embedding for face recognition and clustering," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, MA, USA: IEEE, Jun. 2015, pp. 815–823, ISBN: 9781467369640. DOI: 10.1109/CVPR.2015.7298682.

[29] P. T. Fretwell and P. N. Trathan, "Discovery of new colonies by Sentinel2 reveals good and bad news for emperor penguins," en, *Remote Sensing in Ecology and Conservation*, vol. 7, no. 2, pp. 139–153, 2021, ISSN: 2056-3485. DOI: 10.1002/rse2.176. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/rse2.176 (visited on 10/22/2021).