# PanAf20K: A Large Video Dataset for Wild Ape Detection and Behaviour Recognition

Otto Brookes[1] · Majid Mirmehdi[1] · Colleen Stephens[2] · Samuel Angedakin[2] · Katherine Corogenes[2] · Dervla Dowd[3] · Paula Dieguez[2] · Thurston C. Hicks[6] · Sorrel Jones[2] · Kevin Lee[2] · Vera Leinert[2,3] · Juan Lapuente[2] · Maureen S. McCarthy[2] · Amelia Meier[2] · Mizuki Murai[2] · Emmanuelle Normand[3] · Virginie Vergnes[3] · Erin G. Wessling[4] · Roman M. Wittig[2,7] · Kevin Langergraber[8] · Nuria Maldonado[2] · Xinyu Yang[1] · Klaus Zuberbühler[5] · Christophe Boesch[2,3] · Mimi Arandjelovic[2] · Hjalmar Kühl[2,9] · Tilo Burghardt[1]

## Abstract

We present the PanAf20K dataset, the largest and most diverse open-access annotated video dataset of great apes in their natural environment. It comprises more than 7 million frames across ∼ 20,000 camera trap videos of chimpanzees and gorillas collected at 18 field sites in tropical Africa as part of the Pan African Programme: The Cultured Chimpanzee. The footage is accompanied by a rich set of annotations and benchmarks making it suitable for training and testing a variety of challenging and ecologically important computer vision tasks including ape detection and behaviour recognition. Furthering AI analysis of camera trap information is critical given the International Union for Conservation of Nature now lists all species in the great ape family as either Endangered or Critically Endangered. We hope the dataset can form a solid basis for engagement of the AI community to improve performance, efficiency, and result interpretation in order to support assessments of great ape presence, abundance, distribution, and behaviour and thereby aid conservation efforts. The dataset and code are available from the project website: PanAf20K

✉ Otto Brookes
  otto.brookes@bristol.ac.uk

  Majid Mirmehdi
  m.mirmehdi@bristol.ac.uk

  Hjalmar Kühl
  hjalmar.kuehl@senckenberg.de

  Tilo Burghardt
  tilo@cs.bris.ac.uk

[1] Department of Computer Science, University of Bristol, Bristol, UK

[2] Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany

[3] Wild Chimpanzee Foundation, Leipzig, Germany

[4] Department of Human Evolutionary Biology, Harvard University, Cambridge, MA, USA

[5] School of Psychology & Neuroscience, University of St Andrews, St. Andrews, Scotland

[6] Faculty of Artes Liberales, University of Warsaw, Warsaw, Poland

[7] Institute for Cognitive Sciences Marc Jeannerod, University of Lyon, Lyon, France

[8] Institute of Human Origins, Arizona State University, Tempe, AZ, USA

[9] Senckenberg Museum of Natural History Goerlitz, Goerlitz, Germany

## 1 Introduction

*Motivation* As the biodiversity crisis intensifies, the survival of many endangered species grows increasingly precarious, evidenced by species diversity continuing to fall at an unprecedented rate (Ceballos et al., 2020; Vié et al., 2009). The great ape family, whose survival is threatened by habitat degradation and fragmentation, climate change, hunting and disease, is a prime example (Carvalho et al., 2021). The International Union for Conservation of Nature (IUCN) considers all three member species, that is orangutans, gorillas,

chimpanzees (including bonobos), to be either endangered or critically endangered.

The threat to great apes has far-reaching ecological implications. Great apes contribute to the balance of healthy ecosystems by seed dispersal, consumption of leaves and bark, and shaping habitats by creating canopy gaps and trails (Chappell & Thorpe, 2022; Haurez et al., 2015; Tarszisz et al., 2018). They also form part of complex forest food webs, their removal from which would have cascading consequences for local food chains. In addition, great apes are our closest evolutionary relatives and a key target for anthropological research. We share 97% of our DNA with the phylogenetically most distant orangutans and 98.8% with the closer chimpanzees and bonobos. The study of great apes, including their physiology, genetics, and behaviour, is essential to addressing questions of human nature and evolution (Pollen et al., 2023). Urgent conservation action for the protection and preservation of these emblematic species is therefore essential.

The timely and efficient assessment of great ape presence, abundance, distribution, and behaviour is becoming increasingly important in evaluating the effectiveness of conservation policies and intervention measures. The potential of exploiting camera trap imagery for conservation or biological modelling is well recognised (Kühl & Burghardt, 2013; Tuia et al., 2022). However, even small camera networks generate large volumes of data (Fegraus et al., 2011) and the number and complexity of downstream processing tasks required to perform ecological analysis is extensive. Typically, ecologists first need to identify those videos that contain footage of the target study species followed by further downstream analyses, such as estimating the distance of the animals from the camera (i.e., camera trap distance sampling) to calculate species density or identification of ecologically or anthropologically important behaviours, such as tool use or camera reactivity (Houa et al., 2022). Performing these tasks manually is time consuming and limited by the availability of human resources and expertise, becoming infeasible at large scale. This underlines the need for rapid, scalable, and efficient deep learning methods for automating the detection and assessment of great ape populations and analysis of their behaviours.

To facilitate the development of methods for automating the interpretation of camera trap data, large-scale, open-access video datasets must be available to the relevant scientific communities, whilst removing geographic details that could potentially threaten the safety of animals (Tuia et al., 2022). Unlike the field of human action recognition and behaviour understanding, where several large, widely acknowledged datasets exist for benchmarking (Kay et al., 2017; Kuehne et al., 2011; Soomro et al., 2012), the number of great ape datasets is limited and those that are currently available lack scale, diversity and rich annotations.

*Contribution* In this study, we present the PanAf20K dataset, the largest and most diverse open-access video dataset of great apes in the wild—ready for AI training. The dataset comprises footage collected from 18 study sites across 15 African countries, featuring apes in over 20 distinct habitats (i.e., forests, savannahs, and marshes). It displays great apes in over 100 individual locations (e.g., trails, termite mounds, and water sources) displaying an extensive range of 18 behaviour categories. A visual overview of the dataset is presented in Fig. 1. The footage is accompanied by a rich set of annotations suitable for a range of ecologically important tasks such as detection, action localisation, fine-grained and multi-label behaviour recognition.

*Paper Organisation*. Following this introduction, Sect. 2 reviews existing animal behaviour datasets and methodologies for great ape detection and behaviour recognition. Section 3 describes both parts of the dataset, the PanAf20K and the PanAf500, and details how the data was collected and annotated. Benchmark results for several computer vision tasks are presented in Sect. 4. Section 5 discusses the main findings as well as any limitations alongside future research directions while Sect. 6 summarises the dataset and highlights its potential applications.

## 2 Related Work

*Great Ape Video Datasets for AI Development* While there have been encouraging trends in the creation of new animal datasets (Beery et al., 2021; Cui et al., 2018; Swanson et al., 2015; Van Horn et al., 2018), there is still only a limited number specifically designed for great apes and even fewer suitable for behavioural analysis. In this section, the most relevant datasets are described.

Bain et al. (2021), curated a large camera trap video dataset (> 40 h) with fine-grained annotations for two behaviours; buttress drumming and nut cracking. However, the data and corresponding annotations are not yet publicly available and the range of annotations is limited to two audio-visually distinct behaviours. The Animal Kingdom dataset (Ng et al., 2022), created for advancing behavioural understanding, comprises footage sourced from YouTube (50 h, 30 K videos) along with annotations that cover a wide range of actions, from eating to fighting. The MammalNet dataset (Chen et al., 2023), which is larger and more diverse, is also composed from YouTube footage (18 K videos, 539 h) and focuses on behavioural understanding across species. It comprises taxonomy-guided annotations for 12 common behaviours, identified through previous animal behaviour studies, for 173 mammal categories. While both datasets are valuable resources for the study of animal behaviour, they contain relatively few great ape videos since these species make up only a small proportion of the overall dataset. Animal
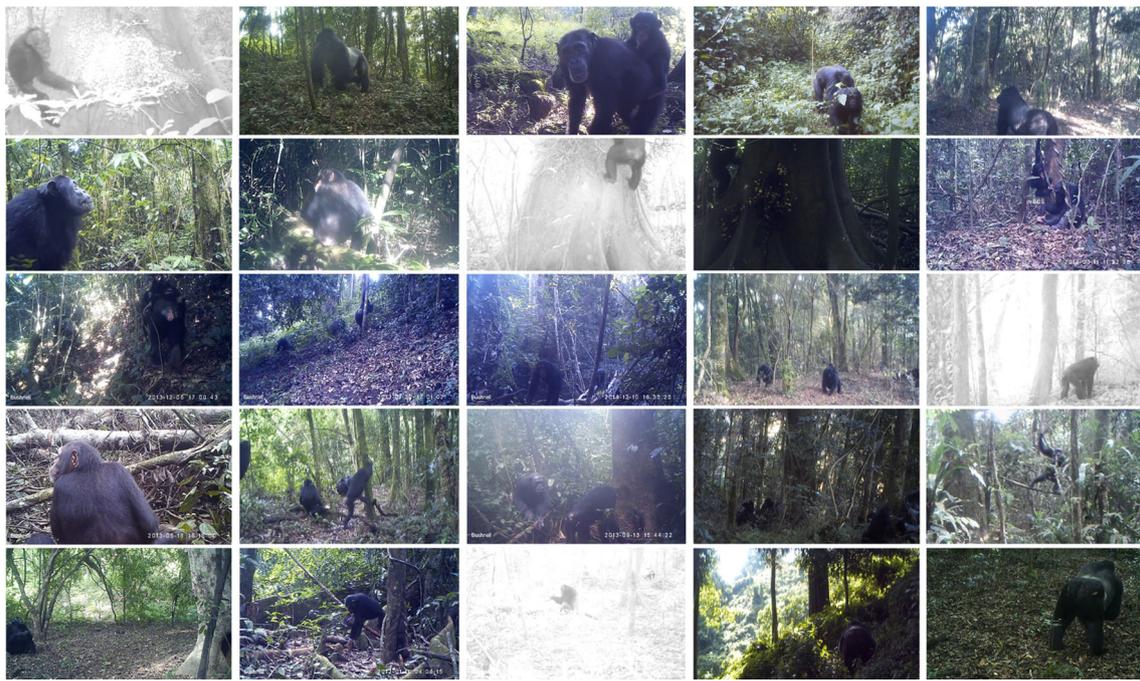
**Fig. 1** PanAf20K visual overview. We present the largest and most diverse open-access video dataset of great apes in the wild. It comprises ∼20,000 videos and more than 7 million frames extracted from camera traps at 18 study sites spanning 15 African countries. Shown are 25 representative still frames from the dataset highlighting its diversity with respect to many important aspects such as behavioural activities, species, number of apes, habitat, day/night recordings, scene lighting, and more

Kingdom spans ∼100 videos while MammalNet includes ∼1000 videos across the whole great ape family, representing ∼0.5% and ∼5% of all videos, respectively. Other work to curate great ape datasets has focused annotation efforts on age, sex, facial location, and individual identification (Brookes & Burghardt, 2020; Freytag et al., 2016; Schofield et al., 2019), rather than behaviour.

For the study of great ape behaviour, the currently available datasets have many limitations. First, they are too small to capture the full breadth of behavioural diversity. This is particularly relevant for great apes, which are a deeply complex species, displaying a range of individual, paired and group behaviours, that are still not well understood (Samuni et al., 2021; Tennie et al., 2016). Secondly, they are not composed of footage captured by sensors commonly used in ecological studies, such as camera traps and drones. This means that apes are not observed in their natural environment and the distribution of behaviours will not be representative of the wild (i.e., biased towards 'interesting' or 'entertaining' behaviours). Additionally, the footage may be biased towards captive or human-habituated animals which display altered or unnatural behaviours and are unsuitable for studying their wild counterparts (Chappell & Thorpe, 2022; Clark, 2011). All these factors may limit the ability of trained models to generalise effectively to wild footage of great apes where conservation efforts are most urgently needed. This,

in turn, limits their practical and *immediate* utility. We aim to overcome these limitations by introducing a large scale, open-access video dataset that enables researchers to develop models for analysing the behaviour of great apes in the wild and evaluate them against established methods.

*Great Ape Detection and Individual Recognition* Yang et al. (2019) developed a multi-frame system capable of accurately detecting the full body location of apes in challenging camera-trap footage. In more recent work, Yang et al. developed a curriculum learning approach that enables the utilisation of large volumes of unlabelled data to improve detection performance (Yang et al., 2023). Several other works focus on facial detection and individual identification. In early research, Freytag et al. (2016) applied YOLOv2 (Redmon & Farhadi, 2017), to localise the faces of chimpanzees. They utilised a second deep CNN for feature extraction (AlexNet (Krizhevsky et al., 2012) and VGGFaces (Parkhi et al., 2015)), and a linear support vector machine for identification. Later, Brust et al. (2017) extended their work utilising a much larger and diverse dataset. Schofield et al. (2019) presented a pipeline for identification of 23 chimpanzees across a video archive spanning 14 years. Similar to Brust et al. (2017), they trained the single-shot object detector, SSD (Schofield et al., 2019), to perform initial localisation, and a secondary CNN model to perform individual classification. Brookes and Burghardt (2020) employed

YOLOv3 (Redmon & Farhadi, 2018) to perform *one-step* simultaneous facial detection and individual identification on captive gorillas.

*Great Ape Action and Behaviour Recognition* To date, three systems have attempted automated great ape behavioural action recognition. The first (Sakib & Burghardt, 2020) was based on the two-stream convolutional architecture by Simonyan and Zisserman (2014) and uses 3D ResNet-18 s for feature extraction and LSTM-based fusion of RGB and optical flow features. They reported a strong top-1 accuracy of 73% across the nine behavioural actions alongside a relatively low average per class accuracy of 42%. The second, proposed by Bain et al. (2021), utilises both audio and video inputs to detect two specific behaviours; buttress drumming and nut cracking. Their system utilises a 3D ResNet-18 and a 2D ResNet-18 for extraction of visual and audio features, respectively, in different streams. They achieved an average precision of 87% for buttress drumming and 85% for nut cracking on their unpublished dataset. Lastly, Brookes et al. (2023) introduced a triple-stream model that utilises RGB, optical flow and DensePose within a metric learning framework, and achieved top-1 and average per-class accuracy of 85% and 65%, respectively.

## 3 Dataset Overview

*Task-Focused Data Preparation* The PanAf20K dataset consists of two distinct parts. The first includes a large video dataset containing 19,973 videos annotated with multi-label behavioural labels. The second part comprises 500 videos with fine-grained annotations across $\sim$ 180,000 frames. Videos are recorded at 24 FPS and resolutions of $720 \times 404$ for 15 s ($\sim$ 360 frames). In this section, we provide an overview of the dataset, including how the video data was originally collected (see Sect. 3.1) and annotated for both parts (see Sect. 3.2).

### 3.1 Data Acquisition

*Camera Trapping in the Wild* The PanAf Programme: The Cultured Chimpanzee has 39 research sites and data collection has been ongoing since January 2010. The data included in this paper samples 18 of these sites and the available data were obtained from studies of varying duration (7–22 months). Grids comprising 20 to 96 $1 \times 1$ km cells were established for the distribution of sampling units (to cover a minimum of 20–50 km$^2$ in rainforest and 50–100 km$^2$ in woodland savannah). An average of 29 (range 5–41) movement-triggered Bushnell cameras were installed per site. One camera was installed per grid cell where possible. However, in larger grids cameras were placed in alternate cells. If certain grid cells did not contain suitable habitat,

such as grassland in forest-savanna mosaic sites, two cameras were placed instead as far away from each other as possible, in cells containing suitable habitat to maximize coverage. In areas where activities of interest (e.g., termite fishing sites) were likely to take place, a second camera was installed to capture the same scene from a different angle. Cameras were placed approx. 1 m high above ground, in locations that were frequently used by apes (e.g., trail, fruit trees). This method ensured a strategic installation of cameras, with maximal chance of capturing footage of terrestrial activity of apes. Both GPS location and habitat type for each location was noted. Footage was recorded for 60 s with a 1 s interval between triggers and cameras were visited every 1–3 months for maintenance and to download the recorded footage throughout the study periods.

### 3.2 Data Annotation

*Fine-grained Annotation of PanAf500* The PanAf500 was ground-truth labelled by users on the community science platform Chimp&See (Arandjelovic et al., 2016) and researchers at the University of Bristol (Sakib & Burghardt, 2020; Yang et al., 2019) (examples are shown in Fig. 2). We re-formatted the metadata from these sources specifically for use in computer vision under reproducible and comparable benchmarks ready for AI-use. The dataset includes frame-by-frame annotations for full-body location, intra-video individual identification, and nine behavioural actions (Sakib & Burghardt, 2020) across 500 videos and $\sim$ 180,000 frames.

As shown in Fig. 3, the number of individual apes varies significantly, from one to nine, with up to eight individuals appearing together simultaneously. Individuals and pairs occur the most frequently while groups occur less frequently, particularly those exceeding four apes. Bounding boxes are categorised according to the COCO dataset (Lin et al., 2014) (i.e., $> 96^2$, $96^2$ and $32^2$ for large, medium and small, respectively) with small bounding boxes occurring relatively infrequently compared to large and medium boxes.

The behavioural action annotations cover 9 basic behavioural actions; sitting, standing, walking, running, climbing up, climbing down, hanging, sitting on back, and camera interaction. We refer to these classes as behavioural actions in recognition of historical traditions in biological and computer vision disciplines, which would consider them behaviours and actions, respectively. Figure 4 displays the behavioural actions classes in focus together with their per-frame distribution. The class distribution is severely imbalanced, with the majority of samples ($> 85\%$) belonging to three *head* classes (i.e., sitting, walking and standing). The remaining behavioural actions are referred to as *tail* classes. The same imbalance is observed at the clip level, as shown in Table 1, although the distribution of classes across
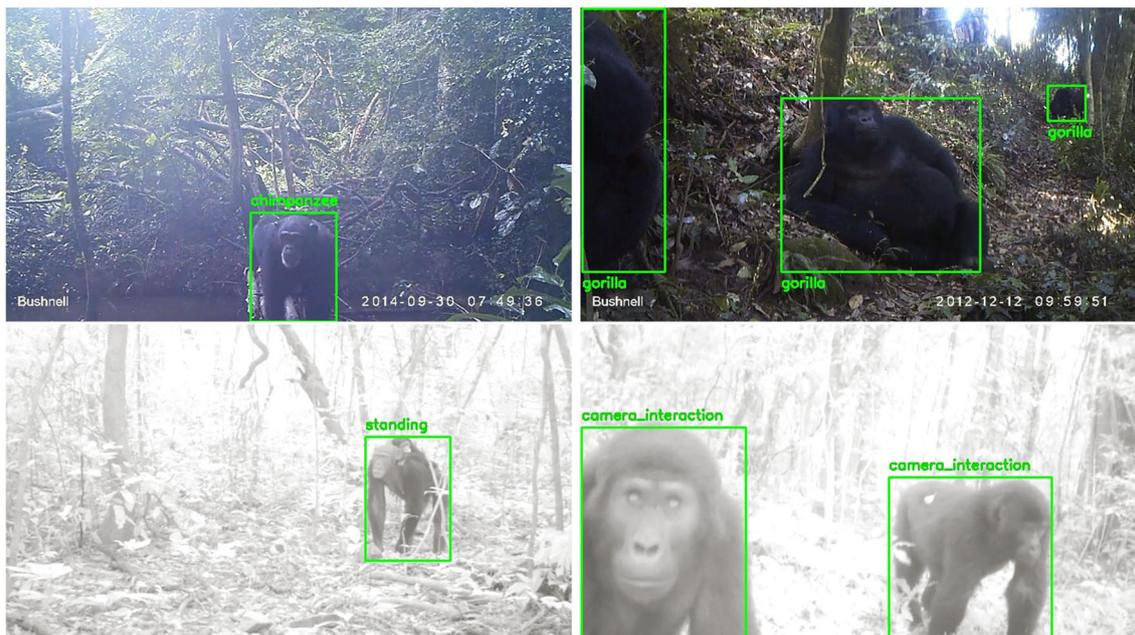
**Fig. 2** Manually annotated full-body location, species and behavioural action labels. Sample frames extracted from PanAf20K videos with species (row 1) and behavioural action annotations (row 2) displayed. Green bounding boxes indicate the full-body location of an ape. Species and behavioural action annotations are shown in the corresponding text
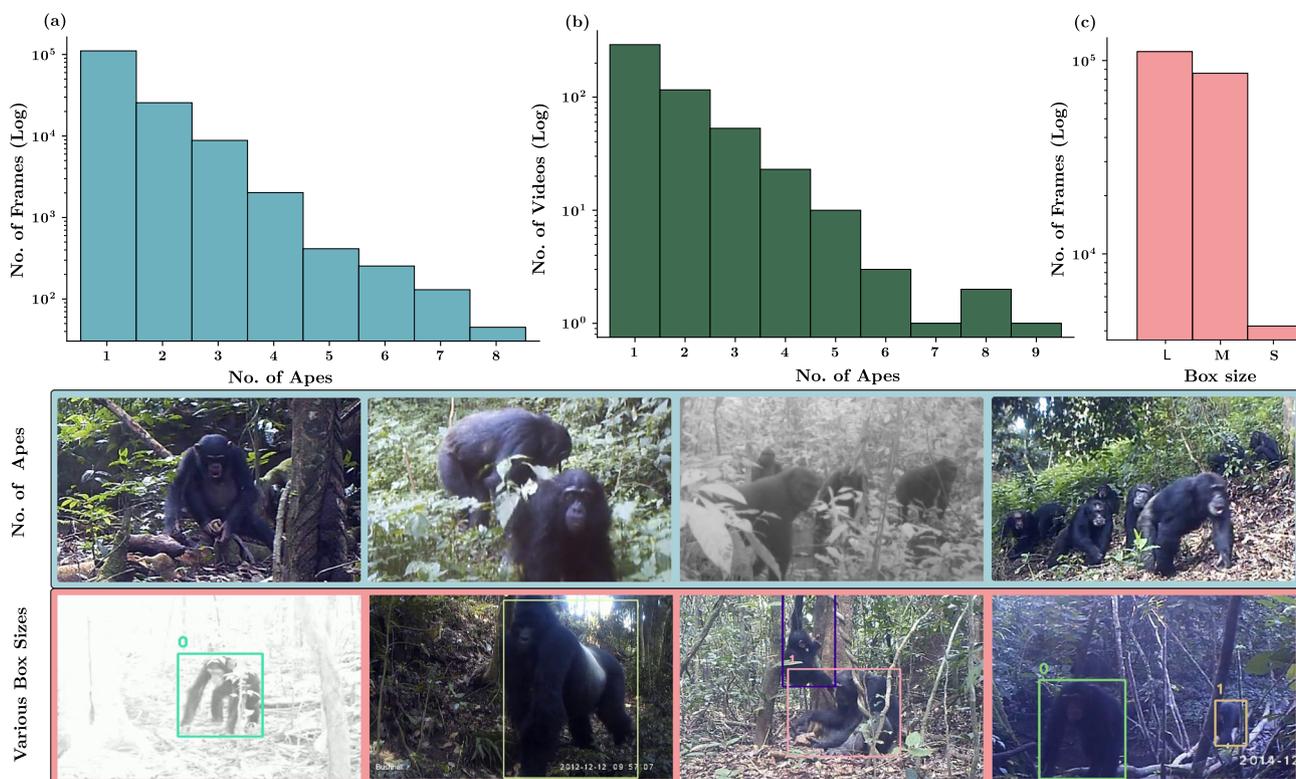


**Fig. 3** Number of apes & bounding box size distribution in the PanAf500 data. The top row shows the distribution of apes across frames and videos in (**a**) and (**b**), respectively, while the distribution of bounding box sizes is shown in (**c**). The middle row shows still frame examples of videos containing one, two, four and eight apes (viewing from left to right). The bottom row demonstrates still frames with bounding boxes of various sizes; the colour of bounding box and associated number represent the intra-video individual IDs
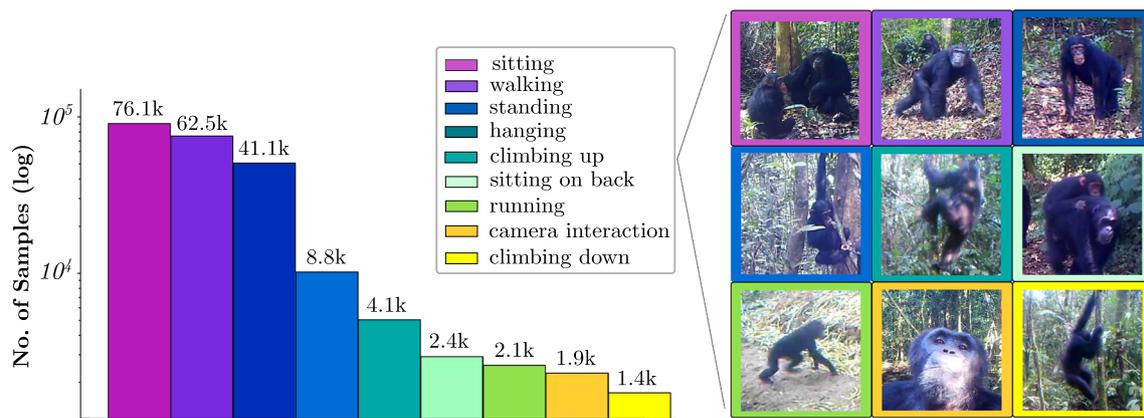
**Fig. 4** Behavioural actions in the PanAf500 data. Examples of each one of the nine behavioural action classes (*right*) and their distribution (*left*) across 500 videos. The total number of per-frame annotations for each behavioural action class is shown on top of the corresponding bar (Color figure online)

**Table 1** Behavioural action class statistics. The total number of clips for each behavioural action alongside the average duration in seconds and frames

| Action | Clips | Time (s) | Frames |
|---|---|---|---|
| walking | 747 | $3.49 \pm 2.94$ | 83.69 |
| standing | 366 | $4.77 \pm 4.59$ | 114.57 |
| sitting | 308 | $10.30 \pm 5.51$ | 247.13 |
| climbing up | 81 | $2.11 \pm 1.67$ | 50.59 |
| hanging | 50 | $7.35 \pm 5.24$ | 176.28 |
| climbing down | 35 | $1.73 \pm 1.24$ | 41.57 |
| running | 34 | $2.61 \pm 2.06$ | 62.59 |
| camera interaction | 32 | $2.52 \pm 3.77$ | 60.59 |
| sitting on back | 26 | $3.89 \pm 4.04$ | 93.46 |

The total number of clips for each behavioural action alongside the average duration in seconds and frames

clips does not match the per-frame distribution exactly. While behavioural actions with longer durations (i.e., sitting) have more labelled frames, this does not necessarily translate to more clips. For example, there are more clips of walking and standing than sitting, and more clips of climbing up than hanging, although the latter have fewer labelled frames.

*Multi-label Behavioural Annotation of PanAf20K* Community scientists on the Chimp&See platform provided multi-label behavioural annotations for ~20,000 videos. They were shown 15-second clips at a time and asked to annotate whether animals were present or whether the clip was blank. To obtain a balance between specificity and keeping the task accessible and interesting to a broad group of people, annotators were presented with a choice of classification categories. These categories allowed focus to be given to ecologically important behaviours such as tool use, camera reaction and bipedalism. Hashtags for behaviours not listed in the classification categories were also permitted, allowing new and interesting behaviours to be added when they were discovered in the videos. The new behaviours were subcat-

egories of the existing behaviours, many of them relating to tool use (e.g., algae scooping and termite fishing in aboreal nests).

To ensure annotation quality and consistency a video was only deemed to be analyzed when either three volunteers marked the video as blank, unanimous agreement between seven volunteers was observed, or 15 volunteers annotated the video. These annotations were then extracted and expertly grouped into 18 co-occurring classes, which form the multi-label behavioural annotations presented here. The annotations follow a multi-hot binary format that indicates the presence of one or many behaviours. It should also be noted that behaviours are not assigned to individual apes or temporally localised within each video. Figure 5 presents examples for several of the most commonly occurring behaviours. Figure 6 shows the full distribution of behaviours across videos, which is highly imbalanced. Four of the most commonly occurring classes are observed in > 60% videos, while the least commonly occurring classes are observed in < 1%. The relationship between behaviours is shown in Fig. 7 which presents co-occurring classes. The behaviours differ from the behavioural actions included in the PanAf500 dataset, corresponding to higher order behaviours that are commonly monitored in ecological studies. For example, instances of travel refer to videos that contain an individual or group of apes travelling, whereas associated behavioural actions such as walking or running may occur in many contexts (i.e., walking towards another ape during a social interaction or while searching for a tool).

Both parts of the dataset are suitable for different computer vision tasks. The PanAf500 supports great ape detection, tracking, action grounding, and multi-class action recognition, while the PanAf20k supports multi-label behaviour recognition. The difference between the two annotation types can be observed in Fig. 8.

**Fig. 5** PanAf20K behaviour examples. Triplets of example frames for six categories (i.e., feeding, travel, camera reaction, social interaction, chimp carrying and tool use) in the PanAf20K dataset are shown. Note that camera reaction, social interaction and chimp carrying have been abbreviated to reaction, social and carrying, respectively

*Machine Labels for Animal Location and IDs* We generated full-body bounding boxes for apes present in the remaining, unlabelled videos using state-of-the-art (SOTA) object detection models evaluated on the PanAf500 dataset (see Sect. 4). Additionally, we assigned intra-video IDs to detected apes using the multi-object tracker, OC-SORT (Cao et al., 2023). Note that these pseudo-labels do not yet associate behaviours with individual bounding boxes.

## 4 Experiments and Results

This section describes experiments relating to the PanAf500 and PanAf20K datasets. For the former, we present benchmark results for great ape detection and fine-grained action recognition. For the latter, we present benchmark results for multi-label behavioural classification. For both sets of experiments, several SOTA architectures are used.

### 4.1 PanAf500 Dataset

*Baseline Models* We report benchmark results for ape detection and fine-grained behavioural action recognition for the PanAf500 dataset, trained and evaluated on SOTA architectures. For ape detection, this entails the MegaDetector (Beery et al., 2019), ResNet-101 (+SCM+TCM) (Yang et al., 2019), VarifocalNet (VFNet) (Zhang et al., 2021), Swin-
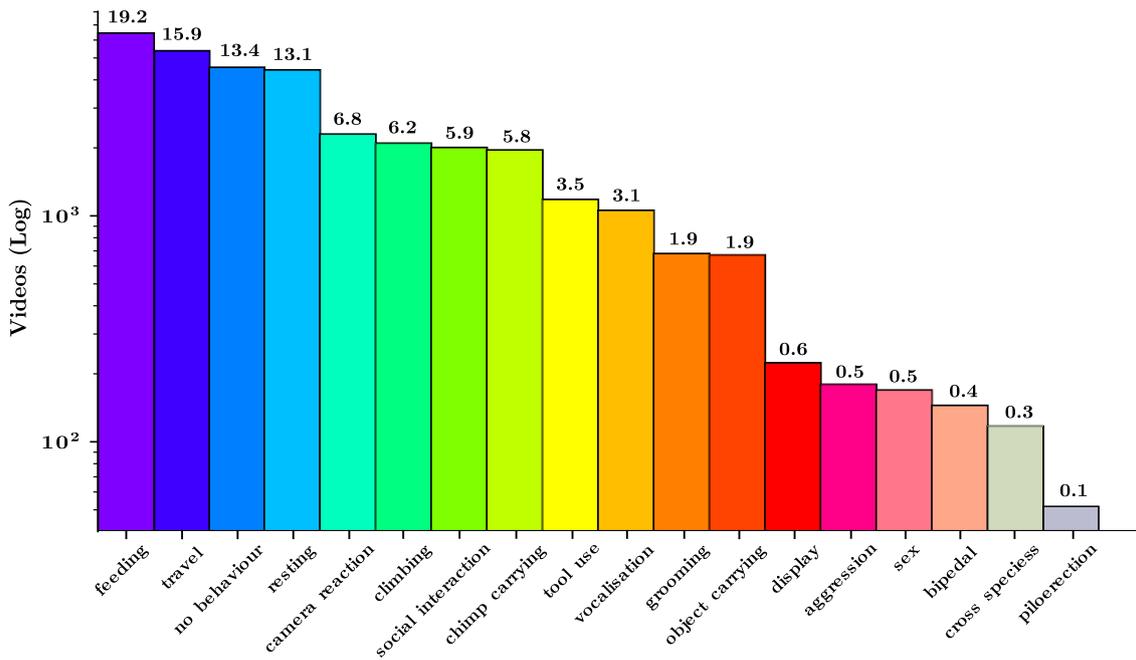
**Fig. 6** Behavioural annotations of the PanAf20K dataset. The distribution of behaviour categories for the PanAf20K dataset is shown. Figures above each bar represent the dataset proportion (%) of each class
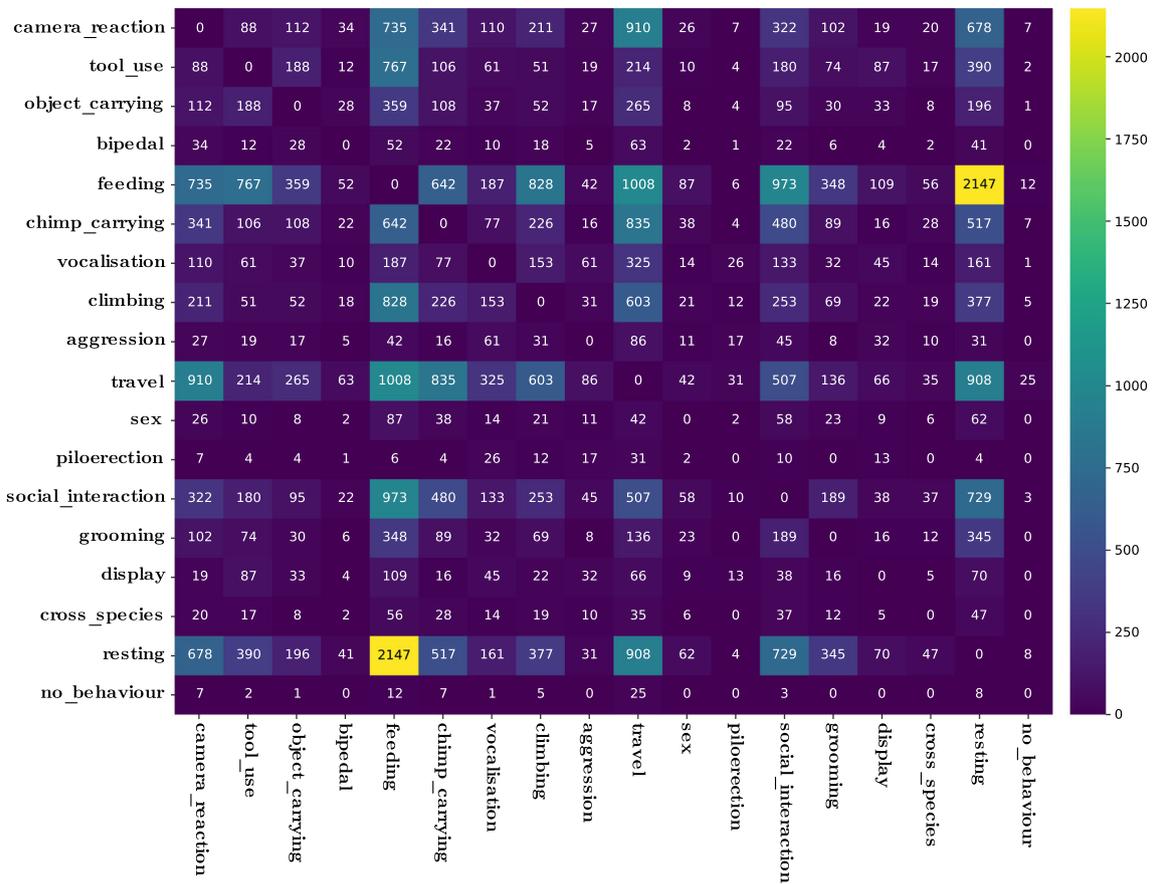


**Fig. 7** Co-occurrence of behaviours in the PanAf20k dataset. A co-occurrence matrix for the PanAf20K behaviours, where each cell reflects the number of times two behaviours occurred together. Diagonal cells are reset to aid visibility

**(a) fine-grained**

**(b) multi-label**



00:00 - 00:10
chimp 1: running

00:00 - 00:05
chimp 2: standing

00:05 - 00:10
chimp 2: running

tool use (nut cracking)
climbing (hanging)
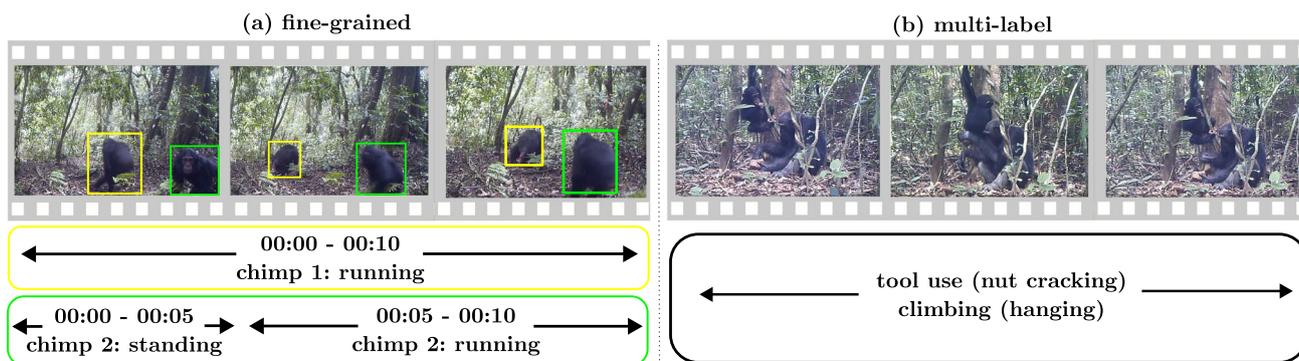
**Fig. 8** Examples of fine-grained and multi-label annotations. For videos with fine-grained annotations, full-body locations and behavioural actions are associated with each ape on a frame-by-frame basis (left). In contrast, multi-label behaviour annotations are provided at the video level (right); behaviours are not localised or assigned specifically to each ape

**Table 2** Ape detection benchmarks. Detection performance on the PanAf500 dataset. Results are reported for the MegaDetector (Beery et al, 2019), ResNet-101 (+SCM+TCM) (Yang et al, 2019), VarifocalNet (Zhang et al, 2021), Swin Transformer (Liu et al, 2021) and ConvNeXt (Liu et al, 2022). The highest scores for each metric are shown in bold

| Model | mAP (%) | | | | Other (%) | | |
|---|---|---|---|---|---|---|---|
| | All | L | M | S | Precision | Recall | F1 |
| MegaDetector | **88.0** | **98.05** | 82.60 | 68.21 | 56.93 | 90.58 | 69.92 |
| ResNet-101 | 81.2 | 77.60 | 88.97 | **88.84** | 42.37 | 88.93 | 57.40 |
| VarifocalNet | 84.1 | 84.50 | 88.73 | 82.07 | 21.73 | 88.57 | 34.90 |
| Swin Transformer | 87.2 | 82.47 | **96.86** | 88.53 | **83.66** | **92.03** | **87.65** |
| ConvNeXt | 86.6 | 83.51 | 95.16 | 81.13 | 81.80 | 91.93 | 86.57 |
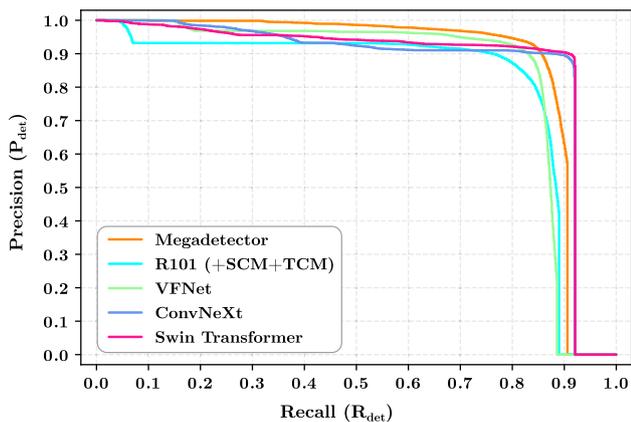


**Fig. 9** Megadetector (Beery et al., 2019) achieves higher precision for the majority of cases although ConvNeXt (Liu et al., 2022) and Swin Transformer (Liu et al., 2021) achieve better precision scores at high recall rates ($R_{det} > 0.84$)
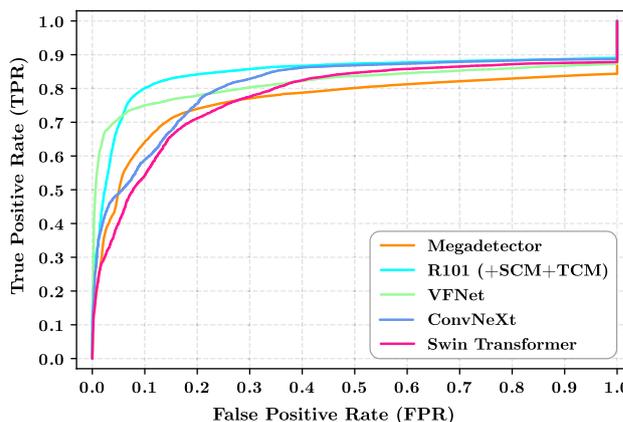


**Fig. 10** R101 (+SCM+TCM) (Yang et al., 2019) and VFNet (Zhang et al., 2021) achieve the highest true positive rates at low false positive rates ($FPR < 0.15$). At higher false positive rates R101(+SCM+TCM) (Yang et al., 2019) performs better

Transformer (Liu et al., 2021) and ConvNext (Liu et al., 2022) architectures. For fine-grained action recognition, we considered X3D (Feichtenhofer, 2020), I3D (Carreira & Zisserman, 2017), 3D ResNet-50 (Tran et al., 2018), Timesformer (Bertasius et al., 2021) and MViTv2 (Li et al., 2022) architectures. Action recognition models were chosen based on SOTA performance on human action recognition datasets and to be consistent with the best performing models on the AnimalKingdom (Ng et al., 2022) and MammalNet datasets (Chen et al., 2023). In all cases, train-val-test (80:05:15) splits were generated at the video-level to ensure generalisation across video/habitat and splits remained consistent across tasks.

**Fig. 11** Megadetector detection examples. A sequence of frames (along each row) extracted from 3 different videos. The ground truth bounding boxes (green) are shown alongside detections (red). The first sequence (*row 1*) shows successful detections. The second set of sequences (*row 2–4*) provide examples of false positive detections. The third set of sequences (*row 5–6*) provide examples of false negative detections
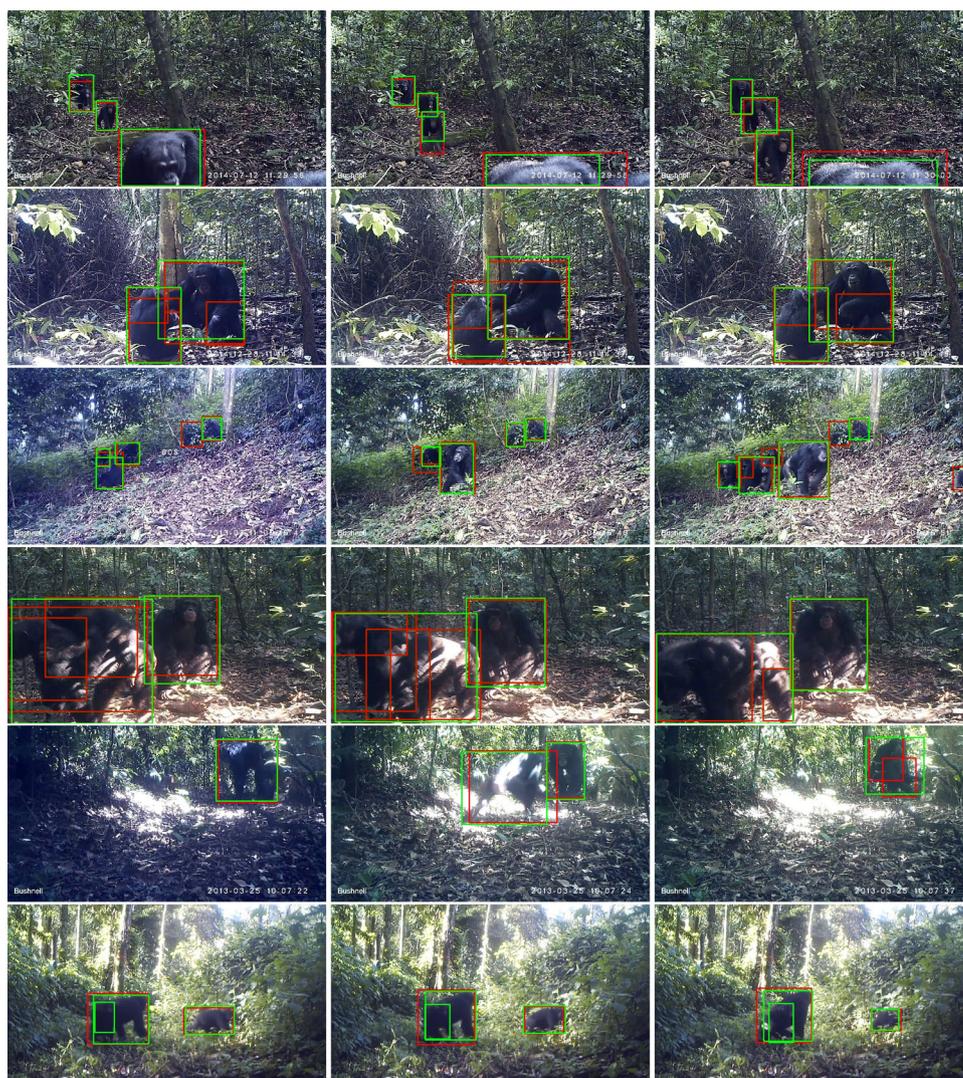
*Great Ape Detection* We initialised all models with pretrained feature extractors. For all models, except the Megadetector, we utilised MS COCO (Lin et al., 2014) pretrained weights. We use the out-of-the-box Megadetector implementation since it is pretrained on millions of camera trap images and provides a strong initialisation for camera trap specific detection tasks. We then fine-tuned each model for 50 epochs using SGD with a batch size of 16. Training was carried out using an input image resolution of $416^2$ and an Intersection over Union (IoU) threshold of 0.5 for non maximum suppression, at an initial learning rate of $1 \times 10^{-2}$ which was reduced by 10% at 80 and 90% of the total training epochs. All ape detection models were evaluated using the commonly used object detection metrics: mean average precision (mAP), precision, recall and F1-scores. All metrics follow the open images standard (Krasin et al., 2017) and are considered in combination during evaluation. Performance is provided separately for small ($32^2$), medium ($96^2$) and large bounding boxes ($> 96^2$), as per the COCO object detection standard, in addition to overall performance.

*Performance* Table 2 shows that the fine-tuned Megadetector achieves the best mAP score overall and for large bounding boxes, although it is outperformed by the Swin Transformer and ResNet-101 (+Cascade R-CNN+SCM+TCM) on medium and small bounding boxes, respectively. This shows that in-domain pre-training of the feature extractor is valuable for fine-tuning since the Megadetector is the only model pretrained on a camera trap dataset, rather than the COCO dataset (Lin et al., 2014). Performance across the remaining metrics, precision, recall and F1-score, is dominated by the Swin Transformer, which shows the importance of modelling spatial dependencies for good detection performance.

The precision-recall (PR) curve displayed in Fig. 9 shows that most models maintain precision of more than 90% ($P_{det} > 0.9$) at lower recall rates ($R_{det} < 0.80$), except
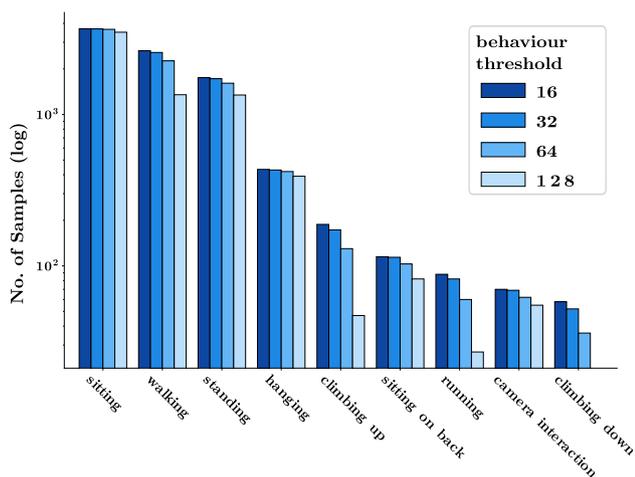
**Fig. 12** Per-class distribution vs. behavioural thresholds. Distribution of each behavioural action class at various behavioural thresholds. Note that tail classes are effected more significantly by longer thresholds than head classes

ResNet-101 (+SCM+TCM) which falls below this at recall of 78% ($R_{det} = 0.78$). The fine-tuned Megadetector achieves consistently higher precision than other models for more than 84% of cases ($R_{det} = 0.84$), outperforming other models by 5% ($P_{det} = 0.05$) on average. However, at higher recall rates ($R_{det} > 0.84$) ConvNeXt and SwinTransformer achieve higher precision, with the latter achieving marginally better performance. The ROC curve presented in Fig. 10 shows that VFNet and ResNet-101 (+SCM+TCM) achieve higher true positive rate than all other models at false positive rates less than 5% ($FPR < 0.05$) and 40% ($FPR < 0.40$), respectively. At higher false positive rates ConvNext and SwinTransformer are competitive with ResNet-101 (+SCM+TCM), with marginally better performance being established by ConvNeXt at very high false positive rates. Figure 11 presents qualitative examples of success and failure cases for the best performing model.

*Behavioural Action Recognition* We trained all models using the protocol established by Sakib and Burghardt (2020). During training we imposed a temporal behaviour threshold that ensures that only frame sequences in which a behaviour is exhibited for $t$ consecutive frames are utilised during training in order to retain well-defined behaviour instances. We then sub-sampled 16-frame sequences from clips that satisfy the behaviour threshold. The test threshold is always kept consistent ($t = 16$). Figure 12 shows the effect of different behaviour thresholds on the number of clips available for each class. Higher behaviour thresholds have a more significant effect on minority/tail classes since they occur more sporadically. For example, there are no training clips available for the climbing down class where $t = 128$. All models were initialised with feature extractors pre-trained on Kinetics-400 (Kay et al., 2017) and fine-tuned for 200 epochs

using the Adam optimiser and a standard cross-entropy loss. We utilised a batch size of 32, momentum of 0.9 and performed linear warm-up followed by cosine annealing using an initial learning rate of $1 \times 10^{-5}$ that increases to $1 \times 10^{-4}$ over 20 epochs. All behavioural action recognition models were evaluated using average top-1 and average per-class accuracy (C-Avg).

*Performance* Table 3 shows the X3D model attains the best top-1 accuracy at behaviour thresholds $t = 16$ and $t = 64$, although similar performance is achieved by MViTV2 and TimeSformer for the latter threshold. It also achieves the best average per-class performance at $t = 64$, while TimeSformer achieves the best performance at $t = 32$ and $t = 128$.

The MVITV2 models realise the best top-1 accuracy at $t = 32$ and $t = 128$, although they do not achieve the best average per-class performance at any threshold. The 3D ResNet-50 achieves the best average per-class performance at $t = 16$. When considering top-1 accuracy, model performance is competitive. At lower behavioural thresholds, i.e., $t = 16$ and $t = 32$, the difference in top-1 performance is 2.55 and 4.68%, respectively, between the best and worst performing models, although this increases to 5.38 and 11.74% at $t = 64$ and $t = 128$, respectively. There is greater variation in average per-class performance and it is rare that a model achieves the best performance across both metrics.

Although we observe strong performance with respect to top-1 accuracy, our models exhibit relatively poor average per-class performance. Figure 13 plots per-class performance against class frequency and shows that the average per-class performance is caused by poor performance on tail classes. The average per-class accuracy across all models for the head classes is 83.22% while only 28.33% is achieved for tail classes. There is significant variation in the performance of models; I3D performs well on hanging and climbing up but fails to classify any of the other classes correctly. Similarly, X3D performs extremely well on sitting on back but achieves poor results on the other classes. None of the models except for TimeSformer correctly classify any instances of running during testing. Figure 14 presents the confusion matrix calculated on validation data alongside examples of misclassified instances.

### 4.2 PanAf20K Dataset

*Data Setup* We generate train-val-test splits (70:10:20) using iterative stratification (Sechidis et al., 2011; Szymanski & Kajdanowicz, 2019). During training, we uniformly subsample $t = 16$ frames from each video, equating to $\sim 1$ frame per second (i.e., a sample interval of 22.5 frames).

*Baseline Models* To establish benchmark performance for multi-label behaviour recognition, we trained the X3D, I3D, 3D ResNet-50s, and MViTv2 models. All models were initialised with feature extractors pre-trained on Kinetics-400

**Table 3** Behavioural action recognition benchmarks. Behavioural action recognition performance on the PanAf500 dataset. Results are reported for X3D (Feichtenhofer, 2020), I3D (Carreira & Zisserman, 2017), 3D ResNet-50 (Hara et al., 2017), MViTV2 (Li et al., 2022), and TimeSformer (Bertasius et al., 2021) models. The highest scores for top-1 and average per-class accuracy are shown in bold

| Model | Top-1 (%) | | | | C-Avg (%) | | | |
|---|---|---|---|---|---|---|---|---|
| | 16 | 32 | 64 | 128 | 16 | 32 | 64 | 128 |
| X3D | **80.00** | 80.04 | **79.40** | 74.24 | 50.35 | 56.10 | **53.02** | 40.89 |
| I3D | 79.29 | 78.48 | 76.90 | 67.45 | 42.15 | 48.14 | 31.65 | 24.46 |
| 3D ResNet-50 | 77.45 | 76.41 | 74.02 | 73.31 | **55.17** | 33.79 | 38.72 | 36.03 |
| MViTV2 | 78.31 | **81.09** | 79.29 | **79.19** | 40.45 | 54.91 | 48.28 | 41.11 |
| TimeSformer | 78.53 | 79.45 | 79.26 | 78.18 | 45.05 | **56.38** | 48.27 | **41.10** |
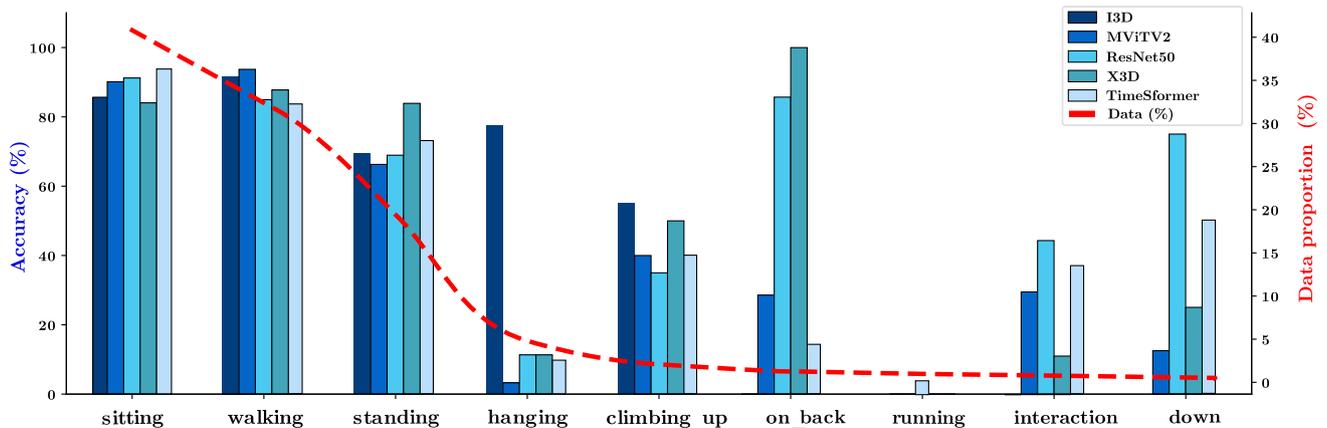


**Fig. 13** Class-wise performance vs. proportion of data. The per-class accuracy for each behavioural action recognition model is plotted against the proportion of data for each class. All models consistently achieve strong performance on the head classes, whereas performance is variable across tail classes

(Kay et al., 2017) and fine-tuned for 200 epochs using the Adam optimiser. We utilised a batch size of 32, momentum of 0.9 and performed linear warm-up followed by cosine annealing using an initial learning rate of $1 \times 10^{-5}$ that increases to $1 \times 10^{-4}$ over 20 epochs. Models were evaluated using mAP, subset accuracy (i.e., exact match), precision and recall. Behaviour classes were grouped, based on class frequency, into head ($> 10\%$), middle ($> 1\%$) and tail ($< 1\%$) segments, and mAP performance is reported for each segment. To address the long-tailed distribution, we substitute the standard loss for those calculated using long-tail recognition techniques. Specifically, we implement (i) focal loss (Cui et al., 2019) $L_{CB}$; (ii) logit adjustment (Menon et al., 2020) $L_{LA}$; and (iii) focal loss with weight balancing via a MaxNorm constraint (Alshammari et al., 2022).

*Multi-label Behaviour Recognition* As shown in Table 4, performance is primarily dominated by the 3D ResNet-50 and TimeSformer models when coupled with the various long-tailed recognition techniques. The TimeSformer (+LogitAdjustment) attains the highest mAP scores for both overall and tail classes, while the MViTV2 (+FocalLoss) and 3D ResNet-50 (+FocalLoss) demonstrate superior performance in terms of head and middle class mAP, respectively. The 3D

ResNet-50 (+FocalLoss) and 3D ResNet-50 (+WeightBalancing) models achieve the best subset accuracy and recall, respectively, while the highest precision is realised by the TimeSformer (+LogitAdjustment) model. Although the 3D ResNet-50 and TimeSformer models perform strongest, it should be noted that the difference in overall mAP across all models is small (i.e., 4.03% between best and worst performing models).

As demonstrated by the head, middle and tail mAP scores, higher performance is achieved for more frequently occurring classes with performance deteriorating significantly for middle and tail classes. Across models, the average difference between head and middle, and middle and tail classes is 35.68 ($\pm$1.88)% and 40.55 ($\pm$3.02)%, respectively. The inclusion of long-tailed recognition techniques results in models that consistently attain higher tail class mAP performance than their standard counterparts (i.e., models that do not use long-tail recognition techniques). The logit adjustment technique consistently results in the best tail class mAP across models, whereas the focal loss results in the best performance on the middle classes for all models except the X3D model. None of the standard models achieve the best performance on any metric.
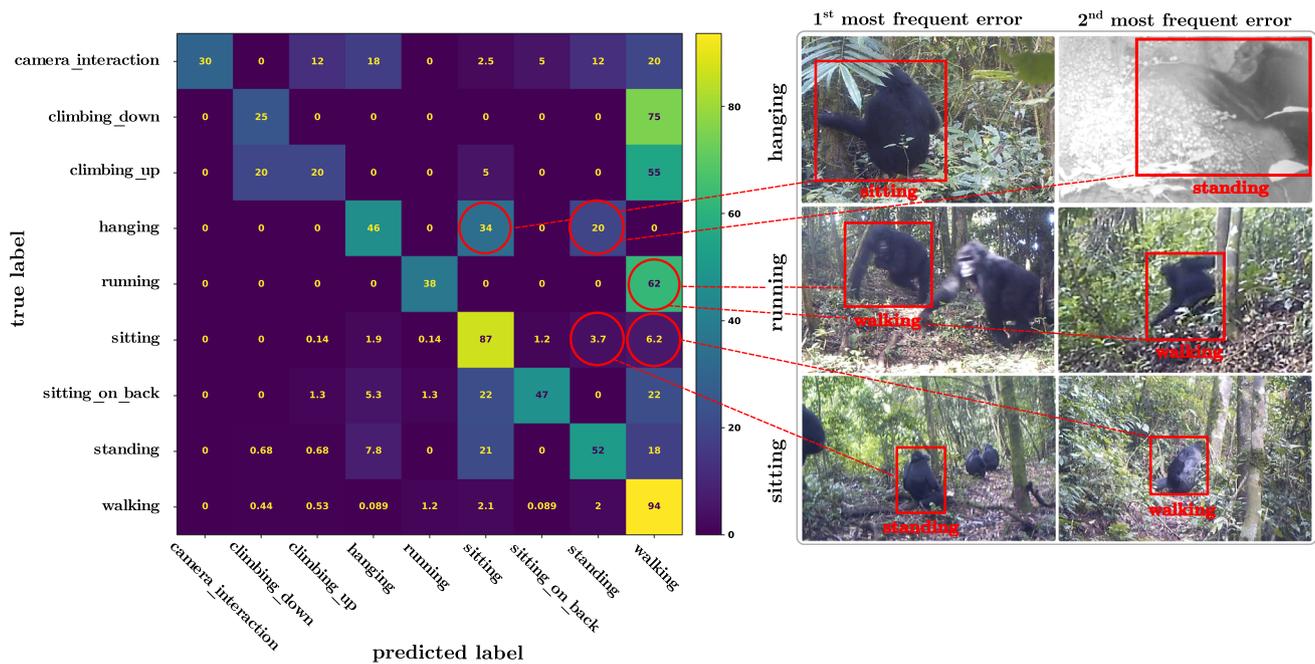
**Fig. 14** Confusion matrix & example errors. The confusion matrix (left) is shown alongside examples of mis-classified frames (right). For misclassified examples, ground truth labels are shown on the y-axis (i.e., hanging, running, sitting) and examples of the classes most likely to be incorrectly predicted for the ground truth class are shown on the x-axis. Note that a high proportion of errors are due to predictions made in favour of majority classes

Figure 15 plots per-class mAP performance of the 3D ResNet-50 and 3D ResNet-50 (+LogitAdjustment) models against the per-class proportion of data. The best performance is observed for the three most commonly occurring classes (i.e., feeding, travel, and no behaviour) whereas the worst performance is obtained by the most infrequently occurring classes (i.e., display, aggression, sex, bipedal, and cross species interaction) with the exception of piloerection. It can also be observed that the ResNet-50 (+LogitAdjustment) model outperforms its standard counterpart on the majority of middle and tail classes, although it is outperformed on tail classes. Examples of success and failure cases by the 3D ResNet-50 model are presented in Fig. 16.

## 5 Discussion and Future Work

*Results* The performance of current SOTA methods is not currently sufficient for facilitating the large-scale, automated behavioural monitoring required to support conservation efforts. The conclusions drawn in ecological studies rely on the highly accurate classification of all observed behaviours by expert primatologists. While the current methods achieve strong performance on head classes, relatively poor performance is observed for rare classes. Our results are consistent with recent work on similar datasets (i.e., AnimalKingdom (Ng et al., 2022) and MammalNet (Chen et al., 2023)) which demonstrate the significance of the long-tailed dis-

tribution that naturally recorded data exhibits (Liu et al., 2019). Similar to (Ng et al., 2022), our experiments show that current long-tailed recognition techniques can help to improve performance on tail classes, although a large discrepancy between head and middle, and head and tail classes still exists. The extent of this performance gap (see Table 4) emphasises the difficulty of tackling long-tailed distributions and highlights an important direction for future work (Perrett et al., 2023). Additionally, the near perfect performance at training time (i.e., > 95% mAP) highlights the need for methods that can learn effectively from a minimal number of examples.

*Community Science and Annotation* Although behavioural annotations are provided by non-expert community scientists, several studies have shown the effectiveness of citizen scientists to perform complex data annotation tasks (Danielsen et al., 2014; McCarthy et al., 2021) typically carried out by researchers (i.e., species classification, individual identification etc.). However, it should be noted that, as highlighted by Cox et al. (2012), community scientists are more prone to errors relating to rare species. In the case of our dataset, this may translate to simple behaviours being identified correctly (e.g., feeding and tool use) whereas more nuanced or subtle behaviours (e.g., display and aggression) are missed or incorrectly interpreted, amongst other problems. This may occur despite the behaviour categories were predetermined by experts as suitable for non-expert annotation.
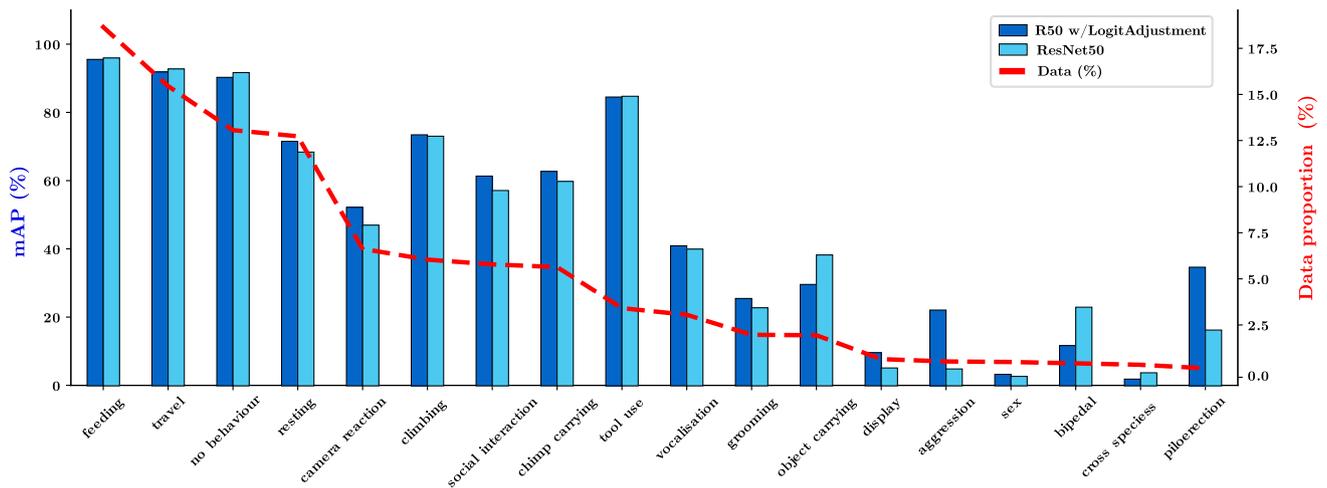
**Fig. 15** Class-wise accuracy vs. proportion of data. The per-class average precision for the 3D ResNet-50 (Hara et al., 2017) and 3D ResNet-50 (+LogitAdjustment) (Hara et al., 2017; Menon et al., 2020) models is plotted against the proportion of data for each class. In general, better model performance is achieved on classes with high data proportions and the ResNet-50 (+LogitAdjustment) model shows improved performance on middle and tail classes

**Fig. 16** Multi-label errors. Frames extracted from three videos exhibit success and failure cases of the 3D ResNet-50 model. Behaviour predictions are shown in light boxes of the first frame of each sequence; true positives are green, false positive are blue, and false negatives are red. In the first video (row 1), the model fails to classify feeding by the chimp visible in frames 1 and 2 whereas in the second video (row 2), it fails to classify tool use by the infant chimp in the final frame. Climbing is predicted incorrectly in the final video (row 3) (Color figure online)
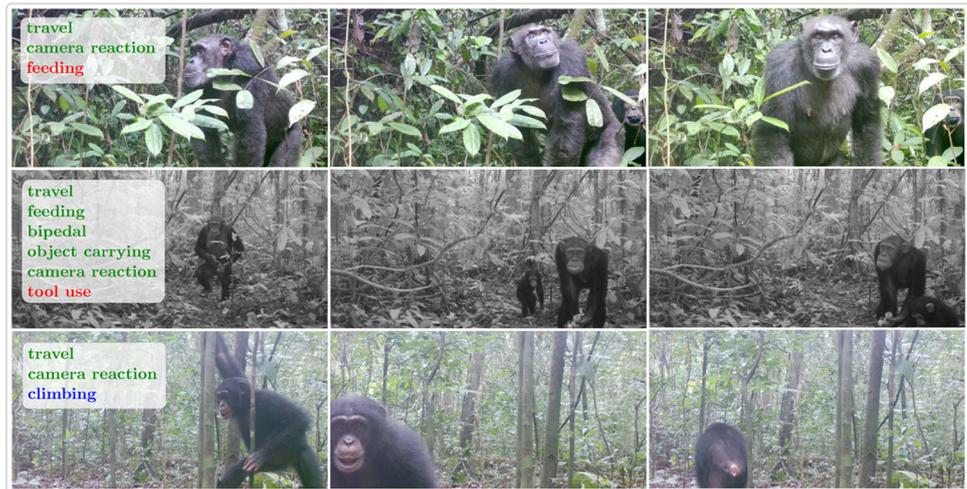


The dataset's rich annotations suit various computer vision tasks, despite key differences from other works. Unlike similar datasets (Chen et al., 2023; Ng et al., 2022), behaviours in the PanAf20K dataset are not temporally located within the video. However, the videos in our dataset are relatively short (i.e., 15 s) in contrast to the long form videos included in other datasets. Therefore, the time stamping of behaviour may be less significant considering it is possible to utilise entire videos, with a suitably fine-grained sample interval (i.e., 0.5–1 s), as input to standard action recognition models. With that being said, behaviours occur sporadically and chimpanzees are often only in frame for very short periods of time. Therefore, future work will consider augmenting the existing annotations with temporal localisation of actions. Moreover, while our dataset comprises a wide range of behaviour categories, many of them exhibit significant intra-class variation. In the context of ecological/primatological studies, this variation often necessitates the creation of separate ethograms for individual behaviours (Nishida et al., 1999; Zamma & Matsusaka, 2015). For instance, within the tool use behaviour category, we find subcategories like nut cracking (utilizing rock, stone, or wood), termite fishing, and algae fishing. Similarly, within the camera reaction category, distinct subcategories include attraction, avoidance, and fixation. In future, we plan to extend the existing annotations to include more granular subcategories.

*Ethics Statement* All data collection, including camera trapping, was done non-invasively, with no animal contact and no direct observation of the animals under study. Full research approval, data collection approval and research and sample permits of national ministries and protected area authorities were obtained in all countries of study. Sample

**Table 4** Multi-label behaviour recognition benchmarks. Results are reported for I3D (Carreira and Zisserman, 2017), 3D ResNet-50 (Hara et al, 2017), X3D (Feichtenhofer, 2020), MViTV2 (Li et al, 2022), and TimeSformer (Bertasius et al, 2021) models with focal loss (Cui et al, 2019), logit adjustment (Menon et al, 2020), and focal loss with weight balancing (Alshammari et al, 2022). The highest scores across all metrics are shown in bold

| Model | mAP (%) All | Head | Middle | mAP (%) Tail | Other (%) Accuracy | Precision | Recall |
|---|---|---|---|---|---|---|---|
| I3D | 45.49 | 87.92 | 53.43 | 6.62 | 41.99 | 51.77 | 38.62 |
| +FocalLoss | 46.65 | 87.67 | 53.51 | 10.17 | 42.51 | 60.02 | 37.46 |
| +LogitAdjustment | 46.81 | 87.52 | 52.54 | 12.05 | 43.02 | 57.99 | 38.88 |
| +WeightBalancing | 46.41 | 88.41 | 51.91 | 11.07 | 41.93 | 57.36 | 35.62 |
| 3D ResNet-50 | 46.03 | 86.12 | 53.22 | 9.73 | 40.76 | 54.13 | 35.87 |
| +FocalLoss | 47.93 | 87.07 | **54.31** | 13.35 | **43.35** | 57.77 | 38.89 |
| +LogitAdjustment | 48.04 | 87.44 | 53.91 | 13.96 | 41.15 | 58.86 | 36.27 |
| +WeightBalancing | 46.68 | 87.09 | 54.06 | 9.90 | 41.54 | 54.05 | **40.62** |
| X3D | 46.06 | 87.32 | 52.95 | 9.36 | 42.70 | 49.26 | 39.25 |
| +FocalLoss | 47.19 | 89.26 | 52.75 | 11.75 | 41.67 | 50.93 | 36.99 |
| +LogitAdjustment | 47.85 | 88.58 | 53.06 | 13.77 | 42.64 | 59.18 | 36.94 |
| +WeightBalancing | 45.64 | 88.45 | 51.15 | 9.75 | 40.96 | 48.40 | 33.29 |
| MViTV2 | 45.71 | 88.72 | 51.16 | 9.76 | 42.38 | 52.03 | 36.55 |
| +FocalLoss | 45.78 | **89.27** | 50.82 | 10.05 | 42.83 | 56.54 | 36.38 |
| +LogitAdjustment | 45.91 | 88.97 | 50.75 | 10.74 | 41.02 | 57.73 | 38.11 |
| +WeightBalancing | 45.36 | 88.58 | 49.82 | 10.59 | 42.44 | 49.77 | 36.20 |
| TimeSformer | 47.24 | 88.83 | 51.91 | 13.29 | 41.60 | 57.66 | 38.63 |
| +FocalLoss | 48.82 | 88.75 | 52.65 | 17.10 | 42.70 | 68.01 | 37.37 |
| +LogitAdjustment | **49.39** | 88.52 | 53.21 | **18.20** | 42.44 | **71.31** | 35.45 |
| +WeightBalancing | 48.17 | 87.98 | 51.81 | 16.78 | 41.86 | 61.16 | 39.36 |

Results are reported for I3D (Carreira & Zisserman, 2017), 3D ResNet-50 (Hara et al., 2017), X3D (Feichtenhofer, 2020), MViTV2 (Li et al., 2022), and TimeSformer (Bertasius et al., 2021) models with focal loss (Cui et al., 2019), logit adjustment (Menon et al., 2020), and focal loss with weight balancing (Alshammari et al., 2022). The highest scores across all metrics are shown in bold

and data export was also done with all necessary certificates, export and import permits. All work conformed to the relevant regulatory standards of the Max Planck Society, Germany. All community science work was undertaken according to the Zooniverse User Agreement and Privacy Policy. No experiments or data collection were undertaken with live animals.

# 6 Conclusion

We present by-far the largest open-access video dataset of wild great apes with rich annotations and SOTA benchmarks. The dataset is directly suitable for visual AI training and model comparison. The size of the dataset and extent of labelling across > 7 M frames and ∼ 20 K videos (lasting > 80 h) now offers the first comprehensive view of great ape populations and their behaviours to AI researchers. Task-specific annotations make the data suitable for a range of associated, challenging computer vision tasks (i.e, animal detection, tracking, and behaviour recognition) which can facilitate ecological analysis urgently required to support conservation efforts. We believe that given its immediate AI compatibility, scale, diversity, and accessibility, the PanAf20K dataset provides an unmatched opportunity for the many communities working in the ecological, biological, and computer vision domains to benchmark and expand great ape monitoring capabilities. We hope that this dataset can, ultimately, be a step towards better understanding and more effectively conserving these charismatic species.

# References

Alshammari, S., Wang, Y. X., Ramanan, D., & Kong, S. (2022). Long-tailed recognition via weight balancing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 6897–6907).

Arandjelovic, M., Stephens, C. R., McCarthy, M. S., Dieguez, P., Kalan, A. K., Maldonado, N., Boesch, C., & Kuehl, H. S. (2016). Chimp&See: An online citizen science platform for large-scale, remote video camera trap annotation of chimpanzee behaviour, demography and individual identification. *PeerJ Preprints*.

Bain, M., Nagrani, A., Schofield, D., Berdugo, S., Bessa, J., Owen, J., Hockings, K. J., Matsuzawa, T., Hayashi, M., Biro, D., & Carvalho, S. (2021). Automated audiovisual behavior recognition in wild primates. *Science Advances,7*(46), eabi4883

Beery, S., Agarwal, A., Cole, E., & Birodkar, V. (2021). The iwildcam 2021 competition dataset. arXiv preprint arXiv:2105.03494

Beery, S., Morris, D., & Yang, S. (2019). Efficient pipeline for camera trap image review. arXiv preprint arXiv:1907.06772

Bertasius, G., Wang, H., & Torresani, L. (2021). Is space-time attention all you need for video understanding? In *Proceedings of the international conference on machine learning (ICML)*.

Brookes, O., & Burghardt, T. (2020). A dataset and application for facial recognition of individual gorillas in zoo environments. In *Workshop on the visual observation and analysis of vertebrate and insect behaviour*. 2012.04689

Brookes, O., Mirmehdi, M., Kühl, H., & Burghardt, T. (2023). Triple-stream deep metric learning of great ape behavioural actions. In *Proceedings of the 18th international joint conference on computer vision, imaging and computer graphics theory and applications* (pp. 294–302).

Brust, C. A., Burghardt, T., Groenenberg, M., Kading, C., Kuhl, H. S., Manguette, M. L., & Denzler, J. (2017). Towards automated visual monitoring of individual gorillas in the wild. In *Proceedings of the IEEE international conference on computer vision workshops* (pp. 2820–2830).

Cao, J., Pang, J., Weng, X., Khirodkar, R., & Kitani, K. (2023) Observation-centric sort: Rethinking sort for robust multi-object tracking. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 9686–9696).

Carreira, J., & Zisserman, A. (2017). Quo vadis, action recognition? A new model and the kinetics dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. pp 6299–6308).

Carvalho, J. S., Graham, B., Bocksberger, G., Maisels, F., Williamson, E. A., Wich, S., Sop, T., Amarasekaran, B., Barca, B., Barrie, A.,

& Bergl, R. A. (2021). Predicting range shifts of African apes under global change scenarios. *Diversity and Distributions, 27*(9), 1663–1679.

Ceballos, G., Ehrlich, P. R., & Raven, P. H. (2020). Vertebrates on the brink as indicators of biological annihilation and the sixth mass extinction. *Proceedings of the National Academy of Sciences, 117*(24), 13596–13602.

Chappell, J., & Thorpe, S. K. (2022). The role of great ape behavioral ecology in one health: Implications for captive welfare and rehabilitation success. *American Journal of Primatology, 84*(4–5), e23328.

Chen, J., Hu, M., Coker, D. J., Berumen, M. L., Costelloe, B., Beery, S., Rohrbach, A., & Elhoseiny, M. (2023). Mammalnet: A large-scale video benchmark for mammal recognition and behavior understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 13052–13061).

Clark, F. E. (2011). Great ape cognition and captive care: Can cognitive challenges enhance well-being? *Applied Animal Behaviour Science, 135*(1–2), 1–12.

Cox, T. E., Philippoff, J., Baumgartner, E., & Smith, C. M. (2012). Expert variability provides perspective on the strengths and weaknesses of citizen-driven intertidal monitoring program. *Ecological Applications, 22*(4), 1201–1212.

Cui, Y., Jia, M., Lin, T. Y., Song, Y., & Belongie, S. (2019). Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 9268–9277).

Cui, Y., Song, Y., Sun, C., Howard, A., & Belongie, S. (2018). Large scale fine-grained categorization and domain-specific transfer learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. pp 4109–4118).

Danielsen, F., Jensen, P. M., Burgess, N. D., Altamirano, R., Alviola, P. A., Andrianandrasana, H., Brashares, J. S., Burton, A. C., Coronado, I., Corpuz, N., & Enghoff, M. (2014). A multicountry assessment of tropical resource monitoring by local communities. *BioScience, 64*(3), 236–251.

Fegraus, E. H., Lin, K., Ahumada, J. A., Baru, C., Chandra, S., & Youn, C. (2011). Data acquisition and management software for camera trap data: A case study from the team network. *Ecological Informatics, 6*(6), 345–353.

Feichtenhofer, C. (2020). X3d: Expanding architectures for efficient video recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 203–213).

Freytag, A., Rodner, E., Simon, M., Loos, A., Kühl, H. S., & Denzler, J. (2016). Chimpanzee faces in the wild: Log-Euclidean CNNs for predicting identities and attributes of primates. In *German conference on pattern recognition* (pp. 51–63). Springer.

Hara, K., Kataoka, H., Satoh, & Y. (2017). Learning spatio-temporal features with 3d residual networks for action recognition. In *Proceedings of the IEEE international conference on computer vision workshops* (pp. 3154–3160).

Haurez, B., Daïnou, K., Tagg, N., Petre, C. A., & Doucet, J. L. (2015). The role of great apes in seed dispersal of the tropical forest tree species Dacryodes normandii (Burseraceae) in Gabon. *Journal of Tropical Ecology, 31*(5), 395–402.

Houa, N. A., Cappelle, N., Bitty, E. A., Normand, E., Kablan, Y. A., & Boesch, C. (2022). Animal reactivity to camera traps and its effects on abundance estimate using distance sampling in the Taï National Park, Côte d'ivoire. *PeerJ, 10*, e13510.

Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P., & Suleyman, M. (2017). The kinetics human action video dataset. arXiv preprint arXiv:1705.06950

Krasin, I., Duerig, T., Alldrin, N., Ferrari, V., Abu-El-Haija, S., Kuznetsova, A., Rom, H., Uijlings, J., Popov, S., Veit, A., & Belongie, S. (2017). Openimages: A public dataset for large-scale

multi-label and multi-class image classification. Dataset available from https://storage.googleapis.com/openimages/web/index.html

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems* (pp. 1097–1105).

Kuehne, H., Jhuang, H., Garrote, E., Poggio, T., Serre, T., (2011). HMDB: A large video database for human motion recognition. In *2011 International conference on computer vision* (pp. 2556–2563). IEEE.

Kühl, H. S., & Burghardt, T. (2013). Animal biometrics: Quantifying and detecting phenotypic appearance. *Trends in Ecology & Evolution, 28*(7), 432–441.

Li, Y., Wu, C.Y., Fan, H., Mangalam, K., Xiong, B., Malik, J., & Feichtenhofer, C. (2022). Mvitv2: Improved multiscale vision transformers for classification and detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 4804–4814).

Lin, T. Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., & Zitnick, C. L. (2014). Microsoft coco: Common objects in context. In *Proceedings of the European conference on computer vision* (pp. 740–755). Springer.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., & Guo, B. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 10012–10022).

Liu, Z., Mao, H., Wu, C. Y., Feichtenhofer, C., Darrell, T., & Xie, S. (2022). A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 11976–11986).

Liu, Z., Miao, Z., Zhan, X., Wang, J., Gong, B., & Yu, S. X. (2019). Large-scale long-tailed recognition in an open world. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 2537–2546).

McCarthy, M. S., Stephens, C., Dieguez, P., et al. (2021). Chimpanzee identification and social network construction through an online citizen science platform. *Ecology and Evolution, 11*(4), 1598–1608.

Menon, A. K., Jayasumana, S., Rawat, A. S., Jain, H., Veit, A., & Kumar, S. (2020). Long-tail learning via logit adjustment. In *Proceedings of the international conference on learning representations*.

Ng, X. L., Ong, K. E., Zheng, Q., Ni, Y., Yeo, S. Y., & 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* (pp. 19023–19034).

Nishida, T., Kano, T., Goodall, J., McGrew, W. C., & Nakamura, M. (1999). Ethogram and ethnography of Mahale chimpanzees. *Anthropological Science, 107*(2), 141–188.

Parkhi, O., Vedaldi, A., & Zisserman, A. (2015). Deep face recognition. In *Proceedings of the British machine vision conference*. British Machine Vision Association.

Perrett, T., Sinha, S., Burghardt, T., Mirmehdi, M., & Damen, D. (2023). Use your head: Improving long-tail video recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 2415–2425).

Pollen, A. A., Kilik, U., Lowe, C. B., & Camp, J. G. (2023). Human-specific genetics: New tools to explore the molecular and cellular basis of human evolution. *Nature Reviews Genetics, 1*–25

Redmon, J., & Farhadi, A. (2017). Yolo9000: Better, faster, stronger. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 7263–7271).

Redmon, J., & Farhadi, A. (2018). Yolov3: An incremental improvement. arXiv preprint arXiv:1804.02767

Sakib, F., & Burghardt, T. (2020). Visual recognition of great ape behaviours in the wild. In *Workshop on the visual observation and analysis of vertebrate and insect behaviour*. 2011.10759.

Samuni, L., Crockford, C., & Wittig, R. M. (2021). Group-level cooperation in chimpanzees is shaped by strong social ties. *Nature Communications, 12*(1), 539.

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

Sechidis, K., Tsoumakas, G., & Vlahavas, I. (2011). On the stratification of multi-label data. In *Machine learning and knowledge discovery in databases* (pp. 145–158). Springer.

Simonyan, K., & Zisserman, A. (2014). Two-stream convolutional networks for action recognition in videos. In *Advances in neural information processing systems*.

Soomro, K., Zamir, A. R., & Shah, M. (2012). Ucf101: A dataset of 101 human actions classes from videos in the wild. arXiv preprint arXiv:1212.0402

Swanson, A., Kosmala, M., Lintott, C., Simpson, R., Smith, A., & Packer, C. (2015). Snapshot Serengeti, high-frequency annotated camera trap images of 40 mammalian species in an African savanna. *Scientific Data, 2*(1), 1–14.

Szymanski, P., & Kajdanowicz, T. (2019). Scikit-multilearn: A scikit-based python environment for performing multi-label classification. *The Journal of Machine Learning Research, 20*(1), 209–230.

Tarszisz, E., Tomlinson, S., Harrison, M. E., Morrogh-Bernard, H. C., & Munn, A. J. (2018). An ecophysiologically informed model of seed dispersal by orangutans: Linking animal movement with gut passage across time and space. *Conservation Physiology, 6*(1), coy013.

Tennie, C., Jensen, K., & Call, J. (2016). The nature of prosociality in chimpanzees. *Nature Communications, 7*(1), 13915.

Tran, D., Wang, H., Torresani, L., Ray, J., LeCun, Y., & Paluri, M. (2018). A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 6450–6459).

Tuia, D., Kellenberger, B., Beery, S., Costelloe, B. R., Zuffi, S., Risse, B., Mathis, A., Mathis, M. W., van Langevelde, F., Burghardt, T., & Kays, R. (2022). Perspectives in machine learning for wildlife conservation. *Nature Communications, 13*(1), 1–15.

Van Horn, G., Mac Aodha, O., Song, Y., Cui, Y., Sun, C., Shepard, A., Adam, H., Perona, P., & Belongie, S. (2018). The inaturalist species classification and detection dataset. In *Proceedings of the IEEE international conference on computer vision* (pp. 8769–8778).

Vié, J. C., Hilton-Taylor, C., & Stuart, S. N. (2009). *Wildlife in a changing world: An analysis of the 2008 IUCN Red List of threatened species*. IUCN

Yang, X., Burghardt, T., & Mirmehdi, M. (2023). Dynamic curriculum learning for great ape detection in the wild. *International Journal of Computer Vision, 1*–19

Yang, X., Mirmehdi, M., & Burghardt, T. (2019). Great ape detection in challenging jungle camera trap footage via attention-based spatial and temporal feature blending. In *Proceedings of the IEEE/CVF international conference on computer vision workshops*.

Zamma, K., & Matsusaka, T. (2015). *Ethograms and the diversity of behaviors* (pp. 510–518). Cambridge University Press.

Zhang, H., Wang, Y., Dayoub, F., & Sunderhauf, N. (2021). Varifocalnet: An iou-aware dense object detector. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 8514–8523).