

OpenApePose: a database of annotated ape photographs for pose estimation

Reviewed Preprint

Published from the original preprint after peer review and assessment by eLife.

About eLife's process

Reviewed preprint posted

July 13, 2023 (this version)

Sent for peer review

February 16, 2023

Posted to arXiv

November 30, 2022

Nisarg Desai , Praneet Bala, Rebecca Richardson, Jessica Raper, Jan Zimmermann, Benjamin Hayden

Department of Neuroscience and Center for Magnetic Resonance Research, University of Minnesota, Minneapolis MN 55455 • Department of Computer Science, University of Minnesota, Minneapolis MN 55455 • Emory National Primate Research Center, Emory University, Atlanta GA 30329

 https://en.wikipedia.org/wiki/Open_access

 <https://creativecommons.org/licenses/by/4.0/>

Abstract

Because of their close relationship with humans, non-human apes (chimpanzees, bonobos, gorillas, orangutans, and gibbons, including siamangs) are of great scientific interest. The goal of understanding their complex behavior would be greatly advanced by the ability to perform video-based pose tracking. Tracking, however, requires high-quality annotated datasets of ape photographs. Here we present *OpenApePose*, a new public dataset of 71,868 photographs, annotated with 16 body landmarks, of six ape species in naturalistic contexts. We show that a standard deep net (HRNet-W48) trained on ape photos can reliably track out-of-sample ape photos better than networks trained on monkeys (specifically, the *OpenMonkeyPose* dataset) and on humans (*COCO*) can. This trained network can track apes almost as well as the other networks can track their respective taxa, and models trained without one of the six ape species can track the held out species better than the monkey and human models can. Ultimately, the results of our analyses highlight the importance of large specialized databases for animal tracking systems and confirm the utility of our new ape database.

eLife assessment

The OpenApePose dataset presented in this manuscript represents an **important** contribution to the field of primate behaviour and computer-vision science with methodological applications that are sure to be applicable for a wide variety of taxa. The analysis supporting the utility of this database is **solid** and **compelling** but would benefit from some additional clarity, particularly with regards to the **landmark** annotations, model parameters and division of the dataset for training, validation and testing.

Introduction

The ability to automatically track moving animals using video systems has been a great boon for the life sciences, including biomedicine (Calhoun and Murthy, 2017 [↗](#); Marshall et al., 2022 [↗](#); Mathis and Mathis, 2020 [↗](#); Periera et al., 2020 [↗](#)). Such systems allow data collected from digital

video cameras to be used to infer the positions of body landmarks such as head, hands, and feet, without the use of specialized markers. In recent years, the field has witnessed the development of sophisticated tracking systems that can track and identify behavior in species important for biological research, including humans, worms, flies, and mice (e.g., [Bohnslav et al., 2021](#); [Calhoun et al., 2019](#); [Hsu and Yttri, 2020](#); [Marques et al., 2019](#)). This problem is more difficult for monkeys, although, even here, significant progress has been made ([Bain et al., 2021](#); [Bala et al., 2020](#); [Dunn et al., 2021](#); [Labuguen et al., 2021](#); [Marks et al., 2022](#); reviewed in [Hayden et al., 2022](#)).

In theory, species-general systems can achieve good performance with small numbers (hundreds or thousands) of hand-annotated sample images. In practice, however, such systems tend to be of limited functionality. That is, they may show brittle performance, and may tend to perform poorly in edge cases, which may wind up being quite common. In general, large and precisely annotated databases (ones with tens of thousands of images or more) may be needed as training sets to achieve robust performance. The monkey tracking in our monkey-specific system (*OpenMonkeyStudio*), for example, required over 100,000 annotated images, and performance continued to improve even at larger numbers of images in the training set ([Bala et al., 2020](#); [Yao et al., 2022](#)).

However, there is no currently publicly available database specifically for non-human apes, which in turn means that readily usable tracking solutions specific to apes do not exist. Although there is hope that models built on related species, such as humans and/or monkeys may generalize to apes, transfer methods remain a work in progress ([Sanakoyeu et al., 2020](#)). Like monkeys, apes are particularly challenging to track due to their homogeneous body texture and exponentially large number of pose configurations ([Yao et al., 2022](#)). We recently developed a novel system for tracking the pose of monkeys ([Bala et al., 2020](#); [Bala et al., 2021](#); [Yao et al., 2022](#)). A critical ingredient of this system was the collection of high-quality annotated images of monkeys, which were used as raw material for training the model. Indeed, the need for high-quality training datasets is a major barrier to progress for much of machine learning ([Deng et al., 2009](#)). Obtaining a database of annotated ape photographs is especially difficult due to apes' relative rarity in captive settings and due to the proprietary oversight common among primatologists.

The lack of such tracking systems represents a critical gap due to the importance of apes in science. The ape (*Hominoidea*) superfamily includes the great apes (among them, humans, *Hominidae* family) and the lesser apes, gibbons and siamangs (*Hylobatidae* family). These species, which represent humans' closest relatives in the animal kingdom, have complex social and foraging behavior, a high level of intelligence, and a behavioral repertoire characterized by flexibility and creativity ([Smuts et al., 2008](#); [Strier, 2016](#)). The ability to perform sophisticated video tracking of apes would bring great benefits to primatology and comparative psychology, as well as to related fields like anthropology and kinesiology ([Hayden et al., 2022](#)). Moreover, tracking systems could be deployed to improve ape welfare and to supplement *in-situ* conservation efforts ([Knaebe et al., 2022](#)).

Here we provide a dataset of annotated ape photographs, which we call *OpenApePose*. This dataset includes four species from the *Hominidae* family: bonobos, chimpanzees, gorillas, orangutans, and several species from the *Hylobatidae* family, pooled into two categories of gibbons and siamangs. This dataset consists primarily of photographs taken at zoos, and also includes images from online sources, including publicly available photographs and videos. Our database is designed to have a rich sampling of poses and backgrounds, as well as a range of image features. We provide high precision annotation of sixteen body landmarks. We show that tracking models built using this database do a good job tracking from a large sample of ape images, and do a better job than networks trained with monkey (*OpenMonkeyPose*, [Yao et al., 2022](#)) or human (*COCO*, [Lin et al.,](#)

2014) databases. We also show that tracking quality is comparable to these two databases tracking their own species (although performance lags slightly behind both). We believe this database will provide an important resource for future investigations of ape behavior.

Results

OpenApePose dataset

We collected several hundred thousand images of five species of apes: chimpanzee, bonobo, gorilla, orangutan, siamang, and a sixth category, including non-siamang gibbons (**Figure 1**). Images were collected from zoos, sanctuaries, and field sites. We also added the ape images from the *OpenMonkeyPose* dataset (16,984 images) to our new dataset, which we call *OpenApePose*. Combined, our final dataset has 71,868 annotated ape images. Our image set contains 11,685 bonobos (*Pan paniscus*), 18,010 chimpanzees (*Pan troglodytes*), 12,905 gorillas (*Gorilla gorilla*), 12,722 orangutans (*Pongo sp.*), and 9274 gibbons (genus *Hylobates* and *Nomascus*) and 7,272 siamangs (*Symphalangus syndactylus*, **Figure 2A**).

We manually sorted and cropped the images such that each cropped image contains the full body of at least one ape while minimizing repetitive poses to ensure a greater diversity of poses in the full dataset. We ensured that all cropped images have a resolution greater than or equal to 300×300 pixels. Next, we used a commercial annotation service (Hive AI) to manually annotate the 16 landmarks (we used the same system in Yao et al., 2022; See **Methods**). The 16 landmarks together comprise a *pose* (**Figure 2B**).

We used these landmarks to infer a bounding box, defined as the distance + 20% pixels between the farthest landmarks on the two axes. Our landmarks were: (1) nose, (2-3) left and right eye, (4) head, (5) neck, (6-7) left and right shoulder, (8-9) elbows, (10-11) wrists, (12) sacrum, that is, the center-point between the two hips, (13-14) knees, and (15-16) ankles. These are the same landmarks we used in our corresponding monkey dataset (Yao et al., 2022), although in that set we also included a landmark for the tip of the tail. (We don't include that here because apes don't have tails). Each data instance is made of: image, species, bounding box, and pose.

Our previous monkey-centered dataset was presented in the form of a *challenge* (Yao et al., 2022). Our ape dataset, by contrast, is presented solely as a resource. The annotations and all 71,868 images are available at <https://github.com/desai-nisarg/OpenApePose>.

Overview of OpenApePose dataset

To illustrate the range of poses in the OpenApePose dataset, we visualize the space spanned by its poses using Uniform Manifold Approximation and Projection (UMAP, McInnes et al., 2018, **Figure 3**). To obtain standard and meaningful spatial representations, we use normalized landmark coordinates based on image size—the x-coordinate normalized using image width and the y-coordinate normalized using the image height. We then center each pose to a reference root landmark (the sacrum), such that the normalized coordinate of each landmark is with respect to the sacrum landmark. We then create the UMAP visualizations by performing dimension

reduction using the

umap function in the

Python package (McInnes et al., 2018). We use the euclidean distance metric with

and



Figure 1

Sampling of annotated images in the *OpenApePose* dataset.

Thirty-two photographs chosen to illustrate the range of photographs available in our larger set, illustrating the variety in species, pose, and background. Each annotated photograph contains an annotation for sixteen different body landmarks (shown here with connecting lines).

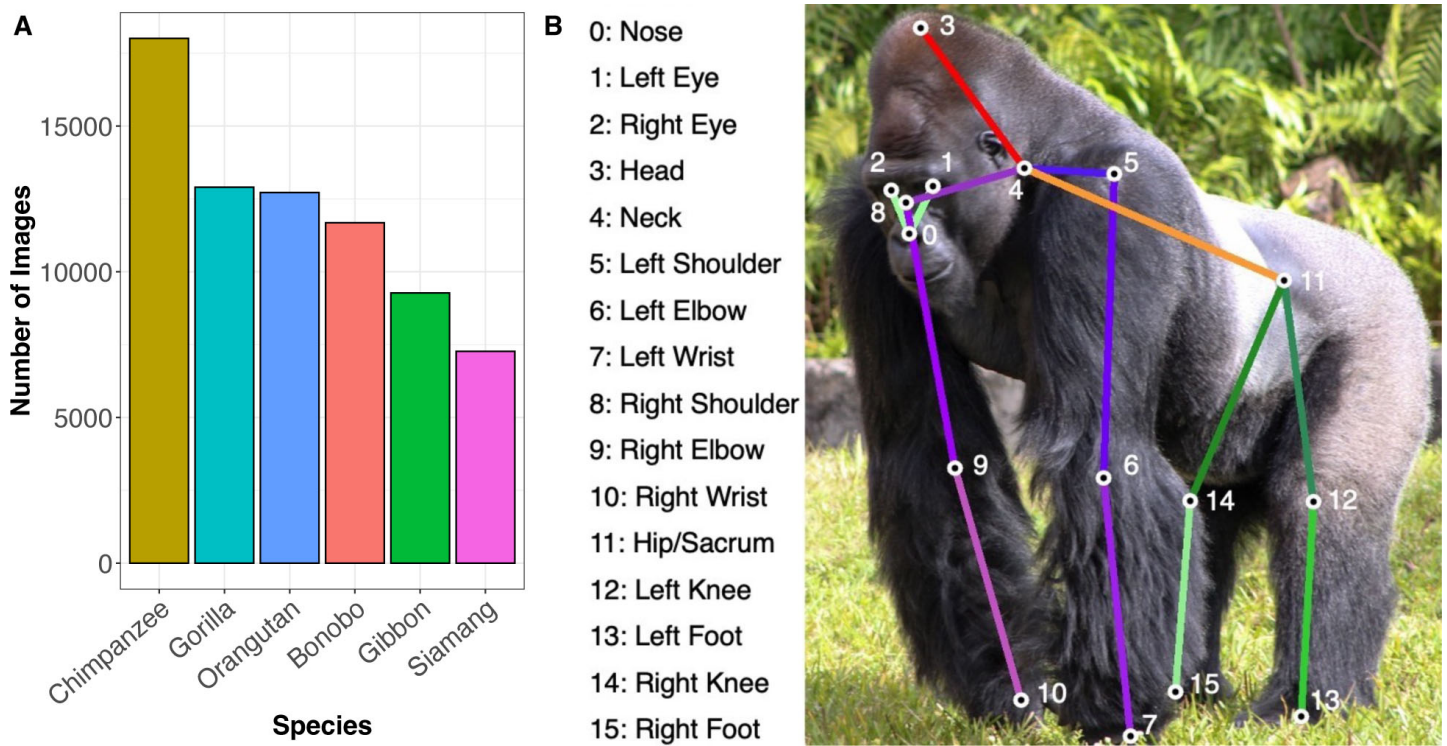


Figure 2

Properties of our OpenApePose database.

A. Number of annotated images per different species in the OpenApePose dataset. **B.** Illustration of our annotations. All 16 annotated points are indicated and labeled on a gorilla image drawn from the database.

, which allowed us a reasonable balance in combining similar poses and separating dissimilar ones.

We label the six different species in the database to visualize their distribution in the dimensions reduced using UMAP. We observe that the *Hylobatidae* family (gibbons and siamangs) form somewhat separate pose clusters from the *Hominidae* family (bonobos, chimpanzees, gorillas, and orangutans, **Figure 3**). These clusters likely reflect the differences in locomotion styles between these families, *Hylobatidae* being true brachiators, whereas *Hominidae* spend more time on moving on the ground. Of the *Hominidae*, the orangutans spend the most time in the trees like the *Hylobatidae*, and this is reflected in the overlap of their poses with the *Hylobatidae*.

Demonstrating the effectiveness of the OpenApePose dataset

We next performed an assessment of the OpenApePose dataset for pose estimation. To do this, we used a standard tracking deep net system HRNet-W48, which currently remains state-of-the-art for pose estimation (Sun et al., 2019). The deep high-resolution net (HRNet) architecture achieves superior performance as it works with high resolution pose representations from the get-go, as compared to conventional architectures that work with lower resolution representations and extrapolate to higher resolutions from low resolutions (*ibid.*). We previously showed that this system does a good job tracking monkeys with a monkey database (Yao et al., 2022).

We split the benchmark dataset into training (43,120 images, 60%), validation (14,374 images, 20%), and testing (14,374 images, 20%) datasets using the

function in the

Python library (Pedregosa et al. 2011).

We first investigated the ability of a model trained on the ape training set to accurately predict landmarks on apes from the test set (that is, a set that contains only images that were not used in training). To evaluate the performance of the HRNet-W48 models trained on this dataset, we used a standard approach of calculating *percent correct keypoints* (PCK) at a given threshold (here, 0.2, see **Methods**) and at a series of other thresholds (0.01-1, at 0.01 increments, **Figure 4A**). The PCK@0.2 for this model was 0.876 and the area under the curve of PCK at all thresholds (AUC) for this model was 0.897. We used a bootstrap procedure to estimate significance and compare the model performance across different datasets (see **Methods**). To assess significance, we calculated the AUCs of 100 random test subsets of 500 images each, sampled from the original held-out test set. We used the standard deviation of the AUCs as the error bars (**Figure 4B**), performed pairwise t-tests on mean AUCs, and used Bonferroni-adjusted p-values to test for significance.

For comparison, we used a model trained on the dataset consisting of 94,550 monkeys, split into training (56,694 images, 60%), validation (18,928 images, 20%), and testing (18,928 images, 20%) to predict apes (specifically, we used OpenMonkeyPose, Yao et al., 2022). (Note that the original OpenMonkeyPose dataset contained some apes; for fair cross-family comparison, we are using a version of OpenMonkeyPose with the apes removed; the 94,550 number above reflects the number of monkeys alone). The monkey dataset showed poorer performance when it comes to estimating landmarks on photos of apes. Specifically, at a threshold of 0.2, the PCK was 0.584, which is lower than the analogous value for OpenApePose (PCK@0.2 = 0.876, p-adjusted < 0.001). Likewise, the area under the curve was also substantially lower (0.743, compared to 0.897 for OpenApePose, p-adjusted < 0.001). In other words, for tracking apes, models trained on monkey images have some value, but they are not nearly as good as models trained on apes.

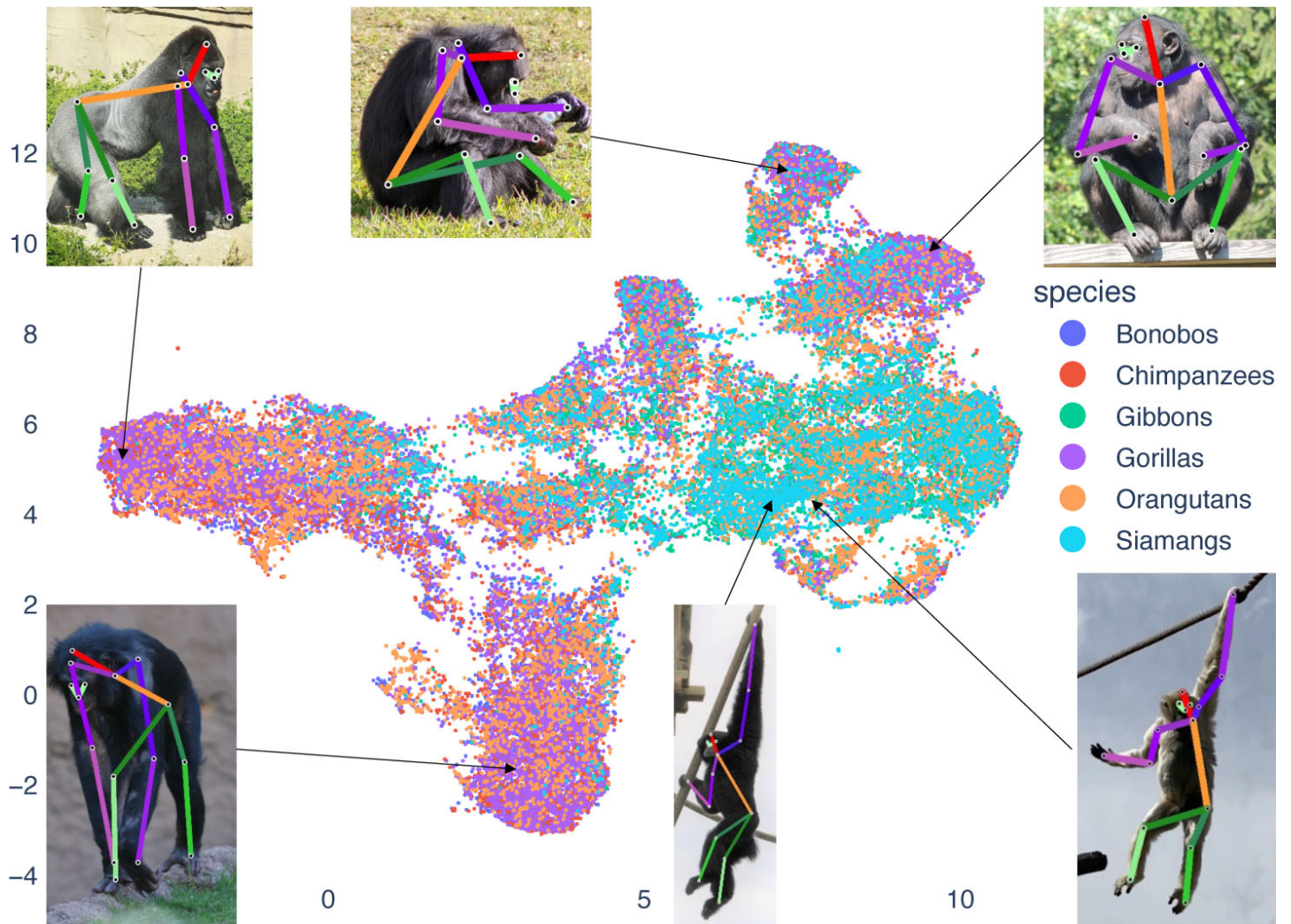


Figure 3

UMAP visualization of the distribution of poses with the species IDs labeled.

X- and Y- dimensions indicate positions in a UMAP space. Each dot indicates a single photograph/pose. Dot colors indicate species (see inscribed legend, right). We include, as insets, example poses, with an arrow pointing to their position in the UMAP plot.

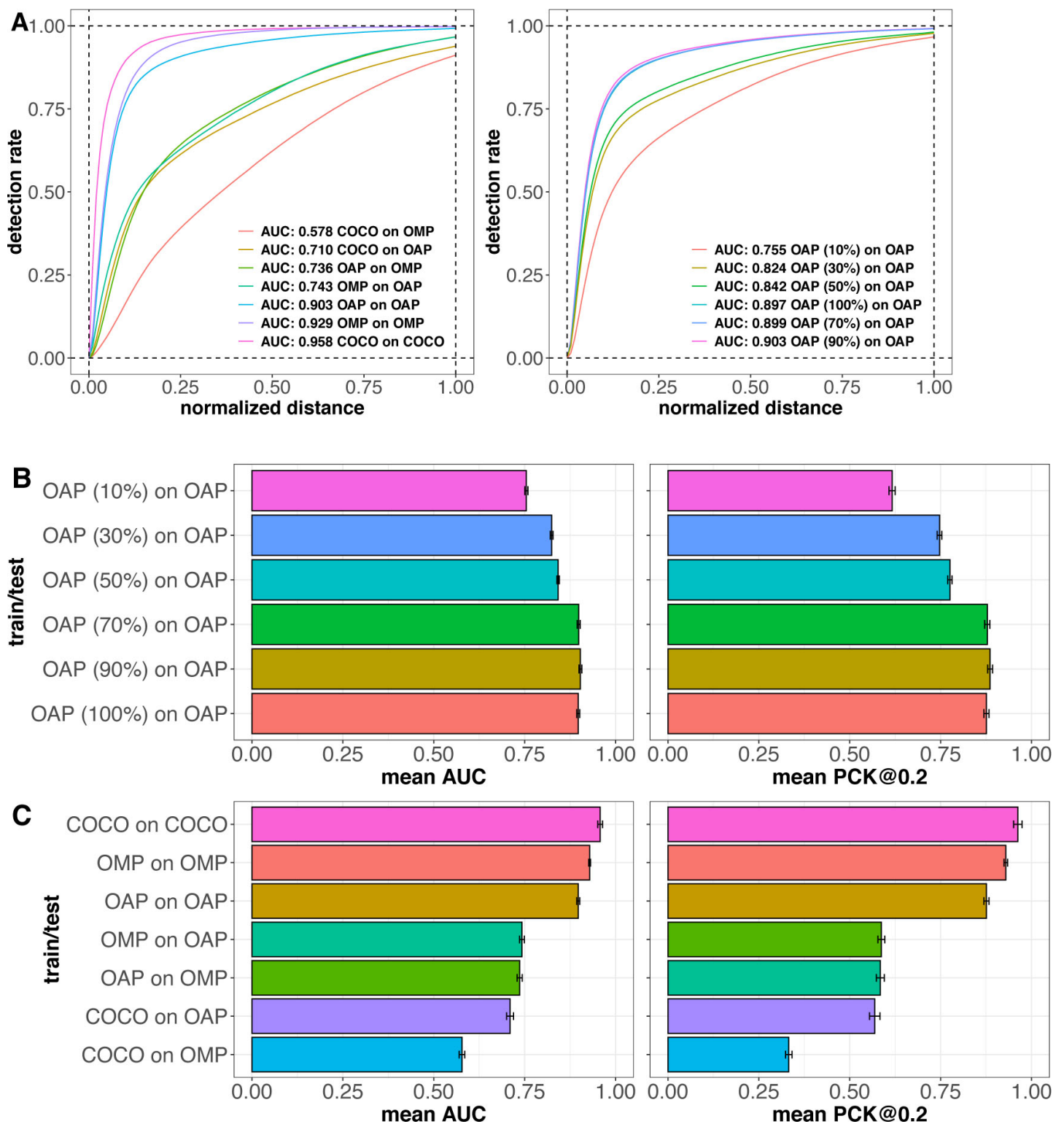


Figure 4

A. Keypoint detection performance of HRNet-W48 models measured using PCK values at different thresholds. **Left:** Models trained on the full training sets of COCO, OpenApePose (OAP), and OpenMonkeyPose (OMP), and tested on the same dataset, as well as across datasets. **Right:** Models trained on different sizes of the full OAP training set, and tested on the OAP testing set. **B.** Barplots showing the keypoint detection performance of state of the art (HRNet-W48) models as measured using percent keypoints correct at 0.2 (PCK@0.2) and area-under-the-curve (AUC) of the PCK curves at thresholds ranging from 0.01-1. Error bars: standard deviation of the performance metrics. Models are trained on different sizes of the full training set of OAP and tested on held-out OAP test sets. **C.** Same as 4B but: models are trained on full training sets of COCO, OAP, and OMP, and tested on the same dataset, as well as across datasets.

Comparison with human pose estimation

A long-term goal of primate pose estimation datasets such as OpenApePose and OpenMonkeyPose is to achieve performance comparable to that of human pose estimation. Hence, as a further comparison, we used a previously published standard model trained on the dataset consisting of 262,465 humans (*COCO*) to predict apes (Lin et al., 2014). This dataset showed poorer performance at predicting landmarks on apes than the model trained on the OAP dataset. Specifically, the PCK@0.2 value of 0.569 was lower than the PCK@0.2 value of 0.876 for OAP (p-adjusted < 0.001) and the AUC value of 0.710 was lower than the AUC value of 0.897 for OAP (p-adjusted < 0.001).

COCO was worse at pose estimation for apes than the OpenMonkeyPose dataset was (PCK@0.2: 0.569 vs. 0.584, p-adjusted < 0.001; and AUC: 0.710 vs. 0.743, p-adjusted < 0.001), despite the fact that it is a much larger dataset (262,465 vs. 56,694 images). Moreover, humans are, biologically speaking, apes, so one may expect the *COCO* dataset to have an advantage on ape tracking over a monkey dataset such as OMP. This does not appear to be the case. However, it is interesting to note that the *COCO* model predicts landmarks on apes better than it predicts landmarks on monkeys (PCK@0.2 values: 0.568 vs 0.332, p-adjusted < 0.001; AUC values: 0.710 vs 0.578, p-adjusted < 0.001). This advantage, at least, does recapitulate phylogeny.

While the OpenApePose-trained model predicted apes at an AUC value of 0.897, the OpenMonkeyPose dataset predicted monkeys at an AUC value of 0.929. These values are close, but significantly different (p-adjusted < 0.001). We surmise that the superior performance of OpenMonkeyPose dataset may be due to the diversity of species, and to its larger size. Finally, the model based on the *COCO* dataset predicted human poses even better still, at an AUC value of 0.956, than either the OMP or OAP within group predictions. This advantage presumably reflects, among other things, the larger size of the dataset.

How big does an ape tracking dataset need to be?

We next assessed the performance of our ape dataset at different sizes (Figure 4B). To do so, we used a decimation procedure in which we assessed the performance of the dataset after randomly removing different numbers of images. Specifically, we subsampled our OpenApePose dataset at a range of sizes (10%, 30%, 50%, 70%, and 90% of the full training set size). Note that our subsampling procedure was randomized to balance across different species. We then tested each of the resulting models on our independent test set.

We found a gradual increase in performance with training set size. Specifically, the performance at 30% was greater than the performance at 10% (PCK@0.2: 0.747 vs. 0.617 and AUC: 0.824 vs. 0.755). Likewise, the performance at 50% was greater than the performance at 30% (PCK@0.2: 0.776 vs. 0.747 and AUC: 0.842 vs. 0.824), performance at 70% was greater than the performance at 50% (PCK@0.2: 0.878 vs. 0.776 and AUC: 0.899 vs. 0.842), and the performance at 90% was comparable to 70% (PCK@0.2: 0.886 vs. 0.878, AUC: 0.903 vs. 0.899), although it too was significantly greater (p-adjusted < 0.001 for all comparisons above). However, the performance at 100% was not significantly greater than the performance at 70% (PCK@0.2: 0.876 vs. 0.878, and AUC: 0.897 vs. 0.899, p-adjusted > 0.9 for both). These results suggest that performance begins to saturate at around 70% size and that increasingly larger sets may not provide additional improvement in tracking and might lead to overfitting.

Interestingly, a similar pattern is observed when tracking monkey poses. While the Convolutional Pose Machines (CPM) models trained on different sizes of the OpenMonkeyPose training sets continue to show improvements as the training set size increases (See Figure 9A in Yao et al., 2022), the HRNet-W48 models show similar saturation beyond 80% training set size (Figure 9B in Yao et al., 2022), just like we observed in the OpenApePose models (see above). (Note that, for OpenMonkeyPose, the HRNet-W48 model performed better across the board, which is why we

prefer it to the CPM approach here). This difference between the two model classes points towards the arms race between dataset size and algorithmic development as the limiting factors for performance. Ultimately, for OpenApePose, future algorithmic developments may facilitate greater performance than increasing the dataset size beyond the number we offer here.

What is the hardest ape species to track?

Finally, we assessed the performance of the model on each species of ape separately. We regenerated the OpenApePose model six times, each time with all images of one of the six taxonomic groups removed. We then tested the models on the images of that group in the OAP test set. Note that this procedure has a second benefit, which is that it automatically ensures that any similar images (such as those collected in the same zoo enclosure or of the same individual) are excluded, and therefore reduces the chance of overfitting artifacts. (However, as we show below, doing this does not markedly reduce performance, suggesting that this type of overfitting is not a major issue in our analyses presented above).

We include a plot including the performance of each of these models on all different species in the supplementary materials ([Figure S1](#)). We also include in the plot the performance of the OpenMonkeyPose model on the species excluded from the OpenApePose dataset. We observe that the OpenApePose model with a specific species removed still performs better on that species than the OpenMonkeyPose model ([Figure S1](#)). Here, we include a plot with performance of the full OpenApePose model on different species, performance of the models with one species removed at a time on that species, and of the OpenMonkeyPose model without apes on each of the species ([Figure 5A-C](#)). Not surprisingly, we find that all the models excluding a species perform worse than the full model on the same species ([Figure 5AB](#); PCK@0.2 and AUC for **Bonobos**: 0.871 vs. 0.881 and 0.896 vs. 0.903; **Chimpanzees**: 0.754 vs. 0.882 and 0.836 vs. 0.902; **Gibbons**: 0.763 vs. 0.855 and 0.827 vs. 0.883; **Gorillas**: 0.869 vs. 0.893 and 0.896 vs. 0.908; **Orangutans**: 0.774 vs. 0.859 and 0.839 vs. 0.886; **Siamangs**: 0.797 vs. 0.869 and 0.848 vs. 0.889; p-adjusted < 0.001 for all comparisons). This result suggests that there is indeed some species-specific information in the model that aids in tracking, and raises the possibility that larger sets devoted to a single species may be superior to our more general multi-species dataset. At the same time, this finding highlights a major finding of this project - that, given current models, large tailored species-specific annotated sets are superior to large multi-species sets. In other words, current models have limited capacity of generalizing across species, even within taxonomic families.

Comparing the different species, we find that the species are all very close in performance ([Figure 5B](#)). Among these close values, the dataset missing gorillas was the most accurate - suggesting that gorillas are the least difficult to track, perhaps because their bodies are the least variable (PCK@0.2: 0.869; AUC: 0.896). Conversely, the dataset missing gibbons was the least accurate, suggesting that gibbons are the most difficult to track (PCK@0.2: 0.763; AUC: 0.827). This observation is consistent with our own intuitions at hand-annotating images - gibbons' habit of brachiation, combined with the variety of poses they exhibit, makes guessing their landmarks particularly tricky for human annotators as well. Overall, however, all ape species were relatively well tracked even when all members of their species were excluded from the dataset.

Note that models with one ape species removed still perform better at tracking the held-out species more accurately than the OpenMonkeyPose model on that species ([Figure 5BC](#); PCK@0.2 and AUC for **Bonobos**: 0.871 vs. 0.542 and 0.896 vs. 0.727; **Chimpanzees**: 0.754 vs. 0.688 and 0.836 vs. 0.803; **Gibbons**: 0.763 vs. 0.587 and 0.827 vs. 0.730; **Gorillas**: 0.869 vs. 0.564 and 0.896 vs. 0.744; **Orangutans**: 0.774 vs. 0.529 and 0.839 vs. 0.707; **Siamangs**: 0.797 vs. 0.556 and 0.848 vs. 0.711; p-adjusted < 0.001 for all comparisons). In other words, the close phylogenetic relationship between ape species does seem to bring about benefits in tracking.

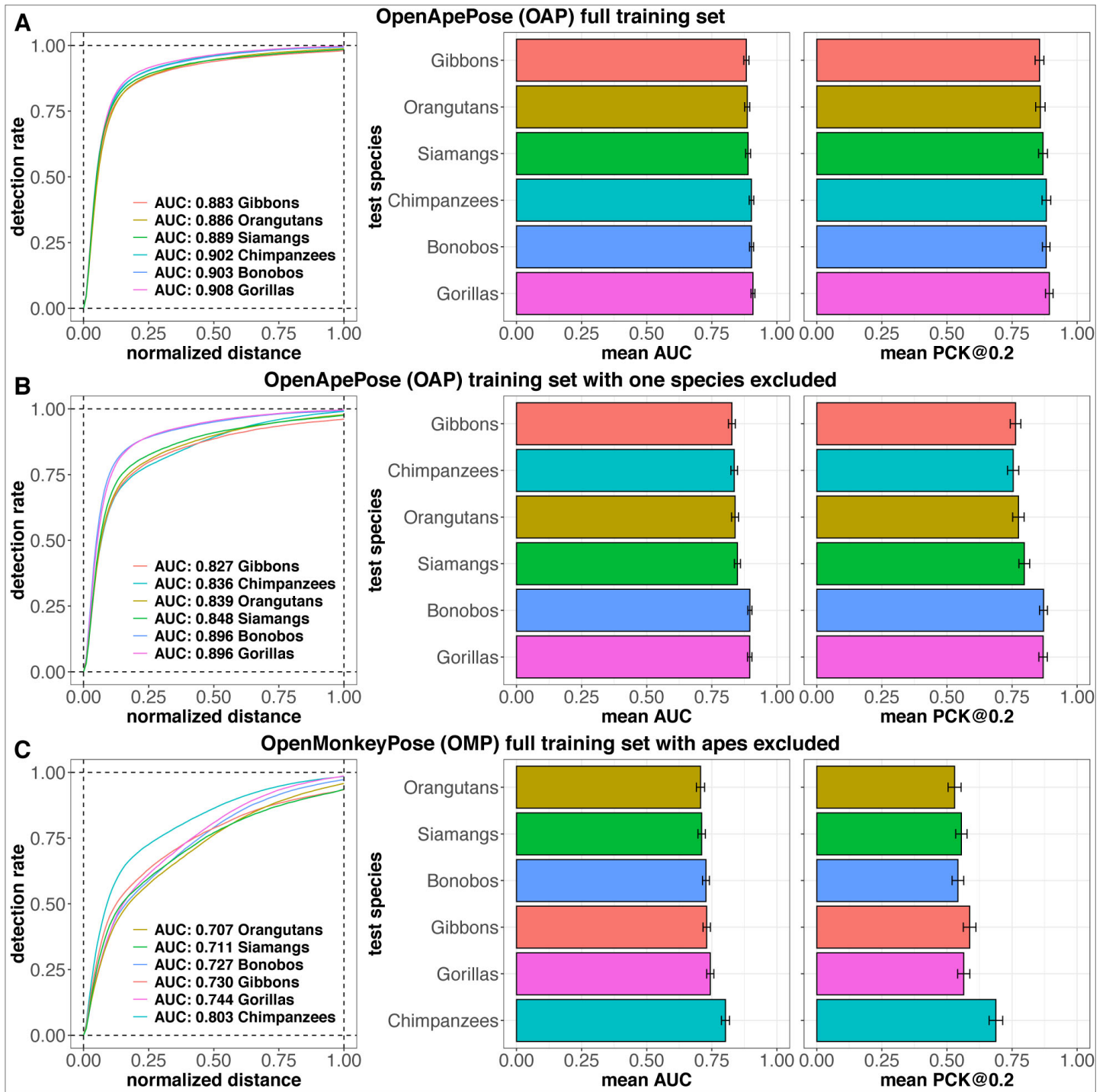


Figure 5

Keypoint detection performance of HRNet-W48 models tested on each species from the OpenApePose (OAP) test set and trained on (A) the full OAP training set, (B) the OAP training set with the corresponding species excluded, and (C) the full OpenMonkeyPose (OMP) dataset with apes excluded.

Left panel includes the probability of correct keypoint (PCK) values at different thresholds ranging from 0-1. **Middle panel** indicates the mean area under the PCK curve for each species. **Right panel** indicates the mean PCK values at a threshold of 0.2 for each species.

The ape superfamily is an especially charismatic clade, and one that has long been fascinating to both the lay public and to scientists. Here we present a new resource, a large (71,868 images) and fully annotated (16 landmarks) database of photographs of six species of non-human apes. These photographs were collected and curated with the goal of serving as a training set for machine vision learning models, especially ones designed to track apes in videos. As such, the apes in our dataset come in a range of poses; photographs are taken from a range of angles, and our photographs have a range of backgrounds. Our database can be found on our website (<https://github.com/desai-nisarg/OpenApePose>).

To test and validate our set, we made use of the HRNet architecture, specifically HRNet-W48. As opposed to architectures such as convolutional pose machines (Wei et al. 2016), hourglass (Newell et al. 2016), simple baselines (ResNet, Xiao et al. 2018), HRNet works with higher resolution feature representations, that facilitate better performance. In contrast, other systems, most famously DeepLabCut, uses ResNets, EfficientNets, and MobileNets V2 as backbones. Pose estimation studies often compare a variety of these architectures to test performance, but increasingly, studies find HRNet to outperform other architectures (Yu et al. 2021; Li et al. 2020). (Our own past work on monkey tracking finds this as well, Yao et al. 2022). Because our goal here is not to evaluate these systems, but rather to introduce our annotated database, we provide data only for the HRNet system.

With growing interest in animal detection, pose estimation and behavior classification (Bain et al. 2021; Sakib et al. 2020; Pereira et al. 2019; Mathis et al. 2021), researchers have leveraged advances in human pose estimation and have made several animal datasets publicly available. For example, there are existing datasets on tigers (n ~ 8000, Li et al., 2020), Cheetahs (n ~ 7500, Joska et al., 2021), horses (n ~ 8000, Mathis et al., 2021), dogs (n ~ 22,000, Briggs et al. 2020, Khosla et al., 2011), cows (n ~ 2000, Russello et al., 2021), five domestic animals (Cao et al., 2019), and fifty-four species of mammals (Yu et al., 2021), and there are large datasets containing millions of frames of rats enabling single and multi-animal 3D pose estimation and behavior tracking (Dunn et al., 2021; Marshall et al., 2021). Relative to these other datasets (with the exception of rat datasets), our ape dataset is much larger (n ~ 71,000). Moreover, our dataset contains multiple closely related species and a wide range of backgrounds and poses. Another major strength of our dataset is that it contains many different of unique individuals, which is rare as most of such datasets include only a few unique individuals.

We anticipate that the main benefit of our database will be for future researchers to develop algorithms that can perform tracking of apes in photos and videos, including videos collected in field sites. Relative to simpler animals like worms and mice, primates are highly complex and have a great deal more variety in their poses. As such, in the absence of better deep learning techniques, the best way to come up with generalizable models is to have large and variegated datasets for each animal type of interest. Our results here indicate that even monkeys and apes - which are in the same order and have superficially similar body shapes and movements - are sufficiently different that monkey photos don't work as well for ape pose tracking. Likewise, despite the remarkable growth of human tracking systems, these systems do not readily generalize to apes, despite our close phylogenetic similarity to them. While there is growing interest in leveraging human-tracking systems to develop better animal-tracking systems, such systems are still in their infancy (Sanakoyeu et al., 2019, Yu et al. 2021, Mathis et al. 2021, Arnkaern et al. 2022, Cao et al. 2019, Kleanthous et al. 2022, Bethell et al. 2022). At the same time, there are better and more usable general pose estimation systems for animals, such as DeepLabCut (Mathis et al. 2018), SLEAP (Pereira et al. 2022), LEAP (Pereira et al. 2019), DeepPoseKit (Graving et al. 2019) that allow pose estimation with small numbers (thousands) of images. These poses can be combined with downstream analysis algorithms and software tools such as MoSeq

(Wiltchko et al. 2020 [↗](#)), SimBA (Nilsson et al. 2020 [↗](#)), and B-SOiD (Hsu et al. 2021 [↗](#)) for behavior tracking. However, it is clear that such systems can benefit from much larger stimulus sets.

While our dataset is readily usable for training pose estimation and behavior tracking models, it has several limitations that could be addressed in the future. First, while we have attempted to include as many backgrounds, poses, and individuals as possible, our dataset is mostly dominated by images taken in captive settings at zoos and sanctuaries. This may not reflect the conditions in wild settings accurately, and may result in reduced performance for applications involving tracking apes in the wild from camera trap footage etc. Nevertheless, OpenApePose still remains the most diverse of currently available datasets. Future attempts at building such datasets should aim to include more images from the wild. Second, this dataset only enables 2D pose tracking as it does not include simultaneous multi-view images that are required for 3D pose estimation (Bala et al., 2020 [↗](#); Kearney et al., 2020 [↗](#); Dunn et al., 2021 [↗](#); Marshall et al., 2021 [↗](#)). Building a dataset that enables 3D pose estimation and has the strengths of OAP in terms of the diversity of individuals and poses would require building multiview camera setups outside of laboratories such as the one at Minnesota zoo by Yao et al., 2022 [↗](#). Third, while many images in our dataset include multiple individuals, we only have one individual labeled in each image. This limits, but does not eliminate, our ability to track multiple individuals simultaneously. Using OpenMMLab, we have had some success tracking multiple individuals using the OAP model. However, datasets with multiple individuals simultaneously will further facilitate multi-animal tracking. Lastly, our dataset does not contain high resolution tracking of finer features, such as face, hands etc. Indeed, many primatologists would be interested in systems that can track facial expression and fine hand movements (Hobaiter et al., 2014 [↗](#); Hobaiter et al., 2021 [↗](#)). Because we have made our image database public, it can be used as a starting point for those researchers seeking to customize to their research goals. Indeed, it may be possible to add hand and face expression annotations to our system to serve these purposes.

There are several important ethical reasons why apes cannot - and should not - serve as subjects in invasive neuroscientific experiments. That does not mean, however, that we cannot draw inferences about their psychology and cognition based on careful observation of their behavior. Indeed, analysis of behavior is an important tool in neuroscience (Niv, 2021 [↗](#); Krakauer et al., 2017 [↗](#)). In our previous work, we have argued for the virtues of primate tracking systems to work hand in hand with invasive neuroscience techniques to improve the reliability of neuroscientific data (Hayden et al., 2022 [↗](#)). However, we have also argued that tracking has another entirely different benefit - it has the potential ability to provide data of such high quality that it can, in some cases, serve to adjudicate between hypotheses that would otherwise require brain measures (Knaebe et al., 2022 [↗](#)). For this reason, tracking data has the potential to reduce the need for non-behavior neuroscientific tools and for invasive and/or stressful recording techniques. We are optimistic that better ape tracking systems will greatly expand the utility of apes in non-invasive studies of the mind and brain. We hope that our dataset will help advance such systems.

Methods

Data collection

The OpenApePose dataset consists of 71,868 photographs of apes. We collected images between August 1, 2021 to September 2022 from zoos, sanctuaries, and internet videos. Note that a subset of these images (16,984 images from the train, validation, and test sets combined) also appeared in the OpenMonkeyPose dataset (Yao et al., 2022 [↗](#)). The remainder are new here.

Zoos and sanctuaries

We obtained images of apes from several zoos. These include zoos in Atlanta, Chicago, Cincinnati, Columbus, Dallas, Denver, Detroit, Erie (Pennsylvania), Fort Worth, Houston, Indianapolis, Jacksonville, Kansas City, Madison, Memphis, Miami, Milwaukee, Minneapolis, Phoenix, Sacramento, San Diego, Saint Paul, San Francisco, Seattle, and Toronto, as well as sanctuaries including the Chimpanzee Conservation Center, Project Chimps, Chimp Haven, and the Ape Initiative (Des Moines). These zoo photographs were taken either by ourselves, our lab members, or by photographers hired on temporary contracts using TaskRabbit (<https://www.taskrabbit.com/>) to take pictures at these zoos. Additionally, several other independent individuals contributed images: Esmay Van Strien, Jeff Whitlock, Jennifer Williams, Jodi Carrigan, Katarzyna Krolik, Lori Ellis, Mary Pohlmann, and Max Block. All photographs were carefully screened for quality and variety of poses first by a specially trained technician and then by N.D.

Internet sources

We also obtained a smaller number of images from internet sources including Facebook, Instagram, and YouTube. From YouTube videos, we took screenshots of apes exhibiting diverse poses during different behaviors. Use of photographic images from these sources is protected by Fair Use Laws and has been expressly approved by the legal office at the University of Minnesota. Specifically, our use of the images satisfies four properties of principles of Fair Use. First, our usage is transformative (a crucial part of their value is in their annotations, which improve their value to scientists); second, they were published in a public forum (YouTube or on public websites); third, we are using a small percentage of the frames in the videos (at 24 fps, we are using at most 1/24 of the frames); fourth, our usage does not reduce the market value for the images, which are, after all, freely available.

Landmark Annotation

We initially obtained hundreds of thousands of images from these sources. The majority of these images (>75%) did not pass our quality checks. Specifically, they were either blurry or too small or were too similar to others or showed too much occlusion images. This process led to 52,946 images in total.

We used a commercial service (Hive AI) to manually annotate 16 landmarks in these images, a process similar to one we used previously (Yao et al., 2022). We use the same set of landmarks as we did in our complementary monkey dataset, with the exception of the tip of the tail (apes don't have tails). The landmarks we used are: (1) nose, (2,3) left and right eye, (4) crown of the head, (5) nape of the neck, (6,7) left and right shoulder, (8,9) left and right elbow, (10,11) left and right wrist, (12) sacrum, or center-point between the hips, (13,14) left and right knee, (15,16) left and right foot. An example image illustrating these annotations is shown in **Figure 2A**. We ensured that the annotations were accurate by visualizing 5 random samples of 100 images with the annotations overlaid on the images, for each batch of 10,000 images, resulting in a total of ~2,500 inspected images. Only one of the five batches showed errors, and we sent the batch back to Hive for correction. Ape images from OpenMonkeyPose were inspected as described in Yao et al., 2022. We converted the annotations in a JSON format that is consistent with our previous OpenMonkeyPose dataset, and similar to other common datasets such as COCO. More details on the annotations are on the GitHub page: <https://github.com/desai-nisarg/OpenApePose>.

Dataset Evaluation

To facilitate the evaluation of generalizability of the OpenApePose dataset, we split the full dataset into 3 sets: training (60%: 43,120 images), validation (20%: 14,374 images), and testing (20%: 14,374 images) using the

function in the

Python library (Pedregosa et al. 2011).

Model training

To train our models, we used the pipelines and tools available in the

Python library (Chen et al., 2019 [↗](#)).

includes a wide range of libraries for computer vision applications including, but not limited to, object detection, segmentation, action recognition, pose estimation etc. For our project, we used the

package in

(MMPose Contributors, 2020 [↗](#)).

supports a range of pose estimation datasets on humans as well as many other animals, and includes pretrained models from these datasets that could be tuned for specific needs. It also provides tools for training a variety of neural network architectures from scratch on existing or new datasets.

In our previous work (Yao et al., 2022 [↗](#)), we tested different top-down neural network architectures for training pose estimation models on our OpenMonkeyPose database (**Figure 9C** in Yao et al. 2022 [↗](#)). This included Convolutional Pose Machines (CPM), Hourglass, ResNet101, ResNet152, HRNet-W32, and HRNet-W48. We found that the best performing architecture was the deep high-resolution net, HRNet-W48 (**Table 2** in Yao et al., 2022 [↗](#)). As opposed to the conventional approaches where higher resolution representations are recovered from lower resolution representations, the deep high-resolution net architecture works with higher resolution representations during the whole learning process. This results in more accurate pose representations for human pose estimation as demonstrated in the original paper (Sun et al., 2019 [↗](#)), and also for primate pose estimation, as we observed for our monkey datasets (Yao et al., 2022 [↗](#); Bala et al., 2020 [↗](#)). HRNet-W48 currently remains the best performing architecture for pose estimation and hence, for this study, we train HRNet-W48 models for comparing the performance on our proposed dataset. We trained all models for 210 epochs.

Other datasets tested

We compared the performance of the HRNet-W48 model trained on our OpenApePose dataset with the performance of the pretrained HRNet-W48 model on the COCO dataset (Sun et al., 2019 [↗](#)), as well as of the HRNet-W48 model trained from scratch on our OpenMonkeyPose dataset (Yao et al., 2022 [↗](#)) with apes removed. The original OpenMonkeyPose dataset included 16,984 images of apes—10,223 in the training, 3378 in the validation, and 3383 in the testing set. Hence, for a fair comparison between HRNet-W48 models trained on monkeys vs. apes, we moved these ape images from the OpenMonkeyPose dataset to the OpenApePose dataset (we provide the annotations for the OpenApePose dataset with the ape images from OpenMonkeyPose included in OpenApePose, as well as separately to enable future replications and comparisons).

On these datasets we performed the following comparisons. First, we performed within-dataset performance comparisons. We compared the performance of OpenApePose in predicting the poses of apes, to the performance of OpenMonkeyPose in predicting the poses of monkeys. Second, we compare the performance of OpenApePose in predicting apes to the performance of the current state-of-the-art human pose estimation model (HRNet-W48 model trained on COCO human keypoint dataset, 2017). Third, we assess the importance of dataset size by systematically reducing

the OpenApePose training set size while keeping the proportion to the species constant. We train an HRNet-W48 model from scratch on training sets 10% (4,312 images), 30% (12,936 images), 50% (21,560 images), 70% (30,184 images), and 90% (38,808 images) of the size of the full training set of 43,120 images. Lastly, to assess if the models were overfitting on the species in the OpenApePose dataset over being generalizable to non-human apes, we train six separate HRNet-W48 models from scratch each with all images from one of the six species (bonobos, chimpanzees, gibbons, gorillas, orangutans, and siamangs) excluded from the training set. We test these models on the test set images of the species excluded from the training set and compare it with the performance of the OpenMonkeyPose model on that species.

Performance Metrics

To evaluate the performance of our models, we used two metrics: (i) the probability of correct keypoint (PCK) at a threshold of 0.2, and (ii) the area under the curve of PCK thresholds ranging from 0-1 in 0.01 increments.

The $PCK@ε$ is the PCK value at a given error threshold ($ε$), defined as:

$\frac{1}{16I} \sum_{i=1}^I \sum_{j=1}^{16} \delta\left(\frac{\|\hat{x}_{ij} - x_{ij}\|}{W} < ε\right)$, where I is the number of images, i indicates the i -th image instance, and j indicates the j -th joint, W is the width of the bounding box, and $\delta(\cdot)$ is the function that returns 1 for a true statement and 0 for a false statement. This formulation ensures that the error tolerance accounts for the size of the image via the size of the bounding box, e.g. for a bounding box that is 300 pixels wide, a $PCK@0.2$ value considers a prediction within $300 \times 0.2 = 60$ pixels, to be a correct prediction.

We calculate the $PCK@ε$ value for $ε$ ranging from 0 to 1 with 0.01 increments. We plot the $PCK@ε$ values for different $ε$ (normalized distances) and calculate the area under the curve to estimate the performance of the HRNet-W48 models.

Statistical significance testing

To perform statistical significance tests of differences in model performance for the aforementioned performance metrics, we take a bootstrap approach. We simulate 100 different test sets by randomly sampling 500 images without replacement, 100 times, from a test set of interest. We then calculate the performance metrics of $PCK@0.2$ and the area under the curve (AUC) of $PCK@ε$ vs. $ε$; for $ε \in [0, 1]$. This allows us to estimate the variation in the performance of the HRNet-W48 models across different test sets. We test the differences in performance using pairwise t-tests for different training and testing set combinations. We report the p-values adjusted for multiple comparisons using Bonferroni correction.

Acknowledgements

We thank the Hayden/Zimmermann lab for valuable discussions and help with taking photographs. We thank Kriti Rastogi and Muskan Ali for their help with ape image collection. We thank Estelle Reballand from Chimpanzee Conservation Center, Fred Rubio from Project Chimps, Adam Thompson from Zoo Atlanta, Reba Collins from Chimp Haven, and Amanda Epping and Jared Tagliatela from Ape Initiative for permissions to take photographs from these sanctuaries as well as contributing images for the dataset. This work was supported by NIH MH128177 (to JZ), P30 DA048742 (JZ, BH), MH125377 (BH), NSF 2024581 (JZ, BH) and a UMN AIRP award (JZ, BH) from the Digital Technologies Initiative (JZ, BH), from the Minnesota Institute of Robotics (JZ), and Emory National Primate Research Center, NIH Office of the Director (P51-OD011132).

Experiment 1	Experiment 2	Mean 1	Mean 2	difference	p.adjusted
OMP (no Apes) on OMP	OAP (100%) on OAP	0.929	0.897	0.031	0
OMP (no Apes) on OMP	OAP (90%) on OAP	0.929	0.903	0.026	0
OMP (no Apes) on OMP	OAP (70%) on OAP	0.929	0.899	0.03	0
OMP (no Apes) on OMP	OAP (50%) on OAP	0.929	0.842	0.087	0
OMP (no Apes) on OMP	OAP (30%) on OAP	0.929	0.824	0.105	0
OMP (no Apes) on OMP	OAP (10%) on OAP	0.929	0.755	0.174	0
OMP (no Apes) on OMP	OAP (100%) on OMP	0.929	0.736	0.193	0
OMP (no Apes) on OMP	OMP (no Apes) on OAP	0.929	0.743	0.186	0
OMP (no Apes) on OMP	COCO on OMP	0.929	0.578	0.351	0
OMP (no Apes) on OMP	COCO on OAP	0.929	0.71	0.219	0
OMP (no Apes) on OMP	COCO on COCO	0.929	0.958	-0.029	0
OAP (100%) on OAP	OAP (90%) on OAP	0.897	0.903	-0.006	0
OAP (100%) on OAP	OAP (70%) on OAP	0.897	0.899	-0.001	1**

Table S1

Pairwise t-tests comparing the AUC values for different training and testing set combinations. AUCs were obtained across 100 random samples of 500 images from the test sets. P-values adjusted for multiple comparisons using Bonferroni correction. ** not significant

OAP (100%) on OAP	OAP (50%) on OAP	0.897	0.842	0.055	0
OAP (100%) on OAP	OAP (30%) on OAP	0.897	0.824	0.073	0
OAP (100%) on OAP	OAP (10%) on OAP	0.897	0.755	0.143	0
OAP (100%) on OAP	OAP (100%) on OMP	0.897	0.736	0.161	0
OAP (100%) on OAP	OMP (no Apes) on OAP	0.897	0.743	0.155	0
OAP (100%) on OAP	COCO on OMP	0.897	0.578	0.32	0
OAP (100%) on OAP	COCO on OAP	0.897	0.71	0.188	0
OAP (100%) on OAP	COCO on COCO	0.897	0.958	-0.06	0
OAP (90%) on OAP	OAP (70%) on OAP	0.903	0.899	0.005	0
OAP (90%) on OAP	OAP (50%) on OAP	0.903	0.842	0.061	0
OAP (90%) on OAP	OAP (30%) on OAP	0.903	0.824	0.079	0
OAP (90%) on OAP	OAP (10%) on OAP	0.903	0.755	0.149	0
OAP (90%) on OAP	OAP (100%) on OMP	0.903	0.736	0.167	0
OAP (90%) on OAP	OMP (no Apes) on OAP	0.903	0.743	0.161	0
OAP (90%) on OAP	COCO on OMP	0.903	0.578	0.326	0
OAP (90%) on OAP	COCO on OAP	0.903	0.71	0.193	0
OAP (90%) on OAP	COCO on COCO	0.903	0.958	-0.054	0
OAP (70%) on OAP	OAP (50%) on OAP	0.899	0.842	0.056	0
OAP (70%) on OAP	OAP (30%) on OAP	0.899	0.824	0.074	0
OAP (70%) on OAP	OAP (10%) on OAP	0.899	0.755	0.144	0
OAP (70%) on OAP	OAP (100%) on OMP	0.899	0.736	0.162	0
OAP (70%) on OAP	OMP (no Apes) on	0.899	0.743	0.156	0

Table S1 (continued)

	OAP				
OAP (70%) on OAP	COCO on OMP	0.899	0.578	0.321	0
OAP (70%) on OAP	COCO on OAP	0.899	0.71	0.189	0
OAP (70%) on OAP	COCO on COCO	0.899	0.958	-0.059	0
OAP (50%) on OAP	OAP (30%) on OAP	0.842	0.824	0.018	0
OAP (50%) on OAP	OAP (10%) on OAP	0.842	0.755	0.087	0
OAP (50%) on OAP	OAP (100%) on OMP	0.842	0.736	0.106	0
OAP (50%) on OAP	OMP (no Apes) on OAP	0.842	0.743	0.1	0
OAP (50%) on OAP	COCO on OMP	0.842	0.578	0.264	0
OAP (50%) on OAP	COCO on OAP	0.842	0.71	0.132	0
OAP (50%) on OAP	COCO on COCO	0.842	0.958	-0.115	0
OAP (30%) on OAP	OAP (10%) on OAP	0.824	0.755	0.07	0
OAP (30%) on OAP	OAP (100%) on OMP	0.824	0.736	0.088	0
OAP (30%) on OAP	OMP (no Apes) on OAP	0.824	0.743	0.082	0
OAP (30%) on OAP	COCO on OMP	0.824	0.578	0.247	0
OAP (30%) on OAP	COCO on OAP	0.824	0.71	0.114	0
OAP (30%) on OAP	COCO on COCO	0.824	0.958	-0.133	0
OAP (10%) on OAP	OAP (100%) on OMP	0.755	0.736	0.019	0
OAP (10%) on OAP	OMP (no Apes) on OAP	0.755	0.743	0.012	0
OAP (10%) on OAP	COCO on OMP	0.755	0.578	0.177	0
OAP (10%) on OAP	COCO on OAP	0.755	0.71	0.045	0

Table S1 (continued)

OAP (10%) on OAP	COCO on COCO	0.755	0.958	-0.203	0
OAP (100%) on OMP	OMP (no Apes) on OAP	0.736	0.743	-0.006	0
OAP (100%) on OMP	COCO on OMP	0.736	0.578	0.159	0
OAP (100%) on OMP	COCO on OAP	0.736	0.71	0.026	0
OAP (100%) on OMP	COCO on COCO	0.736	0.958	-0.221	0
OMP (no Apes) on OAP	COCO on OMP	0.743	0.578	0.165	0
OMP (no Apes) on OAP	COCO on OAP	0.743	0.71	0.033	0
OMP (no Apes) on OAP	COCO on COCO	0.743	0.958	-0.215	0
COCO on OMP	COCO on OAP	0.578	0.71	-0.132	0
COCO on OMP	COCO on COCO	0.578	0.958	-0.38	0
COCO on OAP	COCO on COCO	0.71	0.958	-0.248	0

Table S1 (continued)

Experiment 1	Experiment 2	Mean 1	Mean 2	difference	p.adjusted
OMP (no Apes) on OMP	OAP (100%) on OAP	0.929	0.876	0.053	0
OMP (no Apes) on OMP	OAP (90%) on OAP	0.929	0.886	0.043	0
OMP (no Apes) on OMP	OAP (70%) on OAP	0.929	0.878	0.051	0

Table S2

Pairwise t-tests comparing the PCK@0.2 values for different training and testing set combinations. PCK@0.2 values were obtained across 100 random samples of 500 images from the test sets. P-values adjusted for multiple comparisons using Bonferroni correction. ** not significant

OMP (no Apes) on OMP	OAP (50%) on OAP	0.929	0.776	0.154	0
OMP (no Apes) on OMP	OAP (30%) on OAP	0.929	0.747	0.182	0
OMP (no Apes) on OMP	OAP (10%) on OAP	0.929	0.617	0.313	0
OMP (no Apes) on OMP	OAP (100%) on OMP	0.929	0.587	0.342	0
OMP (no Apes) on OMP	OMP (no Apes) on OAP	0.929	0.584	0.345	0
OMP (no Apes) on OMP	COCO on OMP	0.929	0.332	0.597	0
OMP (no Apes) on OMP	COCO on OAP	0.929	0.569	0.361	0
OMP (no Apes) on OMP	COCO on COCO	0.929	0.962	-0.033	0
OAP (100%) on OAP	OAP (90%) on OAP	0.876	0.886	-0.01	0
OAP (100%) on OAP	OAP (70%) on OAP	0.876	0.878	-0.002	1**
OAP (100%) on OAP	OAP (50%) on OAP	0.876	0.776	0.1	0
OAP (100%) on OAP	OAP (30%) on OAP	0.876	0.747	0.129	0
OAP (100%) on OAP	OAP (10%) on OAP	0.876	0.617	0.26	0
OAP (100%) on OAP	OAP (100%) on OMP	0.876	0.587	0.289	0
OAP (100%) on OAP	OMP (no Apes) on OAP	0.876	0.584	0.292	0
OAP (100%) on OAP	COCO on OMP	0.876	0.332	0.544	0
OAP (100%) on OAP	COCO on OAP	0.876	0.569	0.307	0
OAP (100%) on OAP	COCO on COCO	0.876	0.962	-0.086	0
OAP (90%) on OAP	OAP (70%) on OAP	0.886	0.878	0.008	0

Table S2 (continued)

OAP (90%) on OAP	OAP (50%) on OAP	0.886	0.776	0.11	0
OAP (90%) on OAP	OAP (30%) on OAP	0.886	0.747	0.139	0
OAP (90%) on OAP	OAP (10%) on OAP	0.886	0.617	0.27	0
OAP (90%) on OAP	OAP (100%) on OMP	0.886	0.587	0.299	0
OAP (90%) on OAP	OMP (no Apes) on OAP	0.886	0.584	0.302	0
OAP (90%) on OAP	COCO on OMP	0.886	0.332	0.554	0
OAP (90%) on OAP	COCO on OAP	0.886	0.569	0.317	0
OAP (90%) on OAP	COCO on COCO	0.886	0.962	-0.076	0
OAP (70%) on OAP	OAP (50%) on OAP	0.878	0.776	0.103	0
OAP (70%) on OAP	OAP (30%) on OAP	0.878	0.747	0.131	0
OAP (70%) on OAP	OAP (10%) on OAP	0.878	0.617	0.262	0
OAP (70%) on OAP	OAP (100%) on OMP	0.878	0.587	0.291	0
OAP (70%) on OAP	OMP (no Apes) on OAP	0.878	0.584	0.294	0
OAP (70%) on OAP	COCO on OMP	0.878	0.332	0.546	0
OAP (70%) on OAP	COCO on OAP	0.878	0.569	0.31	0
OAP (70%) on OAP	COCO on COCO	0.878	0.962	-0.084	0
OAP (50%) on OAP	OAP (30%) on OAP	0.776	0.747	0.029	0
OAP (50%) on OAP	OAP (10%) on OAP	0.776	0.617	0.159	0
OAP (50%) on OAP	OAP (100%) on OMP	0.776	0.587	0.189	0
OAP (50%) on OAP	OMP (no Apes) on OAP	0.776	0.584	0.191	0
OAP (50%) on OAP	COCO on OMP	0.776	0.332	0.443	0

Table S2 (continued)

OAP (50%) on OAP	COCO on OAP	0.776	0.569	0.207	0
OAP (50%) on OAP	COCO on COCO	0.776	0.962	-0.187	0
OAP (30%) on OAP	OAP (10%) on OAP	0.747	0.617	0.131	0
OAP (30%) on OAP	OAP (100%) on OMP	0.747	0.587	0.16	0
OAP (30%) on OAP	OMP (no Apes) on OAP	0.747	0.584	0.163	0
OAP (30%) on OAP	COCO on OMP	0.747	0.332	0.415	0
OAP (30%) on OAP	COCO on OAP	0.747	0.569	0.178	0
OAP (30%) on OAP	COCO on COCO	0.747	0.962	-0.215	0
OAP (10%) on OAP	OAP (100%) on OMP	0.617	0.587	0.03	0
OAP (10%) on OAP	OMP (no Apes) on OAP	0.617	0.584	0.032	0
OAP (10%) on OAP	COCO on OMP	0.617	0.332	0.284	0
OAP (10%) on OAP	COCO on OAP	0.617	0.569	0.048	0
OAP (10%) on OAP	COCO on COCO	0.617	0.962	-0.346	0
OAP (100%) on OMP	OMP (no Apes) on OAP	0.587	0.584	0.003	1**
OAP (100%) on OMP	COCO on OMP	0.587	0.332	0.255	0
OAP (100%) on OMP	COCO on OAP	0.587	0.569	0.018	0
OAP (100%) on OMP	COCO on COCO	0.587	0.962	-0.375	0
OMP (no Apes) on OAP	COCO on OMP	0.584	0.332	0.252	0
OMP (no Apes) on OAP	COCO on OAP	0.584	0.569	0.016	0
OMP (no Apes) on OAP	COCO on COCO	0.584	0.962	-0.378	0

Table S2 (continued)

COCO on OMP	COCO on OAP	0.332	0.569	-0.236	0
COCO on OMP	COCO on COCO	0.332	0.962	-0.63	0
COCO on OAP	COCO on COCO	0.569	0.962	-0.394	0

Table S2 (continued)

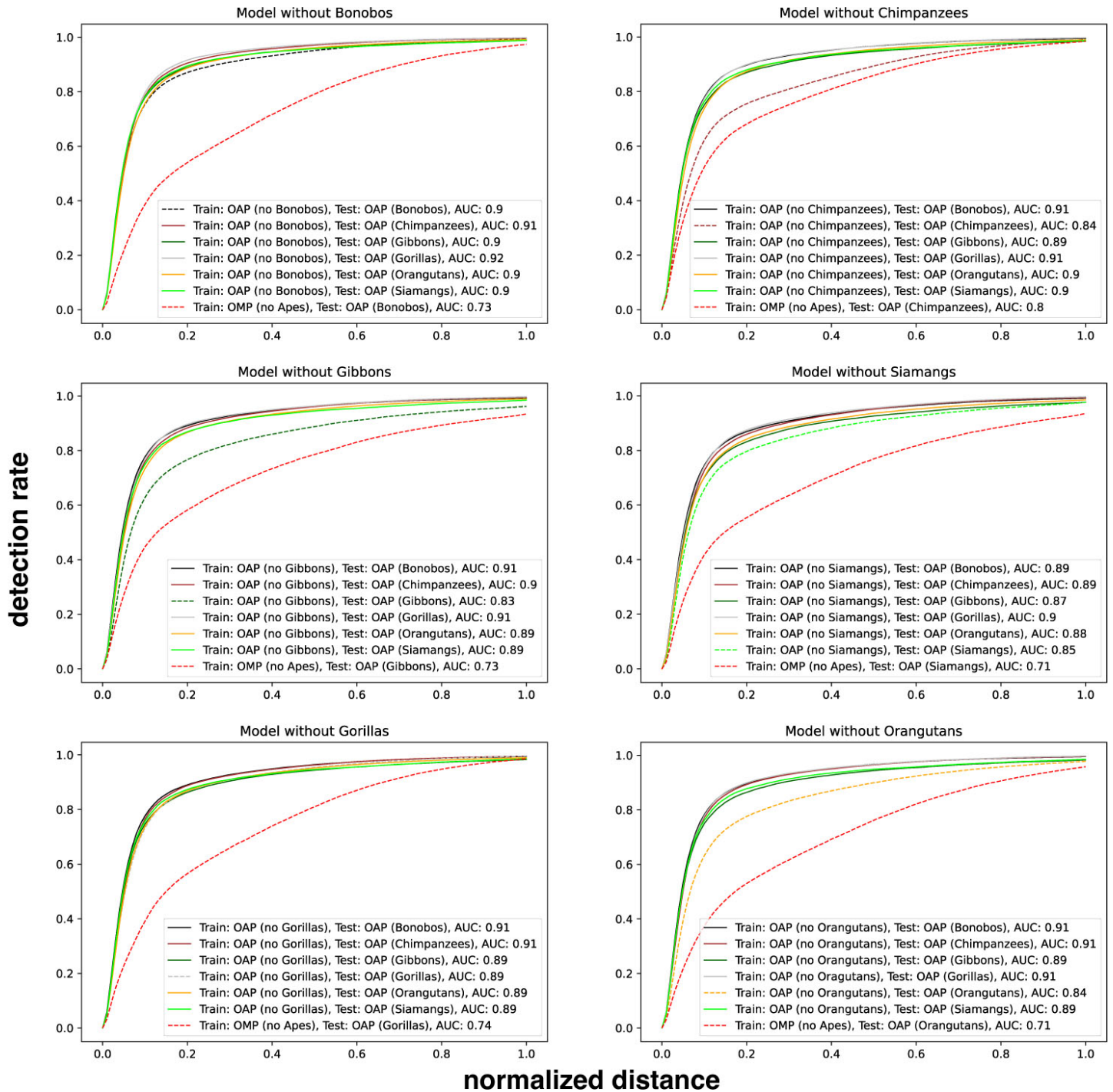


Figure S1

The probability of correct keypoint (PCK) values (y-axis) at different thresholds ranging from 0-1 (x-axis) of HRNet-W48 models tested on each species from the OpenApePose (OAP) test set and trained on the OAP training set with the corresponding species excluded.

Dotted lines indicate the performance on the species excluded from training in the case of OAP and the performance of the OpenMonkeyPose model trained on monkeys on the excluded species.

References

- Arnkærn B., Schoeler S., Ullah M., Cheikh F. A. (2022) **Deep learning-based multiple animal pose estimation** *Electronic Imaging* **34**:1–6
- Azab H., Hayden B. Y. (2018) **Correlates of economic decisions in the dorsal and subgenual anterior cingulate cortices** *European Journal of Neuroscience* **47**:979–993
- Bain M., Nagrani A., Schofield D., Berdugo S., Bessa J., Owen J., Zisserman A. (2021) **Automated audiovisual behavior recognition in wild primates** *Science advances* **7**
- Bala P. C., Eisenreich B. R., Yoo S. B. M., Hayden B. Y., Park H. S., Zimmermann J. (2020) **Automated markerless pose estimation in freely moving macaques with OpenMonkeyStudio** *Nature communications* **11**:1–12
- Bala P. C., Zimmermann J., Park H. S., Hayden B. Y. (2021) **Self-supervised Secondary Landmark Detection via 3D Representation Learning** *arXiv preprint arXiv:2110.00543*
- Bethell E. J., Khan W., Hussain A. (2022) **A deep transfer learning model for head pose estimation in rhesus macaques during cognitive tasks: Towards a nonrestraint noninvasive 3Rs approach** *Applied Animal Behaviour Science* **255**
- Blanchard T. C., Wolfe L. S., Vlaev I., Winston J. S., Hayden B. Y. (2014) **Biases in preferences for sequences of outcomes in monkeys** *Cognition* **130**:289–299
- Blanchard T. C., Piantadosi S. T., Hayden B. Y. (2018) **Robust mixture modeling reveals category-free selectivity in reward region neuronal ensembles** *Journal of neurophysiology* **119**:1305–1318
- Bohnslav J. P., Wimalasena N. K., Clausing K. J., Dai Y. Y., Yarmolinsky D. A., Cruz T., Harvey C. D. (2021) **DeepEthogram, a machine learning pipeline for supervised behavior classification from raw pixels** *Elife* **10**
- Biggs B., Boyne O., Charles J., Fitzgibbon A., Cipolla R. (2020) **Who left the dogs out? 3d animal reconstruction with expectation maximization in the loop** :195–211
- Calhoun A. J., Murthy M. (2017) **Quantifying behavior to solve sensorimotor transformations: advances from worms and flies** *Current opinion in neurobiology* **46**:90–98
- Calhoun A. J., Pillow J. W., Murthy M. (2019) **Unsupervised identification of the internal states that shape natural behavior** *Nature neuroscience* **22**:2040–2049
- Cao J., Tang H., Fang H. S., Shen X., Lu C., Tai Y. W. (2019) **Cross-domain adaptation for animal pose estimation** :9498–9507
- Chen Kai, Jiaqi Wang, Jiangmiao Pang, Yuhang Cao, Yu Xiong, Xiaoxiao Li, Shuyang Sun (2019) **MMDetection: Open mmlab detection toolbox and benchmark** *arXiv preprint arXiv:1906.07155*
- Cisek P., Hayden B. Y. (2022) **Neuroscience needs evolution** *Philosophical Transactions of the Royal Society B* **377**

- Deng Jia, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, Li Fei-Fei (2009) **Imagenet: A large-scale hierarchical image database** :248–255
- Dunn T. W., Marshall J. D., Severson K. S., Aldarondo D. E., Hildebrand D. G., Chettih S. N., Ölveczky B. P. (2021) **Geometric deep learning enables 3D kinematic profiling across species and environments** *Nature methods* **18**:564–573
- Ebitz R. B., Hayden B. Y. (2021) **The population doctrine in cognitive neuroscience** *Neuron* **109**:3055–3068
- Ebitz R. B., Sleezer B. J., Jedema H. P., Bradberry C. W., Hayden B. Y. (2019) **Tonic exploration governs both flexibility and lapses** *PLoS computational biology* **15**
- Farashahi S., Donahue C. H., Hayden B. Y., Lee D., Soltani A. (2019) **Flexible combination of reward information across primates** *Nature human behaviour* **3**:1215–1224
- Fine J. M., Hayden B. Y. (2022) **The whole prefrontal cortex is premotor cortex** *Philosophical Transactions of the Royal Society B* **377**
- Hayden B. Y., Park H. S., Zimmermann J. (2022) **Automated pose estimation in primates** *American journal of primatology* **84**
- Hayden B. Y., Moreno-Bote R. (2018) **A neuronal theory of sequential economic choice** *Brain and Neuroscience Advances* **2**
- Hayden B. Y., Niv Y. (2021) **The case against economic values in the orbitofrontal cortex (or anywhere else in the brain)** *Behavioral Neuroscience* **135**
- Heilbronner S. R., Hayden B. Y. (2016) **Dorsal anterior cingulate cortex: a bottom-up view** *Annual review of neuroscience* **39**
- Hobaiter Catherine, Badihi Gal, Daly Gabriela Bezerra de Melo, Eleuteri Vesta, Graham Kirsty Emma, Grund Charlotte, Henderson Matthew, Rodrigues Evelina Daniela, Safryghin Alexandra, Soldati Adrian, Wiltshire Charlotte (2021) **The Great Ape Dictionary video database (1.0.0) [Data set]** <https://doi.org/10.5281/zenodo.5600472>
- Hobaiter C., Byrne R. W. (2014) **The meanings of chimpanzee gestures** *Current Biology* **24**:1596–1600
- Hsu A. I., Yttri E. A. (2021) **B-SOiD, an open-source unsupervised algorithm for identification and fast prediction of behaviors** *Nature communications* **12**:1–13
- Joska D., Clark L., Muramatsu N., Jericevich R., Nicolls F., Mathis A., Patel A. (2021) **AcinoSet: a 3D pose estimation dataset and baseline models for Cheetahs in the wild** :13901–13908
- Kearney S., Li W., Parsons M., Kim K. I., Cosker D. (2020) **Rgbd-dog: Predicting canine pose from rgbd sensors** :8336–8345
- Khosla A., Jayadevaprakash N., Yao B., Li F. F. (2011) **Novel dataset for fine-grained image categorization: Stanford dogs**
- Kleanthous N., Hussain A., Khan W., Sneddon J., Liatsis P. (2022) **Deep transfer learning in sheep activity recognition using accelerometer data** *Expert Systems with Applications* **207**

- Knaebe B., Weiss C. C., Zimmermann J., Hayden B. Y. (2022) **The promise of behavioral tracking systems for advancing primate animal welfare** *Animals* **12**
- Krakauer J. W., Ghazanfar A. A., Gomez-Marin A., MacIver M. A., Poeppel D. (2017) **Neuroscience needs behavior: correcting a reductionist bias** *Neuron* **93**:480–490
- Labuguen R., Matsumoto J., Negrete S. B., Nishimaru H., Nishijo H., Takada M., Shibata T. (2021) **MacaquePose: A novel “in the wild” macaque monkey pose dataset for markerless motion capture** *Frontiers in behavioral neuroscience* **14**
- Li S., Li J., Tang H., Qian R., Lin W. (2019) **ATRW: a benchmark for Amur tiger re-identification in the wild** *arXiv preprint arXiv:1906.05586*
- Lin Tsung-Yi, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, Zitnick C. Lawrence (2014) **Microsoft coco: Common objects in context** :740–755
- Maisson D. J. N., Cash-Padgett T. V., Wang M. Z., Hayden B. Y., Heilbronner S. R., Zimmermann J. (2021) **Choice-relevant information transformation along a ventrodorsal axis in the medial prefrontal cortex** *Nature communications* **12**:1–14
- Marks M., Jin Q., Sturman O., von Ziegler L., Kollmorgen S., von der Behrens W., 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**:331–340
- Marques J. C., Li M., Schaak D., Robson D. N., Li J. M. (2019) **Internal state dynamics shape brainwide activity and foraging behaviour** *Nature* **577**:239–243 <https://doi.org/10.1038/s41586-019-1858-z>
- Marshall J. D., Klibaite U., Gellis A., Aldarondo D. E., Ölveczky B. P., Dunn T. W. (2021) **The pair-r24m dataset for multi-animal 3d pose estimation**
- Marshall J. D., Li T., Wu J. H., Dunn T. W. (2022) **Leaving flatland: Advances in 3D behavioral measurement** *Current Opinion in Neurobiology* **73**
- Mathis A., Mamidanna P., Cury K. M., Abe T., Murthy V. N., Mathis M. W., Bethge M. (2018) **DeepLabCut: markerless pose estimation of user-defined body parts with deep learning** *Nature neuroscience* **21**:1281–1289
- Mathis M. W., Mathis A. (2020) **Deep learning tools for the measurement of animal behavior in neuroscience** *Current opinion in neurobiology* **60**:1–11
- Mathis A., Biasi T., Schneider S., Yuksekogonul M., Rogers B., Bethge M., Mathis M. W. (2021) **Pretraining boosts out-of-domain robustness for pose estimation** :1859–1868
- McInnes L., Healy J., Melville J. (2018) **Umap: Uniform manifold approximation and projection for dimension reduction** *arXiv preprint arXiv:1802.03426*
- MMPose Contributors (2020) **OpenMMLab Pose Estimation Toolbox and Benchmark**
- Nilsson S. R., Goodwin N. L., Choong J. J., Hwang S., Wright H. R., Norville Z. C., Golden S. A. (2020) **Simple Behavioral Analysis (SimBA)—an open source toolkit for computer classification of complex social behaviors in experimental animals** *BioRxiv*

- Niv Y. (2021) **The primacy of behavioral research for understanding the brain** *Behavioral Neuroscience*
- Pedregosa Fabian, Varoquaux Gaël, Gramfort Alexandre, Michel Vincent, Thirion Bertrand, Grisel Olivier, Blondel Mathieu (1970) **Scikit-learn: Machine learning in Python** *the Journal of Machine Learning Research* **12**:2825–2830
- Pereira T. D., Aldarondo D. E., Willmore L., Kislin M., Wang S. S. H., Murthy M., Shaevitz J. W. (2019) **Fast animal pose estimation using deep neural networks** *Nature methods* **16**:117–125
- Pereira T. D., Shaevitz J. W., Murthy M. (2020) **Quantifying behavior to understand the brain** *Nature neuroscience* **23**:1537–1549
- Pereira T. D., Tabris N., Matsliah A., Turner D. M., Li J., Ravindranath S., Murthy M. (2022) **SLEAP: A deep learning system for multi-animal pose tracking** *Nature methods* **19**:486–495
- Russello H., van der Tol R., Kootstra G. (2022) **T-LEAP: Occlusion-robust pose estimation of walking cows using temporal information** *Computers and Electronics in Agriculture* **192**
- Sakib F., Burghardt T. (2020) **Visual recognition of great ape behaviours in the wild** *arXiv preprint arXiv:2011.10759*
- Sanakoyeu A., Khalidov V., McCarthy M. S., Vedaldi A., Neverova N. (2020) **Transferring dense pose to proximal animal classes** :5233–5242
- Smuts Barbara B., Cheney Dorothy L., Seyfarth Robert M., Wrangham Richard W. (2008) **Primate societies**
- Strier K. B. (2016) **Primate behavioral ecology**
- Sun K., Xiao B., Liu D., and Wang J. (2019) **Deep high-resolution representation learning for human pose estimation** :5693–5703
- Wang M. Z., Hayden B. Y. (2019) **Monkeys are curious about counterfactual outcomes** *Cognition* **189**:1–10
- Wang M. Z., Hayden B. Y. (2021) **Latent learning, cognitive maps, and curiosity** *Current Opinion in Behavioral Sciences* **38**:1–7
- Wang M. Z., Hayden B. Y., Heilbronner S. R. (2022) **A structural and functional subdivision in central orbitofrontal cortex** *Nature communications* **13**:1–12
- Widge A. S., Heilbronner S. R., Hayden B. Y. (2019) **Prefrontal cortex and cognitive control: new insights from human electrophysiology** *F1000Research* **8**
- Wiltshcko A. B., Tsukahara T., Zeine A., Anyoha R., Gillis W. F., Markowitz J. E., Datta S. R. (2020) **Revealing the structure of pharmacobehavioral space through motion sequencing** *Nature neuroscience* **23**:1433–1443
- Yao Y., Bala P., Mohan A., Bliss-Moreau E., Coleman K., Freeman S. M., Park H. S. (2022) **OpenMonkeyChallenge: Dataset and Benchmark Challenges for Pose Estimation of Non-human Primates** *International Journal of Computer Vision* :1–16

Yoo S. B. M., Hayden B. Y., Pearson J. M. (2021) **Continuous decisions** *Philosophical Transactions of the Royal Society B* **376**

Yoo S. B. M., Hayden B. Y. (2020) **The transition from evaluation to selection involves neural subspace reorganization in core reward regions** *Neuron* **105**:712–724

Yoo S. B. M., Tu J. C., Piantadosi S. T., Hayden B. Y. (2020) **The neural basis of predictive pursuit** *Nature neuroscience* **23**:252–259

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

Author information

Nisarg Desai

Department of Neuroscience and Center for Magnetic Resonance Research, University of Minnesota, Minneapolis MN 55455

For correspondence: desai054@umn.edu

Praneet Bala

Department of Computer Science, University of Minnesota, Minneapolis MN 55455

Rebecca Richardson

Emory National Primate Research Center, Emory University, Atlanta GA 30329

Jessica Raper

Emory National Primate Research Center, Emory University, Atlanta GA 30329

Jan Zimmermann

Department of Neuroscience and Center for Magnetic Resonance Research, University of Minnesota, Minneapolis MN 55455

Benjamin Hayden

Department of Neuroscience and Center for Magnetic Resonance Research, University of Minnesota, Minneapolis MN 55455

Editors

Reviewing Editor

Ammie Kalan

University of Victoria, Canada

Senior Editor

George Perry

Pennsylvania State University, United States of America

Reviewer #1 (Public Review):

This work provides a new dataset of 71,688 images of different ape species across a variety of environmental and behavioral conditions, along with pose annotations per image. The authors demonstrate the value of their dataset by training pose estimation networks (HRNet-

W48) on both their own dataset and other primate datasets (OpenMonkeyPose for monkeys, COCO for humans), ultimately showing that the model trained on their dataset had the best performance (performance measured by PCK and AUC). In addition to their ablation studies where they train pose estimation models with either specific species removed or a certain percentage of the images removed, they provide solid evidence that their large, specialized dataset is uniquely positioned to aid in the task of pose estimation for ape species.

The diversity and size of the dataset make it particularly useful, as it covers a wide range of ape species and poses, making it particularly suitable for training off-the-shelf pose estimation networks or for contributing to the training of a large foundational pose estimation model. In conjunction with new tools focused on extracting behavioral dynamics from pose, this dataset can be especially useful in understanding the basis of ape behaviors using pose.

Since the dataset provided is the first large, public dataset of its kind exclusively for ape species, more details should be provided on how the data were annotated, as well as summaries of the dataset statistics. In addition, the authors should provide the full list of hyperparameters for each model that was used for evaluation (e.g., mmpose config files, textual descriptions of augmentation/optimization parameters).

Overall this work is a terrific contribution to the field and is likely to have a significant impact on both computer vision and animal behavior.

Strengths:

- Open source dataset with excellent annotations on the format, as well as example code provided for working with it.
- Properties of the dataset are mostly well described.
- Comparison to pose estimation models trained on humans vs monkeys, finding that models trained on human data generalized better to apes than the ones trained on monkeys, in accordance with phylogenetic similarity. This provides evidence for an important consideration in the field: how well can we expect pose estimation models to generalize to new species when using data from closely or distantly related ones?
- Sample efficiency experiments reflect an important property of pose estimation systems, which indicates how much data would be necessary to generate similar datasets in other species, as well as how much data may be required for fine-tuning these types of models (also characterized via ablation experiments where some species are left out).
- The sample efficiency experiments also reveal important insights about scaling properties of different model architectures, finding that HRNet saturates in performance improvements as a function of dataset size sooner than other architectures like CPMs (even though HRNets still perform better overall).

Weaknesses:

- More details on training hyperparameters used (preferably full config if trained via mmpose).
- Should include dataset datasheet, as described in Gebru et al 2021 (arXiv:1803.09010).
- Should include crowdsourced annotation datasheet, as described in Diaz et al 2022 (arXiv:2206.08931). Alternatively, the specific instructions that were provided to Hive/annotators would be highly relevant to convey what annotation protocols were employed here.
- Should include model cards, as described in Mitchell et al (arXiv:1810.03993).
- It would be useful to include more information on the source of the data as they are collected from many different sites and from many different individuals, some of which may introduce structural biases such as lighting conditions due to geography and time of year.
- Is there a reason not to use OKS? This incorporates several factors such as landmark visibility, scale, and landmark type-specific annotation variability as in Ronchi & Perona 2017 (arXiv:1707.05388). The latter (variability) could use the human pose values (for landmarks

types that are shared), the least variable keypoint class in humans (eyes) as a conservative estimate of accuracy, or leverage a unique aspect of this work (crowdsourced annotations) which affords the ability to estimate these values empirically.

- A reporting of the scales present in the dataset would be useful (e.g., histogram of unnormalized bounding boxes) and would align well with existing pose dataset papers such as MS-COCO (arXiv:1405.0312) which reports the distribution of instance sizes and instance density per image.

Reviewer #2 (Public Review):

The authors present the OpenApePose database constituting a collection of over 70000 ape images which will be important for many applications within primatology and the behavioural sciences. The authors have also rigorously tested the utility of this database in comparison to available Pose image databases for monkeys and humans to clearly demonstrate its solid potential. However, the variation in the database with regards to individuals, background, source/setting is not clearly articulated and would be beneficial information for those wishing to make use of this resource in the future. At present, there is also a lack of clarity as to how this image database can be extrapolated to aid video data analyses which would be highly beneficial as well.

I have two major concerns with regard to the manuscript as it currently stands which I think if addressed would aid the clarity and utility of this database for readers.

1. Human annotators are mentioned as doing the 16 landmarks manually for all images but there is no assessment of inter-observer reliability or the such. I think something to this end is currently missing, along with how many annotators there were. This will be essential for others to know who may want to use this database in the future.

Relevant to this comment, in your description of the database, a table or such could be included, providing the number of images from each source/setting per species and/or number of individuals. Something to give a brief overview of the variation beyond species. (subspecies would also be of benefit for example).

2. You mention around line 195 that you used a specific function for splitting up the dataset into training, validation, and test but there is no information given as to whether this was simply random or if an attempt to balance across species, individuals, background/source was made. I would actually think that a balanced approach would be more appropriate/useful here so whether or not this was done, and the reasoning behind that must be justified.

This is especially relevant given that in one test you report balancing across species (for the sample size subsampling procedure).

And another perhaps major concern that I think should also be addressed somewhere is the fact that this is an image database tested on images while the abstract and manuscript mention the importance of pose estimation for video datasets, yet the current manuscript does not provide any clear test of video datasets nor engage with the practicalities associated with using this image-based database for applications to video datasets. Somewhere this needs to be added to clarify its practical utility.