PriMAT: A robust multi-animal tracking model for primates in the wild

Richard Vogg[1], Matthias Nuske[2], Marissa A. Weis[1], Timo Lüddecke[1], Elif Karakoç[3], Zurna Ahmed[4], Sofia M. Pereira[5,6], Suchinda Malaivijitnond[7,8], Suthirote Meesawat[7], Florentin Wörgötter[2,9,10], Peter M. Kappeler[3,9,11], Alexander Gail[4,9,10,12], Julia Ostner[5,6,9], Oliver Schülke[5,6,9], Claudia Fichtel[3,9], Alexander S. Ecker[1,9,10,13,*]

August 21, 2024

[1]Institute of Computer Science and Campus Institute Data Science, University of Göttingen
[2]Department for Computational Neuroscience, Third Physics Institute, University of Göttingen
[3]Behavioral Ecology & Sociobiology Unit, German Primate Center, Leibniz Institute for Primate Research, Göttingen
[4]Sensorimotor Group, German Primate Center, Leibniz Institute for Primate Research, Göttingen
[5]Behavioral Ecology Department, University of Göttingen
[6]Social Evolution in Primates Group, German Primate Center, Leibniz Institute for Primate Research, Göttingen
[7]National Primate Research Center of Thailand, Chulalongkorn University, Saraburi, Thailand
[8]Department of Biology, Faculty of Science, Chulalongkorn University, Bangkok, Thailand
[9]Leibniz ScienceCampus, German Primate Center, Leibniz Institute for Primate Research, Göttingen
[10]Bernstein Center for Computational Neuroscience, University of Göttingen
[11]Department of Sociobiology/Anthropology, University of Göttingen
[12]Georg-Elias-Müller-Institute of Psychology, University of Göttingen
[13]Max Planck Institute for Dynamics and Self-Organization, Göttingen
[*]corresponding author: ecker@cs.uni-goettingen.de

Abstract

1. Detection and tracking of animals is an important first step for automated behavioral studies in videos. Nowadays, animal tracking is mostly done using deep learning frameworks based on keypoints, which show remarkable results in lab settings with fixed cameras, backgrounds, and lighting. In contrast, multi-animal tracking in the wild is currently underexplored as it presents numerous challenges such as high background variability, lighting conditions, complex motion, and occlusion.

2. We propose a multi-animal tracking model, PriMAT, for nonhuman primates in the wild. The model learns to detect and track primates and other objects of interest from labeled videos or single images using bounding boxes instead of keypoints. Using bounding boxes significantly facilitates data annotation and allows for more robust results. Our one-stage model is conceptually simple but highly flexible, and we add a classification branch that allows us to train individual identification.

3. To evaluate the performance of our model, we applied it in two case studies with Assamese macaques and redfronted lemurs in the wild. We show that with only a few hundred frames labeled with bounding boxes, we can achieve robust tracking results. Combining these results with the classification branch, our model shows an accuracy of 83% in predicting lemur identities.

4. Our approach presents a promising solution for accurately tracking and identifying animals in the wild using deep learning based bounding box tracking, offering researchers a tool to study animal behavior in their natural habitats. Our code, models, training images, and evaluation video sequences are publicly available (https://github.com/ecker-lab/PriMAT-tracking), facilitating its use for animal behavior analyses and future research in this field.

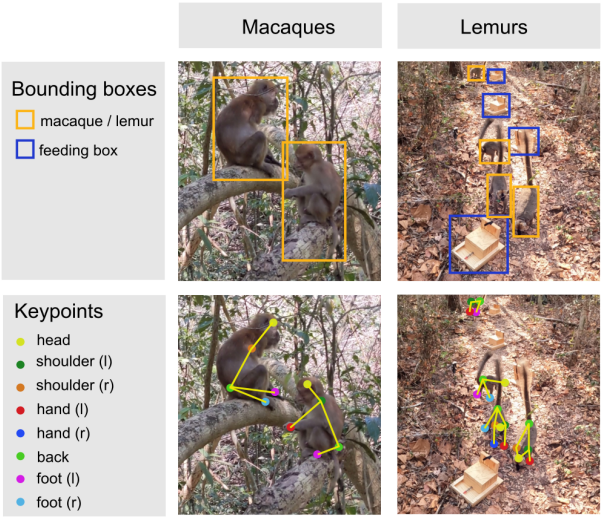

Figure 1: Bounding box based detection vs. keypoint detection on different datasets. Bounding box based tracking has advantages over keypoint tracking when used as a starting point for automated behavior analysis, such as reduced labeling time, higher tracking robustness for videos from the wild, and universal extensibility to other objects of interest.

1 Introduction

Video recordings have been a valuable tool for studying the behavior of nonhuman primates for a long time. Traditionally, researchers have analyzed videos manually, requiring a significant amount of time and effort. In recent years, the availability of large-scale annotated datasets enabled the development and training of deep learning based models (Bala et al., 2020; Brookes and Burghardt, 2020; Labuguen et al., 2021; Yu et al., 2021; Desai et al., 2023; Yang et al., 2022; Ng et al., 2022; Ma et al., 2023). There are models which show promising results on several computer vision tasks in the wild, such as individual identification (Schofield et al., 2019; Bain et al., 2019; Schofield et al., 2023; Paulet et al., 2024), pose estimation (Wiltshire et al., 2023; Marks et al., 2022), object detection (Roy et al., 2023) and tracking (Koger et al., 2023; Pineda et al., 2023). However, most current frameworks used to analyze animal behavior are designed to work in lab settings and rely on pose estimation as an intermediate representation (Vogg et al., 2024). Three of the most used frameworks for such markerless pose estimation are DeepPoseKit (Graving et al., 2019), DeepLabCut (Mathis et al., 2018; Lauer et al., 2022), and SLEAP (Pereira et al., 2022). They allow for multi-animal pose tracking and have user-friendly interfaces that help with annotation and inference. The introduction of these frameworks has contributed to great progress in the area of behavioral analysis.

As valuable as these tools are in the lab, their utility for applications of tracking animals in the wild is limited (Mathis et al., 2021; Perez and Toler-Franklin, 2023). Recording animals in their natural habitat presents challenges such as occlusion, diverse backgrounds, and varying lighting conditions. In addition, the fluidity of the natural environment makes it difficult to track individuals consistently, especially with rapid movements and similar appearances among them. We propose a different path towards behavioral analysis of videos from the wild. Instead of keypoints as an intermediate representation of the objects, our model outputs bounding boxes for all animals and relevant objects in the scene (Fig. 1).

We see three main reasons why object detection via bounding boxes should be a starting point to consider for many applications. First, currently, there are few successful applications of the above-mentioned keypoint-based frameworks on videos from the wild. DeepWild showed that DeepLabCut can learn from around 2000 labeled frames from videos of animals in the wild; however, the results on unseen videos leave room for improvement (Wiltshire et al., 2023) and more labels would be necessary to enhance generalization. A study conducted on horse images showed the limited capability of keypoint-based models to generalize to new, unseen horses (Mathis et al., 2021). A recent survey on keypoint-based methods for animal behavior found that most models do not generalize well to in-the-wild conditions (Perez and Toler-Franklin, 2023). Second, the annotation time for keypoint-based models

is a magnitude larger than for bounding boxes. Wiltshire et al. (Wiltshire et al., 2023) reported that they needed around two hours to label 18 keypoints for all animals in 25 frames. Third, many questions in animal behavior research, especially in the wild, may not require keypoints but instead focus on actions and interactions.

In this paper, we propose PriMAT, a conceptually simple **M**ulti-**A**nimal **T**racking model for **Pri**mates. It builds on the FairMOT model (Zhang et al., 2021) which is trained on individual images and thus does not require entire video sequences to be labeled. Compared to keypoint labeling, it reduces annotation efforts further, as bounding boxes can be drawn fast on a few hundred frames which is enough to do transfer learning to other primate species or camera angles. We added a branch to our model that can learn classification tasks on top of the tracking results such as individual identification. We present two use cases for our model, by applying it to videos of Assamese macaques and redfronted lemurs in the wild. The models learned to detect macaques and lemurs from different camera angles and in videos with highly variable backgrounds and lighting conditions. Additionally, we show two extensions to the base model: First, multi-class tracking allows to detect other objects of interest in the video. We used it to detect lemurs and food boxes during the experiments, which is a crucial step in analyzing interactions. Secondly, we used the introduced classification branch to identify individuals.

2 Methods

2.1 Data

Manual annotations are time-consuming, even though bounding boxes are faster to draw than keypoints. Transfer learning is an efficient way to reduce annotation efforts, by starting with a base model which has been pre-trained on a larger dataset, and fine-tuning with a few hundred annotated frames from the domain of application. We experimented with different pre-training datasets and annotated 500 frames to showcase the applicability of the base model to macaques and lemurs (Fig. 2A). We used tight modal bounding boxes, i.e. the boxes located around the extreme visible pixels of each individual.

Pre-training. Training a model from scratch usually requires large amounts of data. Therefore, we compared different datasets for pre-training our base model. ImageNet (Deng et al., 2009) is a standard dataset for pre-training, containing over 1.2M images from a variety of object classes. We experimented with whether pre-training with the MacaquePose dataset (Labuguen et al., 2021) after ImageNet pre-training would improve performance. MacaquePose consists of 13083 images of macaques mostly taken from zoos. The dataset is completely annotated with instance segmentation masks and keypoints. We constructed bounding boxes from the extreme points of each segmentation mask. We used 80% of the dataset for training the model and kept the remaining images to evaluate the detection performance. To improve generalization, we used the labeled instance segmentation masks, copied the individuals and pasted them to backgrounds from ImageNet (Deng et al., 2009) (Fig. 5A). The resulting images formed a dataset with high background diversity, which we called MacaqueCopyPaste. Additionally, we created MacaqueCopyPasteWild, a dataset where we pasted the monkeys on background images taken in the Phu Khieo Wildlife Sanctuary in Thailand. We later evaluated the performance of the base models on two use cases.

Assamese Macaques. Data on Assamese macaques were collected as part of a long-term project on the behavioral ecology of wild Assamese macaques in the Phu Khieo Wildlife Sanctuary, North East Thailand (Schülke et al., 2011). Data were collected on approximately 250 individually recognizable subjects of different age-sex classes organized in up to five large multi-male/multi-female social groups ranging in size from 40 to 73 individuals (unpublished data). The habitat consists of dense hill evergreen forest, and the monkeys spend the majority of their time in the trees, reducing visibility and creating difficult video-recording conditions. The videos were hand-recorded with resolutions of 1920×1080 using a GoPro Hero 10 camera and 3840×2160 using a Google Pixel 4 camera. More than 300 video snippets have been obtained so far. The videos consist of sequences that are challenging for a tracking model, as they can contain motion blur through camera or animal movement, severe occlusion, and high variability in backgrounds and lighting conditions.

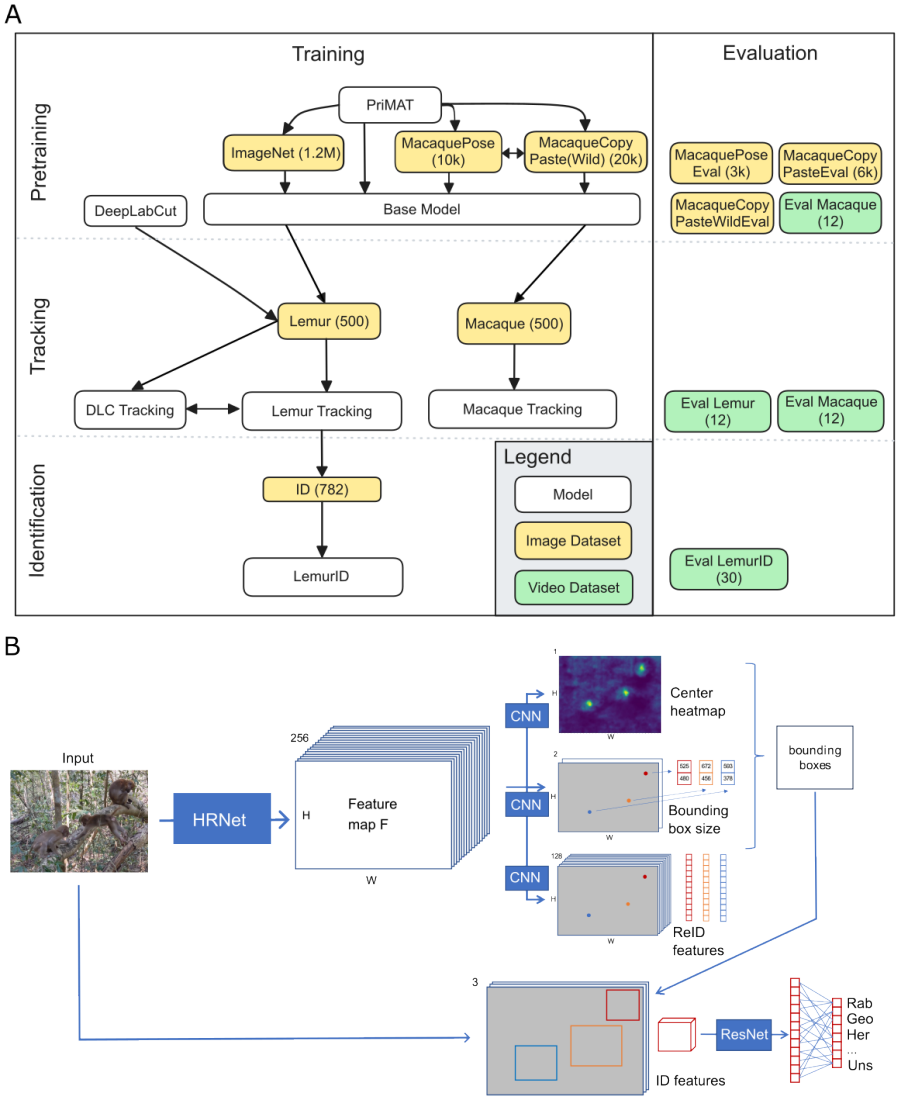

Figure 2: (A) Structure of our datasets and experiments. All models were trained with image datasets, while the performance was evaluated primarily on videos. We compared different datasets for pre-training, and applied them to two use cases: videos of redfronted lemurs and Assamese macaques. We compared our bounding box based approach with DeepLabCut (Lauer et al., 2022). Finally, we demonstrated the use of our additional classification branch for individual identification. (B) Overview of the model architecture: Input images are processed by a convolutional backbone into a feature volume. Different heads learn tracking related tasks. We omitted the offset head for simplicity. Afterwards, individuals can be cut out and processed by the classification branch, for example, for individual identification.

We manually annotated the bounding boxes on 500 frames randomly sampled from the macaque videos. The sampling strategy was based on the pre-trained model on MacaquePose, which yielded an approximate count of individuals in each video. Then, we sampled as many frames as individuals detected, to have more examples from crowded scenes and less from individuals who were by themselves in the video.

We annotated 12 sequences of length 12 seconds completely frame-by-frame to evaluate the tracking performance. The sequences have varying levels of difficulty, and we selected them to all contain some motion or occlusion. We used the CVAT labeling software[1] for the frame-by-frame labeling of the evaluation videos. Every monkey in every frame was annotated if recognizable on the still image, even

[1] https://www.cvat.ai

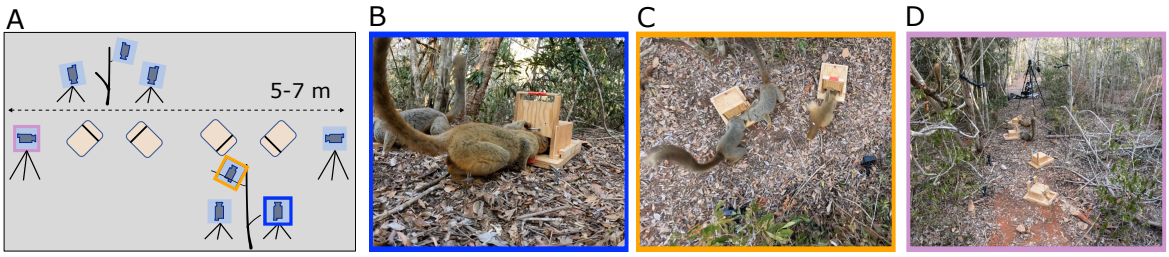

Figure 3: (A) Setup for social learning experiments: Four feeding boxes were placed on the ground and eight cameras were filming from three different perspectives ((B) close, (C) top, (D) far).

without the temporal context. If they were temporally occluded and reappeared, they received the same ID.

For quality control of the ground truth annotations used for evaluation, we assessed the consistency between the labels produced by two human annotators on a subset of three videos of the macaque dataset. We matched the annotated bounding boxes using Hungarian matching and calculated the average intersection over union across all boxes. The average IoU score per video was over 0.8 for the three videos.

Redfronted Lemurs. We conducted social learning experiments in Kirindy Forest, Madagascar, that involved four groups of redfronted lemurs, with group sizes ranging from five to eight individuals. Numerous behavioral studies have been conducted in this area with redfronted lemurs (Schnoell and Fichtel, 2012; Fichtel et al., 2018; Peckre et al., 2022). The data was recorded as part of an ongoing study, in which we used feeding boxes that could be opened with two techniques: lifting a lid or pushing a box to gain access to a food reward. In each lemur group, we presented four feeding boxes in each experiment and repeated experiments ten times over the course of three months. The duration of each experiment varied from 10 to 30 minutes until the lemurs lost interest because the feeding boxes were depleted. We placed eight GoPro Hero 10 cameras to record the interactions of individuals with the feeding boxes from different angles (Fig. 3A). All cameras recorded videos with a resolution of 1920×1080 and 30 frames per second. We turned on Extended Battery Mode for longer recoding times without overheating. Four cameras were placed to record the opening mechanism of each box from a distance of one meter (Fig. 3B). Two cameras mounted on nearby trees were used to capture a top-down view of pairs of boxes (Fig. 3C). Additionally, we placed two cameras on tripods to capture an overview of the experimental set-up from two opposite sites (Fig. 3D). The cameras were voice-controlled to start and stop recording. However, the time it took for each camera to receive and process the voice signal varied and therefore the videos were not completely synchronized. We temporally aligned the videos from the different cameras by cross-correlating the audio signal of each camera with that of one main camera that started recording before all others. This procedure allowed us to reliably determine the temporal offset because an auditory signal recorded by all cameras was used to attract the lemurs to the experiment site (Schnoell and Fichtel, 2012).

We got eight videos from different perspectives for each of the total 40 experimental sessions. Additionally, we did preexperiment sessions with open boxes so the lemurs could get used to the experimental setup. We used a subset of 50 videos from those preexperiment sessions. As the close-up view showed the most variability in lemur appearance, we used 24 videos from this view and 13 frames from the far view and top view, respectively. We manually cut out the largest possible part of each video that continuously contained lemurs, and randomly extracted 10 frames. The 500 resulting frames formed the training set. In these images, we annotated all lemurs by manually drawing the bounding boxes. We decided to exclude the lemur tails from the bounding boxes, as they take a large proportion of the videos and lead to fast changes in bounding box size and location. Additionally, we labeled all feeding boxes in the videos to showcase the multi-class capabilities of our model. When analyzing interactions between lemurs and feeding boxes, it is necessary to detect both classes, however, interaction classification is beyond the scope of this work. For labeling single frames for training, we used the software VoTT[2].

[2]https://github.com/microsoft/VoTT

Additionally, we labeled bounding boxes for all lemurs and feeding boxes in 12 video sequences of length 12 seconds completely frame-by-frame to evaluate the tracking performance of the model. We chose four videos from each camera perspective for evaluation.

Individual Identification. All lemurs wear collars with tags of different shapes and colors to identify individuals. We manually hand selected 350 frames from the close-up videos of one of the groups with six individuals. For this, we watched 40 videos from the different sessions and saved frames in which the collar of at least one individual was identifiable. At maximum, we selected 25 frames from one video to ensure diversity in video background and lighting conditions. Again, we used VoTT to annotate the respective bounding boxes and added the lemur identity. If other lemurs were present in these frames whose collars were not visible, we annotated them with the label "Unsure".

To evaluate the identification performance on temporal tracks, we selected ten close-up camera videos and extracted three one-minute video sequences from each. We manually chose sequences in which the lemur tracking model had no false positives or ID switches.

Ethical approval This study adhered to the Guidelines for the Treatment of Animals in Behavioral Research and Teaching (Buchanan et al., 2012) and the legal requirements of the countries (Madagascar, Thailand) in which the work was carried out. The Malagasy Ministry of the Environment, the Mention Zoologie et Biodiversité, Animale de l'Université d'Antananarivo and the CNFEREF Morondava approved and authorized this research (036/24/MEDD/SG/DGGE/DAPRNE/SCBE.Re). Data collection in Thailand was authorized by the Department of National Parks, Wildlife and Plant Conservation (DNP) and the National Research Council of Thailand (NRCT) (permit numbers 0401/11121 and 0401/13688).

2.2 Model: PriMAT

Due to the existing datasets (Shao et al., 2018; Ionescu et al., 2013) and benchmarks (Milan et al., 2016; Dendorfer et al., 2020; Sun et al., 2022) for multi-object tracking, most models were designed for pedestrian tracking. However, nonhuman primates are very different in appearance, motion patterns, and acceleration. PriMAT is based on the FairMOT (Zhang et al., 2021) architecture, a popular multi-object tracking model initially developed for pedestrian tracking. It belongs to the family of tracking-by-detection approaches. We adapted the tracking strategy to make it more suitable for the appearance and motion of monkeys and lemurs. In particular, we proposed changes to the model to address the challenges of occlusion, low-confidence detections, and abrupt motions such as jumps.

Our model is structured into three parts (Fig. 2B): *(1)* a convolutional backbone used for feature extraction, *(2)* multiple heads that perform different tasks for tracking, and *(3)* a classification branch.

Backbone and heads. The backbone processes the input images $I \in [0, 255]^{W_0 \times H_0 \times 3}$ into feature maps $F \in \mathbb{R}^{W \times H \times C}$, where $W = W_0/4$, $H = H_0/4$ (in our case $W \times H = 272 \times 152$ and $C = 512$). We use HRNet (Sun et al., 2019) as a backbone. Four independent heads process the resulting feature map, each with the same width and height, but with a different number of channels and a different loss, depending on their task. The four heads are:

- Center heatmap (output dimensionality $W \times H \times cl$): This head predicts for each pixel how likely it is to be the center of an object. We adapted the architecture to be a multi-class model by having one center heatmap per class. cl is the number of distinct object classes; in our case $cl = 2$ for lemurs and feeding boxes and $cl = 1$ for the macaque model.

- Bounding box size ($W \times H \times 2$): This head predicts for each pixel the width and height of a hypothetical bounding box if there is a center point at this pixel. The two channels correspond to the width and height of the respective bounding box.

- Bounding box offset ($W \times H \times 2$): The feature map F has a resolution four times smaller than the input image size. Therefore, each pixel on the feature map corresponds to 4×4 pixels in the original image. This head predicts a small offset for each predicted bounding box to correct for the effect of down-sampling.

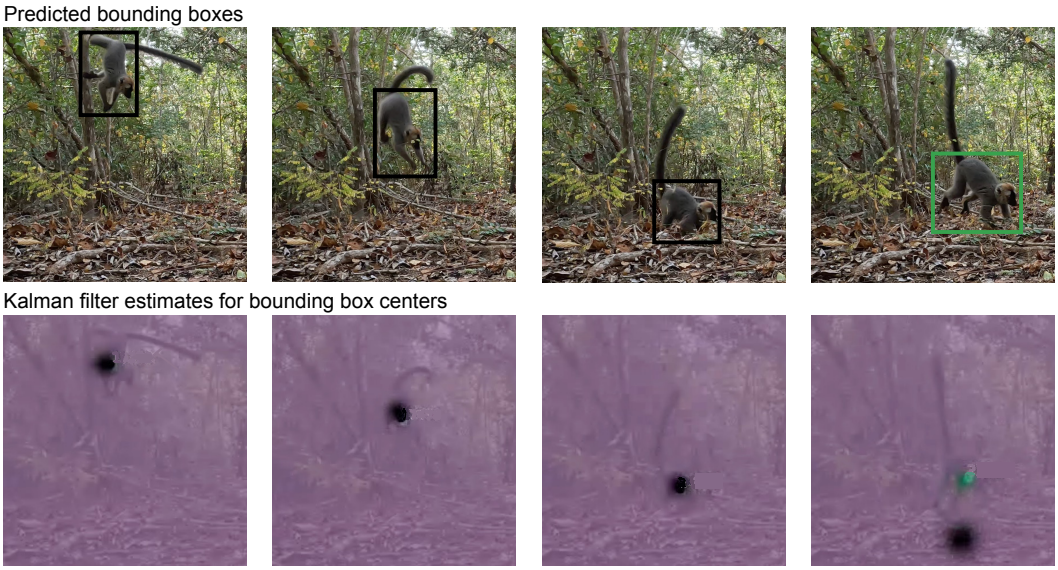

Figure 4: Problem during a jump: The Kalman filter predicts the location of the lemur track based on past velocity. The detection in the last frame cannot be matched to the lemur track. Top row: Detected bounding boxes and assigned tracks identified by the color of the bounding box. The track of the jumping lemur is lost in the last frame (black) and a new track is instantiated (green). Bottom row: Kalman filter predictions for the expected position of the lemur. The Kalman filter assumes linear motion and predicts a different position when the jump ends (black). The newly instantiated track starts with default parameters for the Kalman filter prediction (green).

- Re-identification (ReID) features ($W \times H \times 128$): The ReID head generates a numeric representation as a vector for each detected object. This vector encodes the individual's appearance and is later used as additional information for the association of individuals over time.

For more information about the heads and loss functions, please refer to the detailed description in FairMOT (Zhang et al., 2021).

Association of detections. The general idea for the association is the same as in FairMOT (Zhang et al., 2021): Initially, the model identifies objects from the local maxima in the center heatmap of the first frame, creating individual tracks. In subsequent frames, it generates new detections and ReID features, comparing them to existing tracks for similarity. When they match, the detections are added to the corresponding tracks, while unmatched detections are considered potential candidates for new tracks.

The similarity between existing tracks and new detections is assessed using two criteria, appearance and location. Appearance similarity is calculated via cosine similarity of the ReID features, whereas Intersection-over-Union (IoU) is the measure for similarity in location. Additionally, motion information is modeled using a Kalman filter which predicts the expected location of each bounding box based on velocity in the past frames. The result is a distance matrix between the set of existing tracks (i.e. past detections) and the set of detections in the current frame.

While FairMOT followed a more complicated procedure that involved a sequential application of the two distance measures, we simplified the process as earlier experiments showed that it performed at least equally well. In our model, the final distance matrix is calculated as a linear combination between the IoU distance matrix and the distance matrix obtained from the appearance features. We found a weight of 0.8 for the IoU distance and 0.2 for the appearance feature distance to work well in our applications. The optimal assignment of new detections to existing tracks is obtained using Hungarian matching (Kuhn, 1955). Unmatched tracks are retained for 90 frames (3 seconds in our videos) to facilitate re-detection after occlusion periods.

Monkey tracking. Pedestrians in city scenes are generally easily distinguishable from the background. This makes them easier to track than nonhuman primates in their natural habitat, especially when the primates are in non-standard poses or highly occluded. Therefore, we lowered the confidence threshold for the center heatmap by a factor of 10–40, compared to the original 0.4 used for pedestrian tracking. This adjustment was necessary to maintain track continuity, especially in cases where the animals were moving. As low detection thresholds introduce potential false positive detections, we added a threshold to prevent new boxes from being detected if they have too much overlap with already existing ones. Additionally, we removed detections less than five pixels away from the image border, as most false positives were found in these areas. The appearance of pedestrians is highly variable, and therefore, the original FairMOT model uses only ReID features for the association step. The unmatched tracks and detections are afterwards compared using Intersection-over-Union (IoU). Monkeys have much less appearance variation, which is why for our use cases location is more important than appearance. For this reason, we do the association based on a linear combination of the two distance metrics in one step.

The Kalman filter assumes linear motion of objects and predicts the position of an object based on the velocity in the past frames. During a jump, a lemur moves with high velocity, which then drops close to zero, once the animal lands, which causes problems with the association (Fig. 4). To avoid problems with fast motion changes, we added a potential second stage to the association procedure. If there is no match at the location predicted by the Kalman filter, we additionally check for matches at the location of the last detection.

2.3 Individual identification

To demonstrate how the tracking model can be extended, we added a new branch to the model and trained it to identify individual lemurs. It is important to note that we were not aiming for frame-wise correct identification, which would have been infeasible, as the lemurs were frequently in positions where identification was impossible. Instead, we used the tracks to arrive at a majority vote on which lemur is visible in each sequence of detections. This required the model to be confident in situations where the lemur is clearly identifiable and not be overconfident in situations where the decision could not be made. The final decision for which individual is predicted for a track was made based on the highest number of frames with confidence over 0.3.

We trained the classification branch for individual identification independently of the tracking model. An ImageNet (Deng et al., 2009) pre-trained ResNet18 (He et al., 2016) was used for classification. We fine-tuned the network with the individual lemur images which we cut out from the training images. Instead of taking the exact bounding box, we cut out squared bounding boxes to avoid squeezing rectangular bounding boxes into squared shapes and introducing biases to the training set. We achieved the squared cut-out by increasing the size of the shorter side of the rectangle to match the longer side and resized the cropped individuals to 224×224 pixels. To account for the shift from perfect bounding boxes during training to predicted bounding boxes during inference, we applied random jitter of up to five pixels to the bounding box and random zoom of up to 20% larger to width and height during training.

We trained the first model with 315 hand-selected images of six individuals. The images were extracted from videos in situations where at least one individual was identifiable, either through their face or collar. After training, we used the first model to acquire more training data, by applying it to four additional recordings of experiments and reviewing its predictions per track. We randomly sampled 20 frames per track and went through the selected images manually, only including those in the training set that had no mistakes and were not near duplicates with other selected images. This process resulted in 782 training images. Additionally, we used a small validation set of 35 hand-selected images for selecting the model via early stopping and evaluated the models on 30 snippets of one minute length.

An alternative approach to the classification branch, which can be seen as an independent second model, is to add a new head to the backbone that learns identification. This can be achieved either similarly to the ReID head described above or by cutting out regions of interest from the feature volume and resizing these using the classical ROI-align method (Ren et al., 2015). We compared both model architectures and their performance in lemur identification.

2.4 Evaluation and metrics

Tracking. To evaluate how well our model learned to track, we applied it to the held-out test set which consists of 12 densely annotated video sequences (see Section 2.1 on datasets).

We evaluated the performance of our model with several metrics used for multi-object tracking benchmarks. The main metric is Higher Order Tracking Accuracy (HOTA) (Luiten et al., 2021), as it gives equal importance to detection and association. Additionally, we report Multi-Object Tracking Accuracy (MOTA) (Bernardin and Stiefelhagen, 2008) and IDF1 (Ristani et al., 2016), two traditional metrics for easier comparison to other tracking models.

Compare the detection of bounding box and keypoints. We used a simple metric to compare detection performance between bounding box and keypoint tracking, as we only had bounding box ground truth labeled for complete video sequences. For each ground truth bounding box, we measured if there was a predicted bounding box with an intersection-over-union (IoU) of at least 0.5. Similarly, we counted how often there were at least two keypoints within the bounding box. As the keypoints could be close to the box border, we also counted the keypoints that were within the 1.5× buffered bounding box and reported this as the recall value for each video.

Additionally, we checked for each predicted keypoint whether it was inside a ground truth bounding box, and for each predicted bounding box if there was more than 0.5 IoU with a ground truth box. We saved the success ratio as precision. We reported the F1 score in both cases,

$$F1 = \frac{2 \times precision \times recall}{precision + recall}.$$

Note that this was a very conservative metric from our perspective, as it did not have a minimum requirement for the pose estimation other than that at least two keypoints had to be present.

Detection. We evaluated how well models trained on different datasets were able to generalize to other datasets using mean Average Precision (mAP), a classical metric for object detection tasks (Lin et al., 2014; Shao et al., 2018).

Individual identification. We performed the evaluation for individual identification on 30 videos of length one minute. We evaluated track-wise results and used accuracy as a performance metric. Additionally, we looked at the confusion matrices to check for systematic errors between individuals.

2.5 Training details

The model was implemented in PyTorch (Paszke et al., 2019) and trained with the Adam optimizer (Kingma and Ba, 2014). The training of the base model with MacaquePose was performed on four NVIDIA Quadro RTX 5000 GPUs and took five days for 400 epochs. We started with a learning rate of 5×10^{-5} and reduced it by a factor of 10 after 200 epochs, with a batch size of 8. We used the same settings for fine-tuning with 500 images for both lemurs and macaques, except for an initial learning rate of 10^{-5}. The training took six hours. Early stopping could not be applied as we were training on images, but we were interested in tracking performance on videos.

For the individual identification, we cut out all the lemurs in the dataset and stacked them afterwards, allowing training with a batch size of 128. The model was on a single NVIDIA Quadro RTX 5000 GPU, and we trained it using early stopping with a patience of 10 epochs. Training time with a learning rate of 2×10^{-5} was one hour.

Additionally, we trained two models with DeepLabCut (Lauer et al., 2022) version 2.3.8. For the first model, we annotated seven keypoints (nose, head, neck, spine, start tail, mid tail, end tail) for the lemurs. As the points were mostly annotated along the spine, we called this model DLC-spine. For a second comparison, we annotated 100 frames with 33 keypoints per individual and called the resulting model DLC-full. The time for annotating these frames matches roughly the time needed for annotating 500 bounding boxes. We trained both models for 200,000 training iterations with the data-driven individual assembly method which was recommended for improved performance in the multi-animal setting.

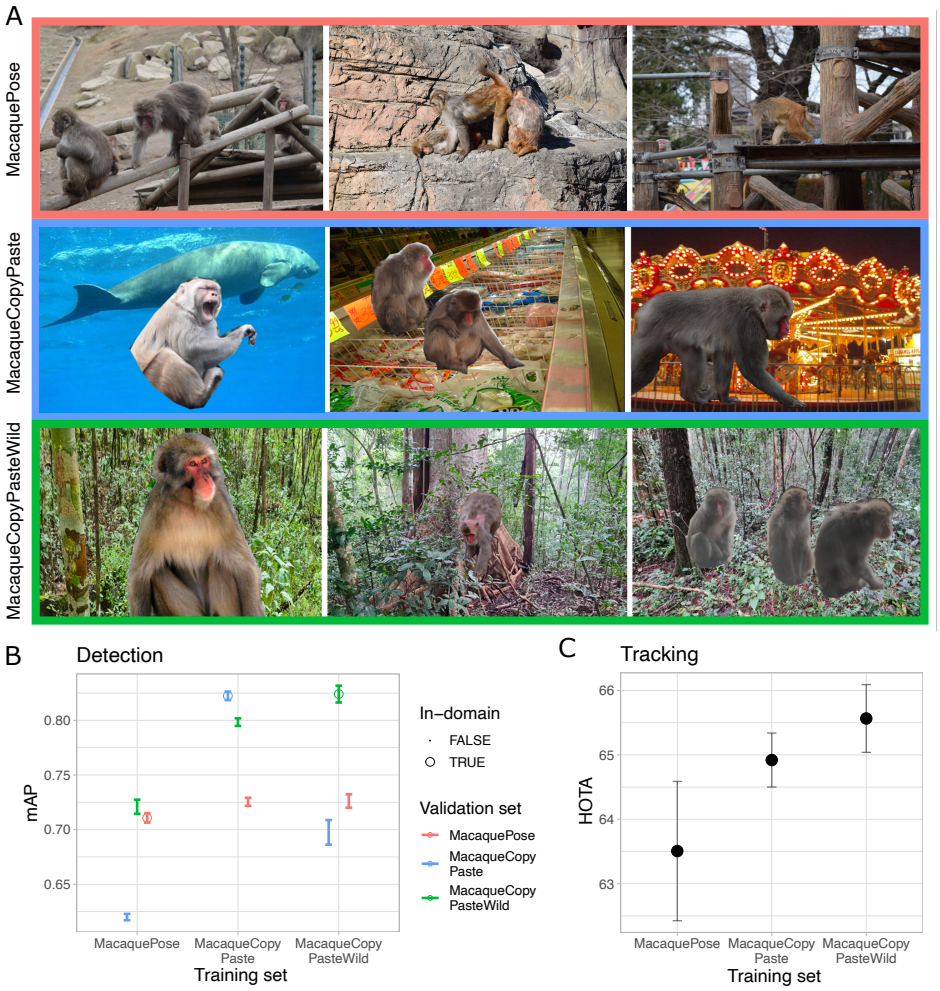

Figure 5: (A) Examples from MacaquePose (Labuguen et al., 2021) (top), MacaqueCopyPaste with ImageNet (Deng et al., 2009) backgrounds (middle), and MacaqueCopyPasteWild with backgrounds from Thailand (bottom). (B) MacaqueCopyPaste detects macaques well on all three datasets. MacaquePose did not generalize well to other datasets. (C) Both Copy-Paste strategies outperform MacaquePose on the validation videos. The error bars show the standard error from training three models with different random seeds.

3 Results

With our multi-animal tracking model PriMAT, we were able to detect and track individual macaques and lemurs in videos from the wild, by only annotating bounding boxes on a few hundred frames. The model was specifically designed for good performance in monkey videos, with solutions for non-standard poses, occlusion, similar appearance, and jumps (see methods). We evaluated the tracking performance of our model with annotated video data from Assamese macaques in Thailand and redfronted lemurs in Madagascar.

Copy-paste data augmentation. To train the base model, we used the images from the MacaquePose (Labuguen et al., 2021) dataset. We experimented with a simple copy-paste strategy for improved generalization to variation in the backgrounds. Previous work has shown that copying and pasting instances onto other backgrounds improved model performance in the area of instance segmentation (Ghiasi et al., 2021). The backgrounds in the MacaquePose images have some variability, however, they are mostly brown or grey (Fig. 5A), so we created MacaqueCopyPaste with ImageNet (Deng et al., 2009) backgrounds and MacaqueCopyPasteWild with backgrounds from the forest in Thailand.

We split the three datasets MacaquePose, MacaqueCopyPaste, and MacaqueCopyPasteWild into

Table 1: Comparison of tracking performance on 12 short video sequences for each, lemurs and macaques. a) MacaqueCopyPaste pre-training resulted in improved results in macaque tracking, but not for lemurs. b) An additional association step to prevent identity switches after jumps or fast motion improved performance for lemurs, but not for macaques, where the validation videos did not contain any jumps. c) A reduced detection threshold showed improved results for both lemurs and macaques.

	Lemurs			**Macaques**		
	HOTA	**MOTA**	**IDF1**	**HOTA**	**MOTA**	**IDF1**
a) pre-training						
ImageNet	**69.8**	**81.1**	**87.3**	68.6	85.8	87.2
MacaqueCopyPaste	62.3	59.9	74.1	**73.9**	**89.7**	**91.6**
no pre-training	61.4	63.4	78.0	64.7	77.9	84.5
b) Association						
Regular	69.2	79.4	86.3	**73.9**	**89.7**	**91.6**
Fast motion	**69.8**	**81.1**	**87.3**	73.2	86.8	91.0
c) Detection threshold						
0.01	**69.8**	81.1	87.3	73.2	86.4	90.8
0.02	69.6	**81.1**	86.7	73.0	87.9	90.3
0.04	69.6	81.1	86.9	**73.9**	89.7	**91.6**
0.1	69.5	80.9	**87.7**	73.1	**90.1**	90.3
0.2	68.5	79.8	86.0	73.1	89.4	90.5
0.4	67.0	78.7	84.4	71.3	87.7	88.3

80% training data and 20% validation data. With the training parts of each of the datasets, we trained three models. We evaluated each model on the validation part of all three datasets. While the models trained with MacaqueCopyPaste performed well on all three datasets, the other models only performed well in their domains (Fig. 5B). The tracking validation videos show similar backgrounds to the one MacaqueCopyPasteWild was trained on, which explains the better tracking performance. Both copy-paste approaches outperformed the models trained on MacaquePose (Fig. 5C).

We further investigated using the base model pre-trained on MacaqueCopyPaste to transfer to the lemur videos. We fine-tuned the base model with the 500 annotated images of lemurs. pre-training with ImageNet was more beneficial than MacaqueCopyPaste for the lemurs (Table 1a). The domain shift from the images of MacaquePose to the lemur videos is large, not only because the species look different, but the lemur experiments were filmed from three different camera angles.

Monkey tracking. The motion patterns and appearance of monkeys and lemurs differ from those of humans. Therefore, PriMAT contains elements that are specifically tailored to videos of primates in the wild. We described these in Section 2.2 and evaluated how they help to improve performance.

First, we implemented an alternative association logic, since the Kalman filter predicts the position of an object based on the velocity in the past assuming linear motion. However, it struggles to accurately track the movement of a monkey during a jump or other fast motion due to the abrupt change in velocity when the monkey lands. To circumvent this problem, we changed the association logic to check for the last visible position of an object in cases where it could not be matched to the current Kalman filter prediction. This change had a positive effect on the performance for the lemur videos (Table 1b). The validation snippets only last 12 seconds each, and there are only a few jumps in the lemur videos recorded. We expect our association approach to show even better results on longer videos with more jumps.

Second, we adapted the detection logic as the primates were often not easily detectable, especially when they were heavily occluded or in non-standard poses. Performance increased when choosing a lower threshold for detections (Table 1c). The original FairMOT model (Zhang et al., 2021) used a threshold of 0.4 for pedestrians, which we recommend to lower by a factor of 10–40 for nonhuman primates.

Bounding box and keypoint based detection. In addition to the studies in the introduction which support the hypothesis that models for keypoint estimation in the wild struggle with gener-

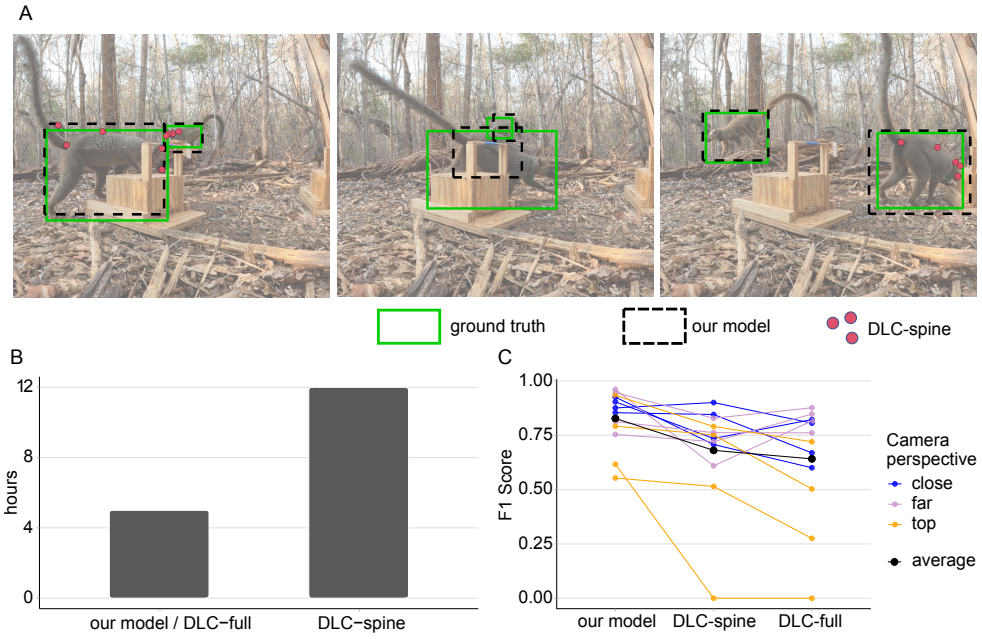

Figure 6: Comparison between our boxes based tracking model and two DeepLabCut models (Lauer et al., 2022). Our model and DLC-spine were both trained with 500 images from three camera perspectives. DLC-spine used seven keypoints along the animal spine. DLC-full was trained using 33 keypoints, and for the same time that was used to label the boxes for our model (5 hours), resulting in 100 frames. (A) A qualitative example where DLC-spine does not detect most keypoints. (B) Annotation time was higher for keypoints. (C) Performance difference on the 12 validation videos. The evaluation metric was chosen in a favorable manner for DLC, as it counts a detection as positive when there are at least two keypoints correctly detected. Missing detections in the skeleton, multiple identities or mixed up skeletons occur in the videos, however, these are not reflected in this metric as such mistakes could not be observed using bounding boxes.

alization, we performed a small experiment. To compare our model with keypoint based tracking, we trained two models with DeepLabCut (Lauer et al., 2022). For the first model (DLC-spine), we annotated seven keypoints (nose, head, neck, spine, start tail, mid tail, end tail) for each individual and the same 500 frames used for our lemur model. The first difference we found was the annotation time. While our annotations took 3–5 hours for 500 frames, depending on the number of individuals, the keypoint annotations for this set of images took around 12 hours (Fig. 6B). For the second model (DLC-full), we annotated 33 keypoints for each individual in 100 frames, which is approximately the time we needed to draw bounding boxes on 500 frames. Even when only looking at a very basic detection metric where two keypoints were needed to count as a successful detection, our model outperformed both keypoint based method (Fig. 6C). In one video, which contained high-contrast shadows and therefore a higher need for generalization, the keypoint based methods were not able to detect any keypoints in the whole sequence. Our bounding box method dropped its performance on this challenging video but still had a final F1 score of 0.62.

Note that the metric we used (see Section 2.4) only evaluates detection. We aimed for a fair evaluation procedure that would not favor any of the approaches. However, qualitative results showed that the keypoint based approach additionally often missed keypoints, had a skeleton distributed across two individuals, or two skeletons within one individual, and showed ID switches, none of which is reflected in this metric. This means that the performance difference for tracking is potentially even larger than the detection results suggest. Apart from these observations, having incomplete skeletons, many missing detections or incomplete tracks make the difference between being usable or not in practical applications. To overcome these types of problem and have a robust and generalizing model, a lot more data would be needed.

Table 2: Accuracy of models trained for individual identification. Having a shared backbone for tracking and individual identification showed poor results with the amount of data used (500 images for tracking, 782 images for individual identification). Feeding the detections into a second model for identification led to improved results.

Architecture	Accuracy
Two separate models	0.83
Shared backbone	0.42

Individual identification. We added a classification branch described in Section 2.2 and applied it to identify individual lemurs in the videos. We used one group of lemurs with six individuals and evaluated the performance on 30 snippets of length one minute, with 96 tracks which had to be classified. The final model had an identification accuracy of 83%. Additionally, we experimented with a different architecture. Instead of having two separate models for tracking and identification, we tested the possibility of adding a new head to the existing backbone. Table 2 shows the need to separate the two tasks, at least for the number of labels we used for training, as the model with shared backbone only reached an accuracy of 42%.

4 Discussion

There are many potential paths towards automated behavioral analysis of videos of animals in the wild. Most current methods in the lab use keypoint detection and tracking as a starting point, as those provide a valuable intermediate representation, namely the pose of the animal. However, so far, keypoint detection based models have struggled to generalize to highly diverse backgrounds in the wild. One solution could be to annotate vastly larger and diverse datasets; however, keypoint annotation is a very time-consuming task.

Therefore, we proposed an alternative approach based on bounding boxes. We demonstrated that our deep learning based multi-animal tracking model, PriMAT, outperforms previous methods based on keypoints on videos from the wild while requiring significantly less labeling time.

Our tracking model is based on the FairMOT architecture (Zhang et al., 2021), which was built to track pedestrians. However, pedestrians move almost exclusively in linear motion and wear distinctive clothing, which influenced the design choices for the model. We addressed primate-specific challenges in the design of our model, including non-standard poses, occlusion, and fast, non-linear motion.

We worked with pre-trained base models that can be used as a starting point to learn to detect and track nonhuman primate species in different settings in the wild using only a few hundred labeled frames. Transfer experiments from macaques to lemurs suggested that ImageNet pre-training led to better results than pre-training on a large macaque dataset. When tracking macaques in the forest however, datasets derivated from MacaquePose yielded better results. We tested strategies of copying and pasting macaques to different backgrounds, and our results showed that when doing transfer learning with a larger domain shift (e.g., very different backgrounds), or a high background variability, the model trained with macaques on ImageNet backgrounds was the best starting point. However, when there are backgrounds from the evaluation domain available, such as with the models trained on MacaqueCopyPasteWild and the videos from Thailand, performance can be improved even further. Our lemur tracking model can track lemurs filmed from different camera perspectives and distances to the camera. As a step towards interaction classification, we adapted the model to track other objects of interest, such as feeding boxes.

Our flexible model can learn additional classification tasks on top of the tracking results, such as individual identification. This can be used in cases where all individuals are known. With a total of 782 images of six individuals, we reached an accuracy of 83% on unseen video tracks. For this task, it was necessary to add a separate branch. Our experiments to learn tracking and identification from the same backbone with an additional head were less successful. This could be explained by the comparably low sample sizes and competing tasks. When focusing on tracking, the model will probably discard unnecessary information, such as the collars, which would be helpful for individual identification. Similarly, tracking performance decreased when giving more weight to the identification head. This approach might work in applications with more annotated data, and the benefit of having only one backbone and different lightweight heads for downstream tasks could be achievable in this

context.

In the future, the flexibility of the model allows the extension of PriMAT to incorporate other tasks such as instance segmentation (Wang et al., 2020) and ultimately interaction classification (Liu et al., 2021) between objects when those interactions are detectable in single frames.

References

Bain, M., Nagrani, A., Schofield, D., and Zisserman, A. (2019). Count, crop and recognise: Fine-grained recognition in the wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*.

Bala, P. C., Eisenreich, B. R., Yoo, S. B. M., Hayden, B. Y., Park, H. S., and Zimmermann, J. (2020). Automated markerless pose estimation in freely moving macaques with openmonkeystudio. *Nature communications*, 11(1):4560.

Bernardin, K. and Stiefelhagen, R. (2008). Evaluating multiple object tracking performance: the clear mot metrics. *EURASIP Journal on Image and Video Processing*, 2008:1–10.

Brookes, O. and Burghardt, T. (2020). A dataset and application for facial recognition of individual gorillas in zoo environments. *arXiv preprint arXiv:2012.04689*.

Buchanan, K., Burt de Perera, T., Carere, C., Carter, T., Hailey, A., Hubrecht, R., Jennings, D., Metcalfe, N., Pitcher, T., Peron, F., et al. (2012). Guidelines for the treatment of animals in behavioural research and teaching.

Dendorfer, P., Rezatofighi, H., Milan, A., Shi, J., Cremers, D., Reid, I., Roth, S., Schindler, K., and Leal-Taixé, L. (2020). Mot20: A benchmark for multi object tracking in crowded scenes. *arXiv preprint arXiv:2003.09003*.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee.

Desai, N., Bala, P., Richardson, R., Raper, J., Zimmermann, J., and Hayden, B. (2023). Openapepose, a database of annotated ape photographs for pose estimation. *Elife*, 12:RP86873.

Fichtel, C., Schnoell, A. V., and Kappeler, P. M. (2018). Measuring social tolerance: an experimental approach in two lemurid primates. *Ethology*, 124(1):65–73.

Ghiasi, G., Cui, Y., Srinivas, A., Qian, R., Lin, T.-Y., Cubuk, E. D., Le, Q. V., and Zoph, B. (2021). Simple copy-paste is a strong data augmentation method for instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2918–2928.

Graving, J. M., Chae, D., Naik, H., Li, L., Koger, B., Costelloe, B. R., and Couzin, I. D. (2019). Deepposekit, a software toolkit for fast and robust animal pose estimation using deep learning. *Elife*, 8:e47994.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Ionescu, C., Papava, D., Olaru, V., and Sminchisescu, C. (2013). Human3. 6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE transactions on pattern analysis and machine intelligence*, 36(7):1325–1339.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Koger, B., Deshpande, A., Kerby, J. T., Graving, J. M., Costelloe, B. R., and Couzin, I. D. (2023). Quantifying the movement, behaviour and environmental context of group-living animals using drones and computer vision. *Journal of Animal Ecology*.

Kuhn, H. W. (1955). The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97.

Labuguen, R., Matsumoto, J., Negrete, S. B., Nishimaru, H., Nishijo, H., Takada, M., Go, Y., Inoue, K.-i., and Shibata, T. (2021). Macaquepose: a novel "in the wild" macaque monkey pose dataset for markerless motion capture. *Frontiers in behavioral neuroscience*, 14:581154.

Lauer, J., Zhou, M., Ye, S., Menegas, W., Schneider, S., Nath, T., Rahman, M. M., Di Santo, V., Soberanes, D., Feng, G., et al. (2022). Multi-animal pose estimation, identification and tracking with deeplabcut. *Nature Methods*, 19(4):496–504.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. (2014). Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer.

Liu, H., Yan, N., Mortazavi, M., and Bhanu, B. (2021). Fully convolutional scene graph generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11546–11556.

Luiten, J., Osep, A., Dendorfer, P., Torr, P., Geiger, A., Leal-Taixé, L., and Leibe, B. (2021). Hota: A higher order metric for evaluating multi-object tracking. *International journal of computer vision*, 129:548–578.

Ma, X., Kaufhold, S., Su, J., Zhu, W., Terwilliger, J., Meza, A., Zhu, Y., Rossano, F., and Wang, Y. (2023). Chimpact: A longitudinal dataset for understanding chimpanzee behaviors. *Advances in Neural Information Processing Systems*, 36:27501–27531.

Marks, M., Jin, Q., Sturman, O., von Ziegler, L., Kollmorgen, S., von der Behrens, W., Mante, V., Bohacek, J., and Yanik, M. F. (2022). Deep-learning-based identification, tracking, pose estimation and behaviour classification of interacting primates and mice in complex environments. *Nature machine intelligence*, 4(4):331–340.

Mathis, A., Biasi, T., Schneider, S., Yuksekgonul, M., Rogers, B., Bethge, M., and Mathis, M. W. (2021). Pretraining boosts out-of-domain robustness for pose estimation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1859–1868.

Mathis, A., Mamidanna, P., Cury, K. M., Abe, T., Murthy, V. N., Mathis, M. W., and Bethge, M. (2018). Deeplabcut: markerless pose estimation of user-defined body parts with deep learning. *Nature neuroscience*, 21(9):1281–1289.

Milan, A., Leal-Taixé, L., Reid, I., Roth, S., and Schindler, K. (2016). Mot16: A benchmark for multi-object tracking. *arXiv preprint arXiv:1603.00831*.

Ng, X. L., Ong, K. E., Zheng, Q., Ni, Y., Yeo, S. Y., and Liu, J. (2022). Animal kingdom: A large and diverse dataset for animal behavior understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19023–19034.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al. (2019). Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

Paulet, J., Molina, A., Beltzung, B., Suzumura, T., Yamamoto, S., and Sueur, C. (2024). Deep learning for automatic facial detection and recognition in japanese macaques: illuminating social networks. *Primates*, pages 1–15.

Peckre, L. R., Michiels, A., Socias-Martinez, L., Kappeler, P. M., and Fichtel, C. (2022). Sex differences in audience effects on anogenital scent marking in the red-fronted lemur. *Scientific Reports*, 12(1):5266.

Pereira, T. D., Tabris, N., Matsliah, A., Turner, D. M., Li, J., Ravindranath, S., Papadoyannis, E. S., Normand, E., Deutsch, D. S., Wang, Z. Y., et al. (2022). Sleap: A deep learning system for multi-animal pose tracking. *Nature methods*, 19(4):486–495.

Perez, M. and Toler-Franklin, C. (2023). Cnn-based action recognition and pose estimation for classifying animal behavior from videos: A survey. *arXiv preprint arXiv:2301.06187*.

Pineda, R. R., Kubo, T., Shimada, M., and Ikeda, K. (2023). Deep mantra: deep learning-based multi-animal tracking for japanese macaques. *Artificial Life and Robotics*, 28(1):127–138.

Ren, S., He, K., Girshick, R., and Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28.

Ristani, E., Solera, F., Zou, R., Cucchiara, R., and Tomasi, C. (2016). Performance measures and a data set for multi-target, multi-camera tracking. In *European conference on computer vision*, pages 17–35. Springer International Publishing Cham.

Roy, A. M., Bhaduri, J., Kumar, T., and Raj, K. (2023). Wildect-yolo: An efficient and robust computer vision-based accurate object localization model for automated endangered wildlife detection. *Ecological Informatics*, 75:101919.

Schnoell, A. V. and Fichtel, C. (2012). Wild redfronted lemurs (eulemur rufifrons) use social information to learn new foraging techniques. *Animal Cognition*, 15:505–516.

Schofield, D., Nagrani, A., Zisserman, A., Hayashi, M., Matsuzawa, T., Biro, D., and Carvalho, S. (2019). Chimpanzee face recognition from videos in the wild using deep learning. *Science advances*, 5(9):eaaw0736.

Schofield, D. P., Albery, G. F., Firth, J. A., Mielke, A., Hayashi, M., Matsuzawa, T., Biro, D., and Carvalho, S. (2023). Automated face recognition using deep neural networks produces robust primate social networks and sociality measures. *Methods in Ecology and Evolution*, 14(8):1937–1951.

Schülke, O., Pesek, D., Whitman, B. J., and Ostner, J. (2011). Ecology of assamese macaques (macaca assamensis) at phu khieo wildlife sanctuary, thailand. *Journal of Wildlife in Thailand*, 18(1):1–15.

Shao, S., Zhao, Z., Li, B., Xiao, T., Yu, G., Zhang, X., and Sun, J. (2018). Crowdhuman: A benchmark for detecting human in a crowd. *arXiv preprint arXiv:1805.00123*.

Sun, K., Xiao, B., Liu, D., and Wang, J. (2019). Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5693–5703.

Sun, P., Cao, J., Jiang, Y., Yuan, Z., Bai, S., Kitani, K., and Luo, P. (2022). Dancetrack: Multi-object tracking in uniform appearance and diverse motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20993–21002.

Vogg, R., Lüddecke, T., Henrich, J., Dey, S., Nuske, M., Hassler, V., Murphy, D., Fischer, J., Ostner, J., Schülke, O., et al. (2024). Computer vision for primate behavior analysis in the wild. *arXiv preprint arXiv:2401.16424*.

Wang, Y., Xu, Z., Shen, H., Cheng, B., and Yang, L. (2020). Centermask: single shot instance segmentation with point representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9313–9321.

Wiltshire, C., Lewis-Cheetham, J., Komedová, V., Matsuzawa, T., Graham, K. E., and Hobaiter, C. (2023). Deepwild: Application of the pose estimation tool deeplabcut for behaviour tracking in wild chimpanzees and bonobos. *Journal of Animal Ecology*.

Yang, Y., Yang, J., Xu, Y., Zhang, J., Lan, L., and Tao, D. (2022). Apt-36k: A large-scale benchmark for animal pose estimation and tracking. *Advances in Neural Information Processing Systems*, 35:17301–17313.

Yu, H., Xu, Y., Zhang, J., Zhao, W., Guan, Z., and Tao, D. (2021). Ap-10k: A benchmark for animal pose estimation in the wild. *arXiv preprint arXiv:2108.12617*.

Zhang, Y., Wang, C., Wang, X., Zeng, W., and Liu, W. (2021). Fairmot: On the fairness of detection and re-identification in multiple object tracking. *International Journal of Computer Vision*, 129:3069–3087.