International Journal of Computer Vision
https://doi.org/10.1007/s11263-022-01698-2

OpenMonkeyChallenge: Dataset and Benchmark Challenges for Pose Estimation of Non-human Primates

Yuan Yao[1] · Praneet Bala[1] · Abhiraj Mohan[1] · Eliza Bliss-Moreau[4] · Kristine Coleman[2] · Sienna M. Freeman[3] · Christopher J. Machado[4] · Jessica Raper[3] · Jan Zimmermann[5] · Benjamin Y. Hayden[5] · Hyun Soo Park[1]

Received: 24 September 2021 / Accepted: 22 September 2022

Abstract
The ability to automatically estimate the pose of non-human primates as they move through the world is important for several subfields in biology and biomedicine. Inspired by the recent success of computer vision models enabled by benchmark challenges (e.g., object detection), we propose a new benchmark challenge called OpenMonkeyChallenge that facilitates collective community efforts through an annual competition to build generalizable non-human primate pose estimation models. To host the benchmark challenge, we provide a new public dataset consisting of 111,529 annotated (17 body landmarks) photographs of non-human primates in naturalistic contexts obtained from various sources including the Internet, three National Primate Research Centers, and the Minnesota Zoo. Such annotated datasets will be used for the training and testing datasets to develop generalizable models with standardized evaluation metrics. We demonstrate the effectiveness of our dataset quantitatively by comparing it with existing datasets based on seven state-of-the-art pose estimation models.

Keywords Behavioral tracking · Deep learning · Non-human primates · Dataset and benchmark challenge

Communicated by Matej Kristan.

Jan Zimmermann, Benjamin Y. Hayden, Hyun Soo Park are co-last authors.

✉ Praneet Bala
balax007@umn.edu

Yuan Yao
yaoxx340@umn.edu

Abhiraj Mohan
mohan056@umn.edu

Eliza Bliss-Moreau
eblissmoreau@ucdavis.edu

Kristine Coleman
colemank@ohsu.edu

Sienna M. Freeman
sienna.freeman@emory.edu

Christopher J. Machado
cjmachado@ucdavis.edu

Jessica Raper
jraper@emory.edu

Jan Zimmermann
janz@umn.edu

Benjamin Y. Hayden
benhayden@gmail.com

Hyun Soo Park
hspark@umn.edu

[1] Computer Science and Engineering, University of Minnesota, Minneapolis, USA

[2] Oregon National Primate Research Center, Beaverton, USA

[3] Emory National Primate Research Center, Atlanta, USA

[4] California National Primate Research Center, Davis, USA

[5] Neuroscience, University of Minnesota, Minneapolis, USA

1 Introduction

Recent years have seen great advances in systems that can automatically detect major landmarks in moving animals without fiducial markers, that is, *pose* (Mathis & Mathis, 2020; Dunn et al., 2021; Wiltschko et al., 2015; Karashchuk et al., 2020; Günel et al., 2019). Such pose estimation systems have greatly benefited research in fields that study the tracked species (e.g., rodents, flies, and fishes). However, the ability to estimate the pose of non-human primates has lagged, rendering the primate order a major outstanding problem in the field (Bala et al., 2020; Hayden et al., 2021). At the same time, non-human primates remain of great interest in biomedicine and related fields, including in neuroscience and

Published online: 16 October 2022

psychology, as well as in anthropology, epidemiology, and ecology. Automated pose estimation can also benefit animal welfare programs, veterinary medical practice and, indeed, conservation projects (Knaebe et al., 2022).

Estimating pose of non-human primates (NHPs) is particularly challenging due to their homogeneous body texture and exponentially large pose configurations (Bala et al., 2020). Two major innovations are needed to solve the pose estimation problem in NHPs. (1) Algorithmic innovation: pose models are expected to learn a generalizable visual representation that encodes the complex relationship between the visual appearance and spatial landmarks, which allows detecting poses in images with diverse primate identities, species, scenes, backgrounds, and poses in the wild environment. Existing deep learning models including convolutional pose machine (Wei et al., 2016), stacked hourglass model (Newell et al., 2016), DeeperCut (Insafutdinov et al., 2016), and AlphaPose (Fang et al., 2017) incorporate a flexible representation with a large capacity, which have shown strong generalization on human subjects. However, these models are not applicable to the image samples of NHPs from the out-of-training-distribution due to their characteristics (homogeneous appearance and complex pose). (2) Data innovation: the pose estimation models learn the visual representation from a large annotated dataset that specifies the locations of landmarks. Existing publicly available datasets including OpenMonkeyPose (200K multiview macaque images in a specialized laboratory environment) (Bala et al., 2020) and MacaquePose (13K in-the-wild macaque images) (Labuguen et al., 2021) are important resources for the development of pose estimation algorithms, and as such, extend the boundary of pose tracking performance of NHPs. However, due to limited data diversity (appearance, pose, viewpoint, environment, and species), existing datasets are currently insufficient for learning generalizable estimation models (See Figure 8 for model generalization across datasets).

Here we describe a novel dataset consisting of 111,529 images of NHPs in natural contexts with 17 landmark annotations. These datasets are obtained from various sources including the Internet, three National Primate Research Centers, and the Minnesota Zoo. Our motivation for developing this dataset includes inspiration from the recent success of computer vision models for human pose estimation (von Marcard et al., 2018), object detection (Lin et al., 2014), and visual question answering (Antol et al., 2015), enabled by standard benchmark challenges. For instance, the COCO benchmark challenges on object detection, segmentation, and localization have facilitated collective community effort through an annual competition, which in turn has been a driving force to advance computer vision models (Lin et al., 2014). In these domains, such datasets have served as a common comparison for friendly competitions, as a goal for experimentation, and as a benchmark to evaluate innovations. At the same time, such datasets tend to be difficult and expensive to generate, so sharing them makes economic sense for the field. Making them public greatly lowers the barriers to entry for new teams with innovative ideas.

With our dataset, we present a new benchmark challenge called *OpenMonkeyChallenge* for NHP pose estimation (http://openmonkeychallenge.com). It is an open and ongoing competition where the performance of each model is measured by the standard evaluation metrics (MPJPE (Iskakov et al., 2019) and PCK (Cao et al., 2019)). We leverage our unprecedentedly large annotated dataset, which includes diverse poses, species, appearances, and scenes as shown in Fig. 1. We split the dataset into the training and testing datasets where the testing dataset is used to evaluate the performance of competing models. We demonstrate that our dataset addresses the limitation on data diversity in the existing datasets. Specifically, we show the effectiveness of our dataset quantitatively by comparing it with existing datasets (e.g., OpenMonkeyPose and MacaquePose) based on state-of-the-art pose estimation models.

We organize this paper in the following way. We introduce the OpenMonkeyChallenge dataset, including data format, distribution, collection and annotation method in Sect. 3. Based on the dataset, we formulate the evaluation protocol in Sect. 4. We validate the usefulness of our dataset by comparing with existing datasets including MacaquePose (Labuguen et al., 2021) and OpenMonkeyPose (Bala et al., 2020), and study the estimation performance of existing models in Sect. 5.

2 Related Work

OpenMonkeyChallenge aims to advance non-human primate pose estimation through community efforts facilitated by a benchmark challenge.

2.1 Animal Pose Estimation

Understanding behaviors of animals is one of the main goals of multiple research domains including medicine, neuroscience, biology, and animal husbandry. For instance, ethogramming (Sade, 1973) is a major tool in neuroscience to categorize behavioral states and their transitions, e.g., sitting, standing, and running. Standard ethogramming involves manual annotations by experienced researchers. It is a costly and labor-intensive process, which limits the repeatability and makes it difficult to scale up. The difficulty of animal pose estimation contrasts sharply with human pose estimation, in which computer vision enables pose estimation at massive scale in a fully automated fashion. There exists various detection frameworks such as convolutional pose

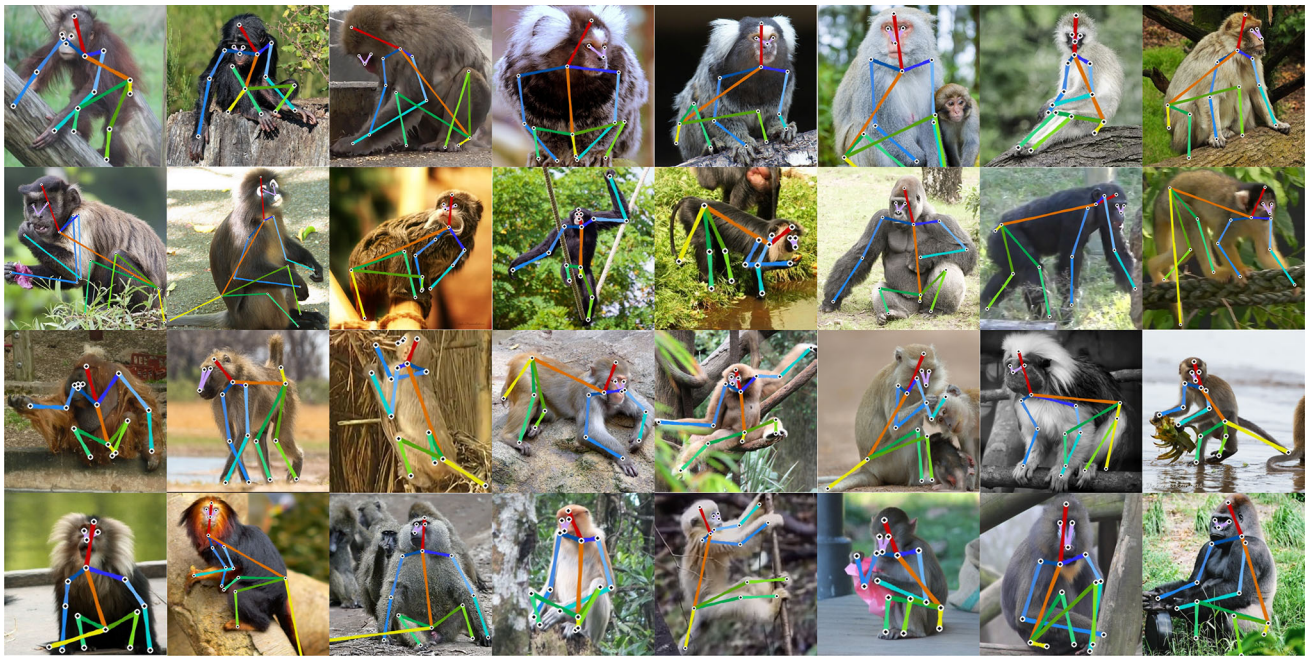


Fig. 1 We present an OpenMonkeyChallenge using 111,529 annotated images of non-human primates (26 species), obtained from the Internet, three National Primate Research Centers, and the Minnesota Zoo. 17 landmarks are manually annotated for each image. OpenMonkeyChallenge aims to extend the boundary of pose estimation for non-human primates across multiple species through an annual competition to build generalizable pose estimation models

machine (Wei et al., 2016), hourglass network (Newell et al., 2016), DeepCut / DeeperCut (Pishchulin et al., 2016; Insafutdinov et al., 2016), Openpose (Cao et al., 2019), DeepPose (Toshev & Szegedy, 2014), Densepose (Güler et al., 2018), recurrent human pose (Belagiannis & Zisserman, 2017), and deep fully-connected part-based models (de Bem et al., 2018), that have improved the performance boundary of human pose estimation by leveraging large-scale benchmark datasets. Recently, equivalent CNN models have been designed to estimate animal poses. DeepLabCut (Mathis et al., 2018) retargeted a convolutional neural network (CNN) trained from a generic image recognition task to detect pose of animals given a partially labeled dataset. Due to the strong generalizability of CNNs, unlabeled images can be successfully annotated. LEAP (Pereira et al., 2018) took a new step by designing an efficient CNN architecture that can be readily integrated in a realtime graphical user interface. This allows users to easily interact with the CNN, facilitating semi-automatic pose annotation. These approaches are agnostic to the animal kinematic structure, which allows estimating poses of diverse species such as flies, cheetahs, fishes, and mice. However, due to the nature of supervised learning, it still requires substantial amount of data to annotate and shows inferior performance when applying to a new target video. Self-supervised learning can be a viable solution. For instance, multiview self-supervision (Günel et al., 2019; Yao et al., 2019; Bala et al., 2020) that uses multiview geometry to constraint the pose, which allows using unlabeled data for training. Notably, OpenMonkeyStudio (Bala et al., 2020) designed a large multi-camera system called OpenMonkeyStudio to track dexterous non-human primates by multiview bootstrapping.

2.2 Animal and Primate Datasets

What makes human pose detection in computer vision different from that of other animals is the existence of a large annotated dataset. In the human domain, the datasets such as MPII (Andriluka et al., 2014), COCO (Lin et al., 2014), FLIC (Sapp & Taskar, 2013), and PoseTrack (Iqbal et al., 2017; Andriluka et al., 2018), HiEve (Lin et al., 2020) constitute millions of images across diverse poses, appearance, occlusion, and background. This allows learning a CNN pose estimator that can be readily applicable to a new pose and scene. Further, a benchmark challenge such as the COCO keypoint detection challenge facilitates community effort to improve the detector performance every year. We summarize the different datasets in Table 1. Similar to human pose estimation, there exists animal pose datasets with specific target objectives, such as benchmarks for Amur Tiger re-identification (Li et al., 2020), and animal behavior understanding (Ng et al., 2022). However, due to the diversity of species in the animal kingdom, size of animal datasets are lacking in comparison to that of human datasets. Since non-

Table 1 Overview of publicly available datasets for articulated human and primate pose estimation.

Dataset	# of Poses	Data type
We are family (Eichner & Ferrari, 2010)	3131	Humans
FLIC (Sapp & Taskar, 2013)	20,928	Humans
MS COCO Keypoints (Lin et al., 2014)	250,000	Humans
PoseTrack17 (Iqbal et al., 2017)	16,219	Humans
PoseTrack18 (Andriluka et al., 2018)	153,615	Humans
HiEve (Lin et al., 2020)	1,099,357	Humans
OMS (Bala et al., 2020)	195,228	Primates (Rhesus Macaque)
MacaquePose (Labuguen et al., 2021)	13,000	Primates (Rhesus Macaque)
Proposed (OMC)	111,529	Primates (26 species)

For each dataset we report the number of annotated poses and species

human primates play a pivotal role in biomedicine and related fields, including neuroscience, psychology, anthropology and ecology, the paper inherently focuses on non-human primate pose estimation. Existing datasets, in particular, for non-human primates, are rather small and domain specific, which precludes learning a generalizable CNN model. For example, OpenMonkeyStudio (Bala et al., 2020) included 200K multiview images that are captured from controlled and specialized laboratory conditions, which is not generalizable to primates in natural habitats. The MacaquePose dataset (Labuguen et al., 2021) includes 13K annotated images from the Internet that span diverse environments and poses. However, it is limited to one species. This paper presents a new large dataset of multiple primates including 26 species in natural habitats and formulates a benchmark challenge to advance primate tracking in the wild.

3 OpenMonkeyChallenge Benchmark Dataset

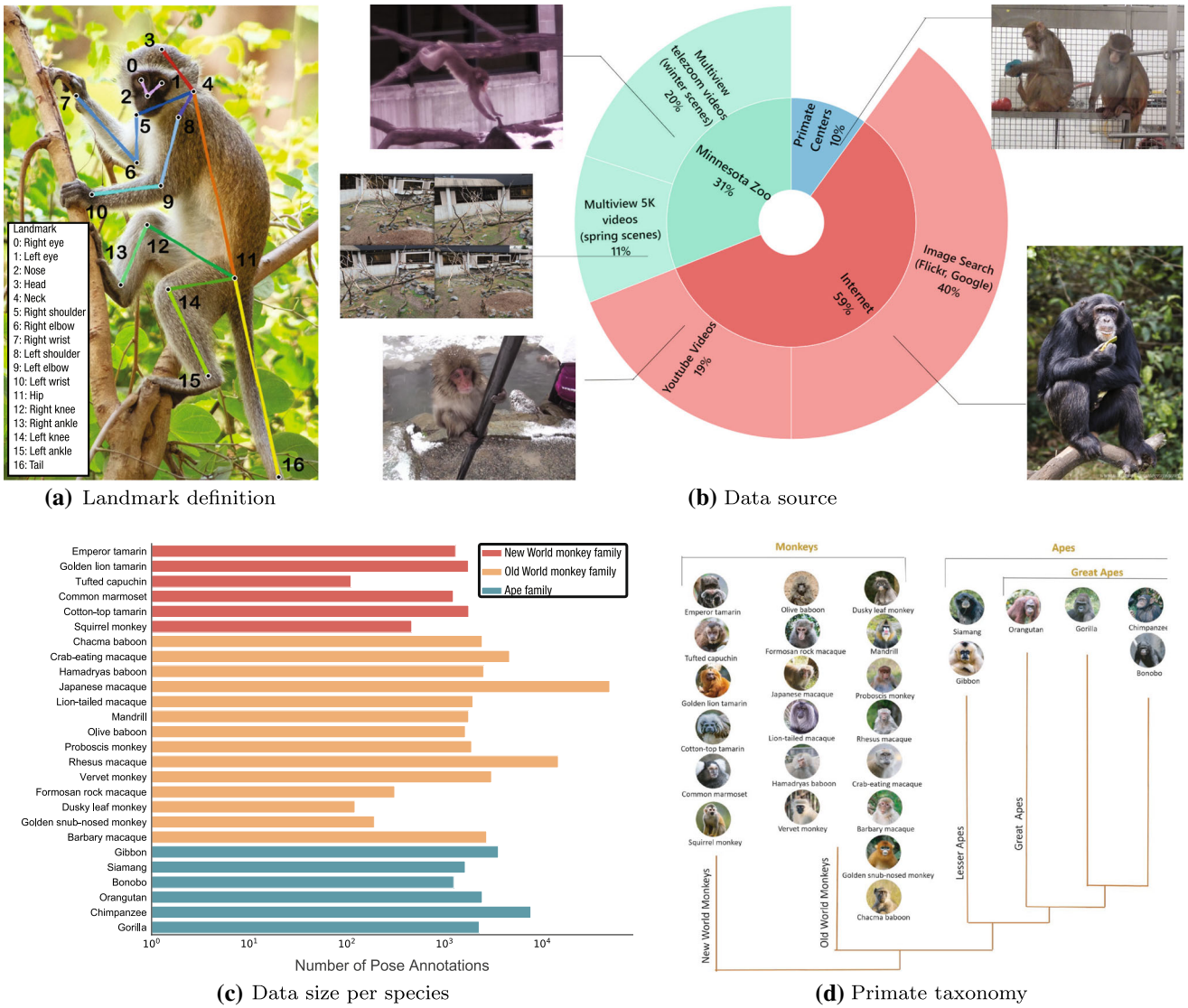
We collected 111,529 images of 26 species of primates (6 New World monkeys, 14 Old World monkeys, and 6 apes), including Japanese macaques, chimpanzees, and gorillas from (1) internet images and videos, such as Flickr and YouTube, (2) photographs of multiple species of primates from three National Primate Research Centers, and (3) multiview videos of 27 Japanese macaques in the Minnesota Zoo (Fig. 2b and d). For each photograph, for example, in Fig. 1, we cropped the region of interest such that each cropped image contains at least one primate. We ensured that all cropped images have a higher resolution than 500×500 pixels.

We identify the region of interest (i.e., bounding box detection) by bootstrapping with a weak monkey detector (Redmon & Farhadi, 2018) followed up by manual refinement and use a commercial annotation service (Hive AI) to manually annotate the 17 landmarks. The 17 landmarks together comprise a pose. Our landmarks include Nose, Left eye, Right eye, Head, Neck, Left shoulder, Left elbow, Left wrist, Right shoulder, Right elbow, Right wrist, Hip, Left knee, Left ankle, Right knee, Right ankle, and Tail. Each data instance is made of a triplet, image, species, pose as shown in Fig. 2a.

We split the benchmark dataset into training (66,917 images, 60%), validation (22,306 images, 20%), and testing (22,306 images, 20%) datasets. We minimize visually similar image instances across splits by categorizing them using the time of capture, video and camera identification numbers, and photographers. Fig. 2c illustrates the data distribution across species, and each species includes more than 100 annotated images. We also visualize the distribution of bounding box sizes in Fig. 3. The bounding box sizes indicate the diagonal length of the bounding boxes for each primate instance in the image.

Data Statistics The OpenMonkeyChallenge dataset contains a diversity of species, poses, and appearances. We use Uniform Manifold Approximation and Projection (UMAP) (McInnes et al., 2018) to reduce the high dimensional pose ($\mathbb{R}^{34}$ for 17 landmarks) into two dimensions as shown in Fig. 4. To generate a spatially meaningful distribution, we normalize the pose coordinates. Specifically, the coordinates of each pose (17 landmarks) are normalized by centering the root landmark (hip joint), i.e., the landmark coordinate is relative with respect to the hip joint. These relative coordinates are normalized by the size of the bounding box to account for different sizes of images. Further, we align the orientation such that all poses have the same facing directions. This results in coherent clusters with poses.

The primates are classified into three types based on their families: New World monkeys, Old World monkeys, and apes. Poses are distributed across species, which are highly correlated with the semantically meaningful poses such as sitting, standing, and climbing. For each cluster, we visualize average images by aligning the poses. Overall, we find



(a) Landmark definition **(b)** Data source

(c) Data size per species **(d)** Primate taxonomy

Fig. 2 **a** We annotated the 17 landmarks that describe the pose of the primate in an image. **b** We collect image data from diverse sources: Internet image searches and YouTube videos, professional photographs from three National Primate Research Centers, and multiview videos from the Minnesota Zoo. The original images are cropped to include at least one primate and ensured to have higher than 500×500 resolution. **c** Our dataset is composed of 26 species of monkeys and apes, and more than 100 images are annotated for each species. We split the data into training, validation, and testing datasets, approximately 6:2:2 ratio, respectively. **d** Primate taxonomy. Our dataset includes diverse species of monkeys and apes

that the majority of data consists of sitting poses from a variety of views.

The clustering results also highlight the difference in locomotion patterns among primate families. For example, Old World monkeys (orange) heavily outnumber the other two families and dominate most of the clusters, and a few clusters of which the average pose is vertical climbing are by large composed of the apes (green). Other actions, such as sitting, walking, and standing, are common in all the primate families.

3.1 Data Collection Method

We collected images from three sources: internet images and videos, photographs from National Primate Research Centers, and multiview videos from the Minnesota Zoo.

Internet images Approximately 59% of our dataset were collected from the Internet through Image and video search engines. For instance, we used the Flickr API to scrape the list of image URLs and YouTube search engine to find relevant videos using species name keywords. We ensure visual diversity (shapes, poses, viewpoints, sizes, colors, and envi-

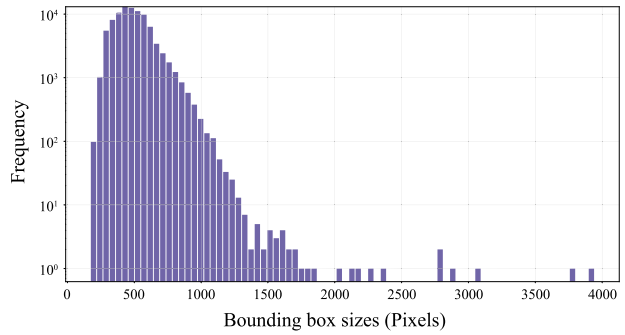


Fig. 3 We visualize the distribution of bounding box sizes, where the bounding box sizes is the diagonal length of the bounding box

ronments) and quality (image resolution, blurriness, lighting, and occlusion) of the scraped data via manual inspection. For the common species such as rhesus macaque, mandrill, and gorilla, image searches were sufficient. For the rarer species such as marmoset, we leveraged the video search features and extracted image frames from the videos. Not only does this approach allow us to obtain more images of the rarer species, but we also collected images that are less iconic than those from search engines. We hired two annotators for image and video searches. After image collections, we annotated the bounding boxes that contain the primate instances. For a subset of internet images, we do not own the copyright of the images. We specify the terms and conditions of use in the website.

Photographs from national primate centers We made use of high quality images of primates photographed by staff at two National Primate Centers: Emory National Primate Research Center and the Oregon National Primate Research Center. The photographers were asked to capture primate images from diverse viewpoints and poses at high resolution (>2K pixel resolution) and often made use of a tele-zoom lens. 10,500 images are captured from the professional photographers across the primate centers. Further, we collected videos from California National Primate Research Center. Still images were extracted from a video library developed at the California National Primate Research Center (Machado et al., 2011; Bliss-Moreau et al., 2013). Video footage of monkeys behaving was recorded at the center's large 0.5 acre outdoor enclosures and from images of monkeys in the laboratory. Videos were edited to be 30 s in duration and included a range of behaviors, including aggression, grooming, feeding, resting, and affective displays. Still images were captured from the videos for use in this project.

Multiview videos from the Minnesota Zoo We used video cameras to capture video images of a large troop ($n = 27$ individuals) of snow monkeys (*Macaca fuscata*) at the Minnesota Zoo (Apple Valley, MN) for a long duration (1 week). Unlike the images taken by photographers who precisely control focal length and viewpoint to ensure high resolution images, these video cameras passively observe the scene. The monkeys inhabit a large arena that facilitates natural social interactions among them. It is a large open space (bigger than 600 m^2), which leads to a new challenge as monkeys appear small in images (10–50 pixel size) if a wide field of view lens is used to cover the large area. We address this challenge by using a multi-camera system made of 20–30 cameras where each camera observes a small area (up to 5×5m) using a narrow field of view (long or tele-zoom focal length). We identified the regions of the enclosure that frequently involve diverse activities (e.g., trails, ponds, and playgrounds) to

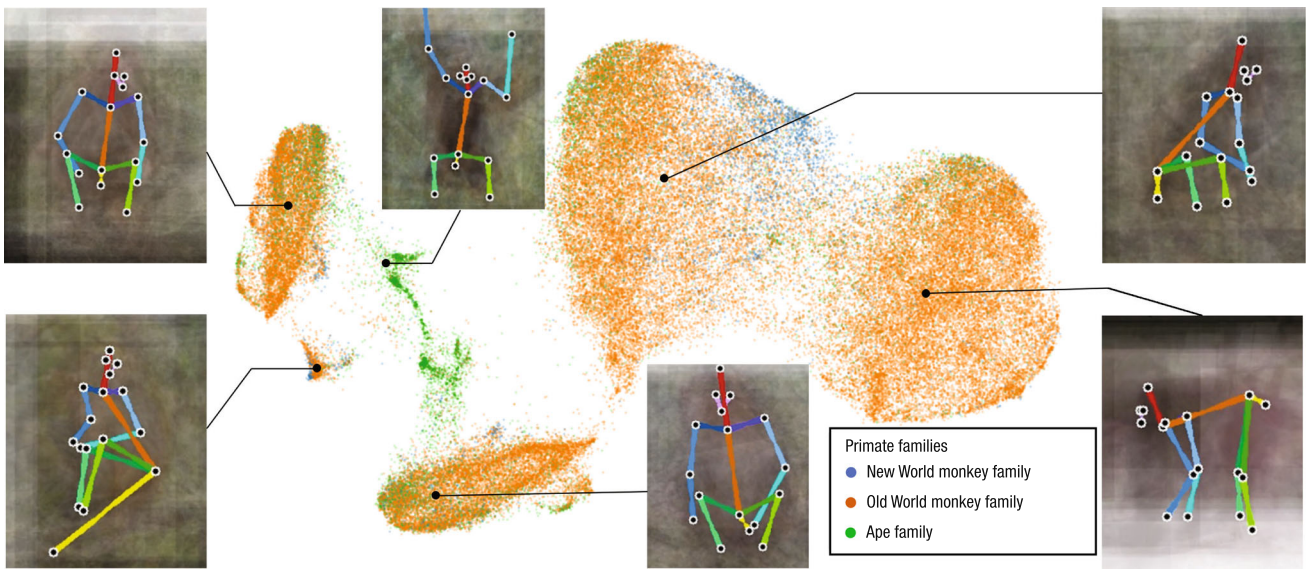


Fig. 4 We visualize a distribution of poses of the OpenMonkeyChallenge dataset using UMAP for dimension reduction. For each cluster, we show an average image overlaid with the median pose to illustrate its visual pattern

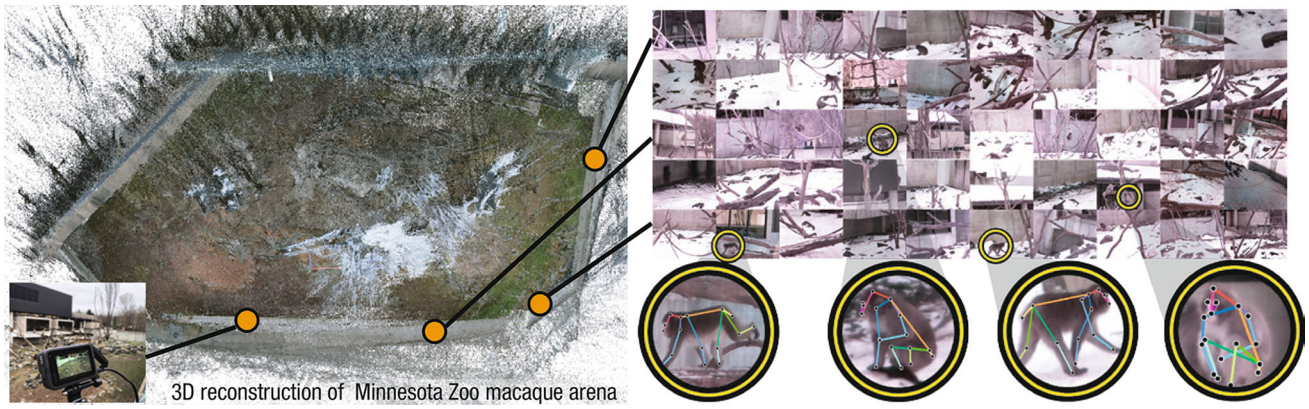


Fig. 5 We show the 3D reconstruction of the Minnesota Zoo macaque arena using the multiview cameras mounted along the enclosure for data capture. The multiview images and four arbitrary cropped images superimposed with the projection of the reconstruction are also shown

maximize the monkey appearance in images. Because videos were multiview videos, we used a monkey bounding box detection algorithm to identify the monkeys and then refined these boxes manually. We collected the image data from two seasons (winter and spring) to maximize diversity of background visual appearance (Fig. 5).

3.2 Semi-automatic Annotation

Identifying images that contain primate instances from videos and annotating their landmarks are prohibitively labor intensive tasks. For instance, fewer than 2% of the frames in the videos from narrow field of view (FOV) cameras used in the zoo data contain primate instances. Watching every frame in videos to annotate bounding boxes for primate instances is time-consuming, e.g., one day zoo videos is equivalent to approximately 5000 hours (∼6,000,000 images) of labor. Instead, we leverage an iterative bootstrapping approach to address the bounding box annotation task.

Bounding box proposal We trained a weak primate detector that can predict the bounding box of a primate instance given an image. The bounding box (left-top corner coordinate, width, and height) of 3000 internet images are manually annotated, and used to train a YOLOv3 model (Redmon & Farhadi, 2018) that can recognize primate bounding boxes. We use a lower threshold for bounding box detection such that the false positives are slightly more common than the false negatives. This bounding box prediction automates identifying image frames that contain primate instances, so that a majority of image frames without primates can be pruned, which significantly reduces the required labor. Further, it provides bounding box candidates for each image.

Bounding box refinement Given the bounding box proposals, we designed a graphic user interface to visualize and refine bounding boxes as shown in Fig. 6. The interface shows an image with bounding box candidates. The annotators are asked to find false positives and redundant poses from the previous frames (green bounding box with red cross). Further, they can add bounding boxes (red bounding boxes). Human helpers can perform this task in 5∼15 seconds per image. With this manual refinement, we ensure all cropped images include at least one primate. Once we have the refined bounding boxes, we incrementally increase the size of data to re-train the bounding box detection model to adapt to the target environments.

Landmark annotation Given the bounding box annotations, we used a commercial annotation service (Hive AI) to annotate 17 landmarks from cropped images. When the landmarks are occluded, the annotators are instructed to specify the best guess location and to indicate visibility.

Every one of the photographs that were annotated professionally was checked by hand by two experts who have a background in neuroscience or primatology. Photographs for which there was doubt about accuracy were removed from the dataset, or else in some cases returned to the annotation service for re-annotation. We estimate that the proportion of photographs that failed this test was about 1%.

4 Benchmark Evaluation Protocol

The annotations for the training and validation datasets are publicly available while that for the testing dataset is hidden. We have established the evaluation server to automatically evaluate the performance of the competing models on the testing dataset and maintain the leader. Specifically, the species landmark detection result on training/validation/testing datasets is uploaded to the evaluation server in a pre-defined file format, and the evaluation result is generated by the server. Users are asked to post their results in the leaderboard that sorts the performance based on three standard keypoint metrics: mean per joint position

Fig. 6 We design a graphic user interface to refine bounding boxes. Given an image and bounding box proposals (green boxes) from a weak detector, the annotators are asked to remove false positives and redundant poses from previous frames (green bounding boxes with red cross) and to add false negatives (red bounding boxes) (Color figure online)

error (MPJPE), probability of correct keypoint (PCK) metric at error tolerance, and average precision (AP) based on object keypoint similarity (OKS).

Mean per joint position error (MPJPE) (Iskakov et al., 2019) measures normalized error between the detection and ground truth for each landmark (the smaller, the better):

$$\text{MPJPE}_i = \frac{1}{J}\sum_{j=1}^{J}\frac{\|\widehat{\mathbf{x}}_{ij} - \mathbf{x}_{ij}\|}{W}$$

where MPJPE_i is the MPJPE for the ith landmark, J is the number of image instances, $\widehat{\mathbf{x}}_{ij} \in \mathbb{R}^2$ is the ith predicted landmark in the jth image, $\mathbf{x}_{ij} \in \mathbb{R}^2$ is its ground truth location, and W is the width of the bounding box. Note that MPJPE measures the normalized error relative to the bounding box size W, e.g., 0.1 MPJPE for 500×500 bounding box corresponds to 50 pixel error.

Probability of correct keypoint (PCK) (Cao et al., 2019) is defined by the detection accuracy given error tolerance (the bigger, the better):

$$\text{PCK}@\epsilon = \frac{1}{17J}\sum_{j=1}^{J}\sum_{i=1}^{17}\delta\left(\frac{\|\widehat{\mathbf{x}}_{ij} - \mathbf{x}_{ij}\|}{W} < \epsilon\right)$$

where $\delta(\cdot)$ is an indicator function that outputs 1 if the statement is true and zero otherwise. ϵ is the spatial tolerance for correct detection. Note that PCK measures the detection accuracy given the normalized tolerance with respect to the bounding box width, e.g., PCK@0.2 with 200 pixel bounding box size refers to the detection accuracy where the detection with the error smaller than 40 pixels is considered as a correct detection.

For the sake of rigor, we also provide results for different variations of PCK. The formulation for the same can be found as below,

$$\text{PCKd}@\epsilon = \frac{1}{17J}\sum_{j=1}^{J}\sum_{i=1}^{17}\delta\left(\frac{\|\widehat{\mathbf{x}}_{ij} - \mathbf{x}_{ij}\|}{d} < \epsilon\right)$$

Note that PCKd measures the detection accuracy with respect to the diagonal length of the bounding box (d). PCKh measures the detection accuracy with respect to the head size (hs). For the purpose of this paper, hs is calculated using the ground truth head and neck landmarks.

$$\text{PCKh}@\epsilon = \frac{1}{17J}\sum_{j=1}^{J}\sum_{i=1}^{17}\delta\left(\frac{\|\widehat{\mathbf{x}}_{ij} - \mathbf{x}_{ij}\|}{hs} < \epsilon\right)$$

Average precision (AP) measures detection precision (the bigger, the better):

$$\text{AP@}\epsilon = \frac{1}{17J}\sum_{j=1}^{J}\sum_{i=1}^{17}\delta(\text{OKS}_{ij} \geq \epsilon)$$

where OKS measures keypoint similarity (Lin et al., 2014):

$$\text{OKS}_{ij} = \exp\left(-\frac{\|\widehat{\mathbf{x}}_{ij} - \mathbf{x}_{ij}\|^2}{2W^2k_i^2}\right)$$

where OKS_{ij} is the keypoint similarity of the jth image of the ith landmark. k_i is the ith landmark relative tolerance. Unlike PCK, OKS measures per landmark accuracy by taking into account per landmark variance k_i (visual ambiguity of landmarks), e.g., eye is visually less ambiguous than hip. We define k_i based on COCO keypoint challenge and augment the tail landmark such that $k_{tail} = k_{wrist}$.

We created a website http://openmonkeychallenge.com/ that shares the dataset and benchmark challenges. The training/validation/testing datasets can be downloaded from the website. The annotations are available for the training and validation datasets. The testing results (landmark detection on the testing data) from the developed models can be submitted to the evaluation server in JSON file format:

```
{\image_id'' = int,
\file_name''  = str,
\landmarks''  = [x1,y1,...,x17,y17]}
```

where x_i and y_i are x, y coordinates of the ith landmark. The evaluation server will return the performance on the testing data using MPJPE, PCK, and AP metrics. The evaluation results will be posted in the leaderboard that sorts the algorithms based on the performance. Optionally, the users can opt out. The website includes step-by-step description of the evaluation process, file format, and visualization code.

5 Dataset Evaluation

We evaluate OpenMonkeyChallenge data in three aspects: (1) generalization across datasets via cross-dataset evaluation; (2) data performance gap between humans and primates; and (3) baseline performance across state-of-the-art pose estimation.

5.1 Cross-dataset Evaluation

To evaluate the generalizability of our dataset, we conduct a cross-dataset evaluation with OpenMonkeyPose (Bala et al., 2020) and MacaquePose (Labuguen et al., 2021). OpenMonkeyPose (Bala et al., 2020) consists of 195,228 annotated images simultaneously captured by 62 precisely arranged high-resolution video cameras. The dataset involves inanimate objects (barrels, ropes, feeding stations), two background colors (beige and chroma-key green), and four rhesus macaque subjects varying in size and age (5.5–12 kg). MacaquePose (Labuguen et al., 2021), a dataset with more than 13,083 images of macaque, is collected by searching for images with a 'macaque' tag in Google Open Images and captured in zoos and the Primate Research Institute of Kyoto University.

We split each dataset into training (60%), validation (20%), and testing (20%) sets. We train a convolutional pose machine (CPM) (Wei et al., 2016) using the training data from one of the datasets with spatial data augmentation (translation and rotation) until it starts to overfit based on the model performance on the validation data, and test that model on the testing data from each dataset. Fig. 7 summarizes the performance in MPJPE. The CPM model trained by the OpenMonkeyChallenge dataset achieves the lowest MPJPE on the OpenMonkeyChallenge and MacaquePose (Labuguen et al., 2021) test datasets, which indicates that the diversity and generalizability of our training dataset (outperforming MacaquePose own testing data). For the OpenMonkeyPose testing dataset, it achieves the second best close to the OpenMonkeyPose. This is mainly caused by the domain difference: the images of OpenMonkeyStudio were captured by a controlled lab environment that has a homogeneous background and monkey texture. For the same reason, this model has poor performance on the other two datasets due to its low generalizability.

Each dataset has its own bias, i.e., it is expected to perform best on the model trained on its own training dataset, e.g., MP trained model on MP testing data. Therefore, for fair comparison, we employ the cross data evaluation, e.g., both MP and OMC trained models on OMP testing data, while unfair comparison would be comparing the performance of cross evaluation with that of self evaluation, e.g., comparing the performance of OMC trained model on MP testing data with that of MP trained model on MP testing data. Given the comparison protocol, the model trained on OMC significantly outperforms on other datasets as shown in Fig. 8: (1) Compared to the model trained on MP, the OMC trained model outperforms with 72% error reduction on the OMP dataset; (2) Compared to the model trained on OMP, the OMC trained model outperforms with 25% error reduction on the MP dataset.

The model trained on OMC performs competitively: (3) OMC trained model when tested on MP testing dataset results in 4% error reduction compared to MP trained model tested on MP dataset; (4) Compared to the model trained on OMP, the OMC trained model underperforms with 13% greater error on the OMP testing dataset. Given the comparison, introduction of OMC dataset is not a trivial addition. Its data

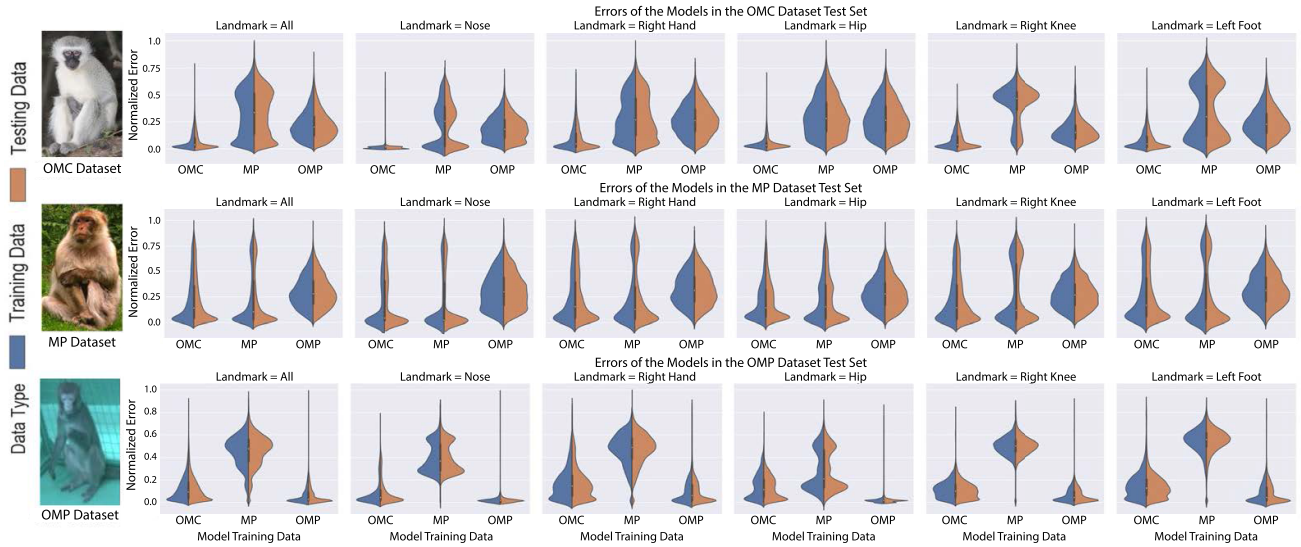


Fig. 7 Three detection models are trained on OpenMonkeyChallenge (OMC), MacaquePose (MP), and OpenMonkeyPose (OMP), respectively. In each box, we visualize three violin plots corresponding to the detection models. Each violin plot shows the normalized error histogram of landmarks on training (blue) and testing (brown) data (first row: OMC dataset; second row: MP dataset; third row: OMP dataset). The model trained on OMC (left violin plot in each box) is the most generalizable (inverted T shape histogram) (Color figure online)

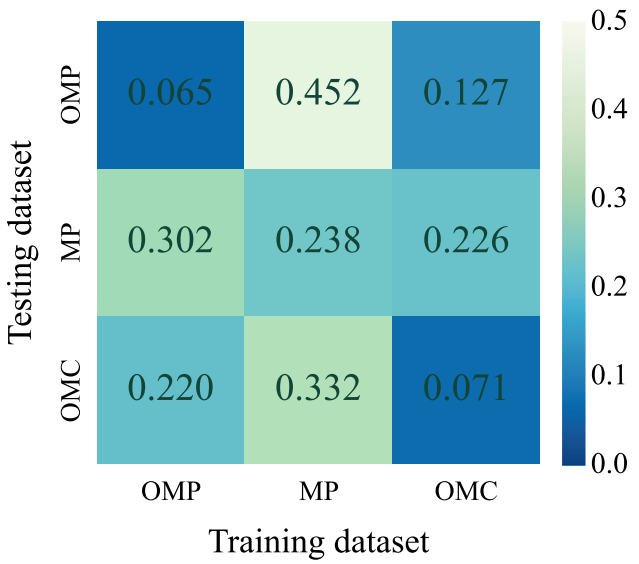


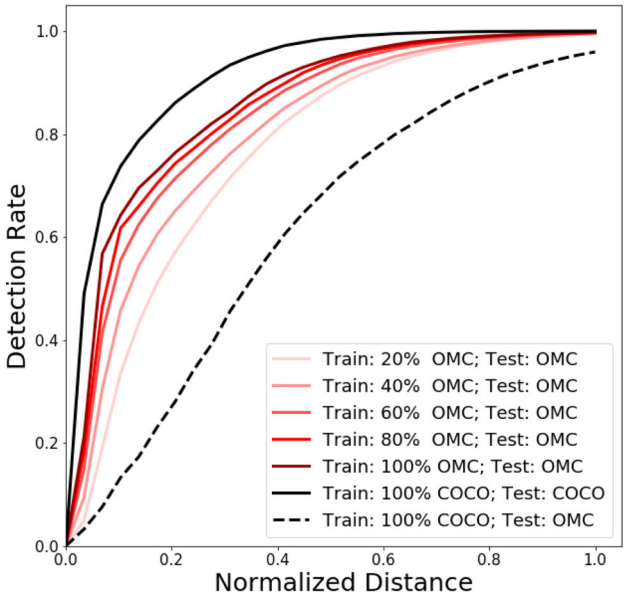
Fig. 8 We summarize the cross-dataset evaluation to show the generalizability using the normalized error in a confusion matrix, e.g., the second row of the third column shows the normalized error of the MP testing data for the model trained on OMC training dataset. The model trained on OMC dataset shows the smallest error or comparable to the model that is testing on its own training data

diversity substantially improves the generalizability of the model. In addition, we show the analysis of the performance of a model trained on the three datasets together. This has been indicated in the Fig. 9b. We also evaluate the impact of pre-training as human datasets can be beneficial for training low level features. In Fig. 9b, a key observation is that the impact introduced by pre-training is minimal because OpenMonkeyChallenge dataset is sufficiently large to properly learn low level features.

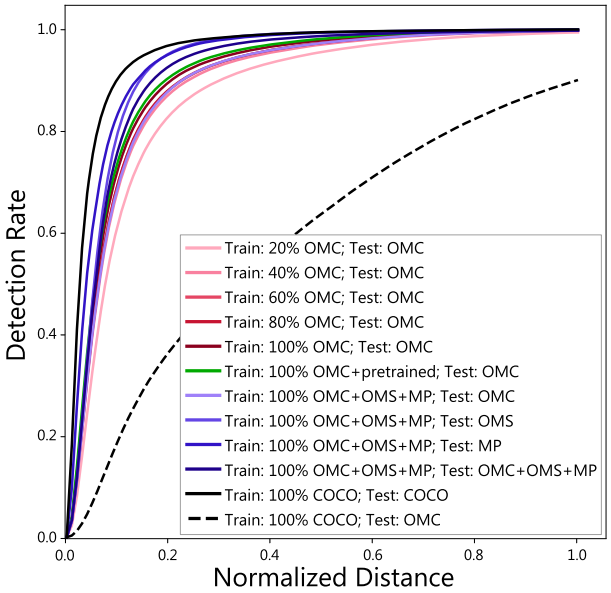
5.2 Comparison with Human Pose Estimation

The distal goal of our benchmark challenge is to achieve a performance comparable to human pose estimation. For instance, a state-of-the-art human pose detector (CPM) trained on the COCO-keypoint dataset (Lin et al., 2014) produces 0.061 MPJPE or 0.849 PCK@0.2 (upper bound performance). Without a nontrivial modification, a CPM trained on our dataset achieves 0.074 MPJPE or 0.761 PCK@0.2 as reported in Fig. 9a. In other words, there exists a considerable performance gap between human and primate pose estimation. The major performance gap is attributed to the size and diversity of the dataset. COCO dataset includes 250 k annotated image instances while OMC dataset includes 111 k instances. The OMC dataset is precise as each annotation was reviewed by at least two experts on neuroscience and primatology. Monkeys and primates are, in general, more agile, producing diverse poses than humans. Further, unlike humans who wear clothes that provide a strong semantic cue for joint localization, the appearance of monkeys and primates is, largely, homogeneous. This poses a main challenge of identifying the landmark locations. The goal of this paper is to identify this performance gap, and proposes a community effort to develop semi-supervised learning frameworks that can leverage unlabeled data to address this limitation.

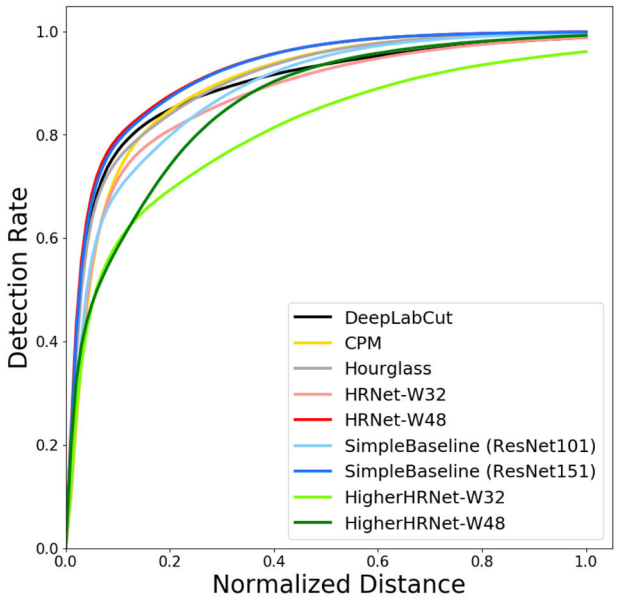
Further, we evaluate the human detection model on our dataset, which achieves 0.197 MPJPE or 0.265 PCK@0.2 for reference (lower bound performance). We propose that the



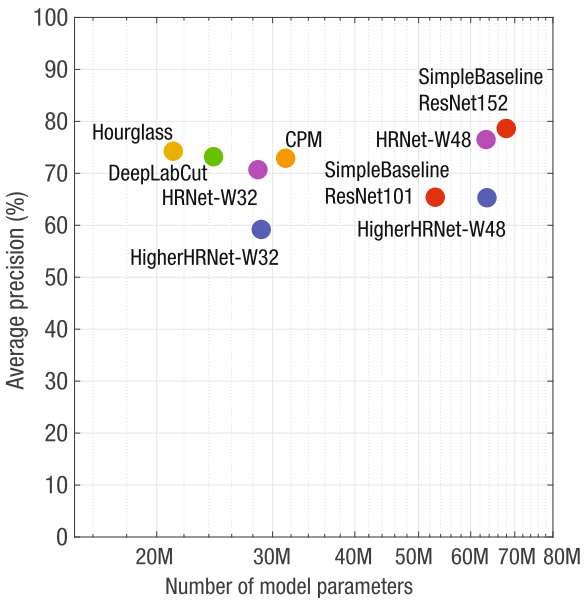
(a) Comparison with human pose estimation on CPM model



(b) Comparison with human pose estimation on HRNet model



(c) Comparison with baseline algorithms



(d) AP vs. model size

Fig. 9 **a** We use PCK to measure keypoint detection performance on CPM models. The black solid line shows the performance of the human landmark detector (train and test on COCO) that forms the upper bound of the primate landmark detector. The black dotted line shows the testing performance of the human landmark detector (trained on COCO) on OMC data without retraining, which forms the lower bound. OMC dataset allows us to train a primate specific model that shows significant performance improvement from the lower bound. Yet, there still exists a large gap between the human and primate landmark detectors. We also visualize the performance improvement as increasing the number of OMC training data. **b** We use PCKd to measure keypoint detection performance on HRNet models. The plot comprises of curves generated using similar experimental setup as shown in Fig. 9a. The solid green line shows the performance of primate landmark detector trained using pretrained model weights on OMC dataset. We also visualize the performance of the primate landmark detector (trained on OMC, OMS and MP dataset) on different datasets. **c** Six state-of-the-art pose estimation models are trained with OMC datasets. These are PCK curves in the test set from these models. **d** We show the average precision (AP) of state-of-the-art models as a function of the number of model parameters. If the data size is large enough, a larger model is likely to learn complex visual patterns

Table 2 Model comparison with MPJPE metric of each landmark with top-down and bottom-up methods on the OpenMonkeyChallenge test set

Method	Eye	Nose	Head	Neck	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	Tail	Mean
Top-down												
DeepLabCut	0.042	0.044	0.056	0.066	0.079	0.090	0.115	0.107	0.096	0.117	0.158	0.089
CPM	0.022	0.024	0.048	0.060	0.077	0.093	0.112	0.081	0.075	0.087	0.118	0.074
Hourglass	0.018	0.019	0.040	0.064	0.084	0.093	0.089	0.082	0.070	0.081	0.108	0.069
HRNet-W48	**0.016**	**0.018**	0.042	0.055	**0.076**	**0.082**	**0.082**	**0.076**	**0.065**	**0.077**	**0.096**	**0.064**
HRNet-W32	0.017	0.020	0.042	0.059	0.078	0.086	0.089	0.082	0.066	0.080	0.102	0.067
SimpleBaseline (ResNet152)	0.017	0.020	0.043	**0.054**	0.078	0.083	0.085	0.077	0.067	0.079	0.099	0.065
SimpleBaseline (ResNet101)	0.021	0.025	0.031	0.094	0.094	0.111	0.117	0.102	0.079	0.094	0.136	0.083
Bottom-up												
HigherHRNet-W32	0.035	0.040	**0.022**	0.119	0.128	0.162	0.108	0.151	0.069	0.110	0.183	0.102
HigherHRNet-W48	0.023	0.034	0.034	0.109	0.122	0.126	0.098	0.124	0.086	0.100	0.161	0.092

The values represented in bold signify the performance of the best model for a given metric/landmark

major benefits associated with human pose estimation is the progress in developing, efficient and generalizable models with self-supervised methods (Yang et al., 2021; Sumer et al., 2017; Jakab et al., 2020; Ludwig et al., 2021; Wan et al., 2019; Ren & Lee, 2018). We anticipate that a similar algorithmic innovation will close the gap.

We also conduct an ablation study to evaluate the impact of large data, i.e., how the amount of training data affects the landmark detection accuracy on the testing dataset. Given the training data, we incrementally reduce the amount of the training images used for model training by 20% at each time and measure the model performance using PCK metric. Fig. 9a shows the impact of the data increments, i.e., the model trained on 100% training data achieves the highest PCK result, outperforming the model with 20% of training data by 15% at PCK@0.2.

In addition to CPM, we evaluate the dataset using HRNet (Sun et al., 2019) as shown in Fig. 9b. HRNet has a higher capacity, which allows learning a more generalizable model, achieving higher accuracy (PCKd@0.1: 0.895). Nonetheless, the trend remains the same: OpenMonkeyChallenge dataset is far smaller than the human dataset, which introduces a fundamental performance gap between humans and primates.

5.3 State-of-the-art Detection Model Performance Evaluation

We conduct a comparative study on the performance of the state-of-the-art pose detection models using the OpenMonkeyChallenge dataset. We train nine pose estimation models until it starts to overfit based on the performance on the validation data. These models can be categorized into the top-down methods and the bottom-up methods. The top-down models (DeepLabCut with ResNet (Mathis et al., 2018), CPM (Wei et al., 2016), Hourglass (Newell et al., 2016), HRNet-W32 (Sun et al., 2019), HRNet-W48, SimpleBaseline with ResNet101 (Xiao et al., 2018), and SimpleBaseline with ResNet152) detect the keypoints of a single primate given the bounding box. In contrast, the bottom-up models (HigherHRNet-W32 (Cheng et al., 2020) and HigherHRNet-W48) localize the landmarks without a bounding box and group them to form poses, specialized for multi-primate detection. For all models, we use their own pretrained model and training procedural protocol, i.e., the DeepLabCut model is pretrained on ImageNet. The top-down models, in general, show stronger performance because of resolution while it shows weaker performance when multiple primates are present. Table 2 summarizes the normalized MPJPE of each landmark in the testing dataset predicted by six models across models. Table 3 reports the PCK@0.2 of each landmark in the testing dataset, and Fig. 9c shows the PCK curve of each model. In short, there is no clear winner. All models use a variant of high capacity convolutional neural networks that can effectively memorize and generalize the training data through fully supervised learning. SimpleBaseline (Xiao et al., 2018) slightly outperforms other models (the lowest MPJPE and the highest PCK@0.2). Fig. 9d shows AP comparison as a function of the model parameters. In general, when the number of data is sufficiently large, larger and deeper models outperform small and shallow models because more complex visual patterns can be learned. Table 4 reports the performance of each model based on the OKS of each landmark. Table 5 reports the PCK@0.1 of each landmark for models trained on varying training sets.

One of the major characteristics of OpenMonkeyChallenge data is a wide range of poses across diverse species. Each species has at least more than 100 annotated images as shown in Fig. 2c. We evaluate the model performance for each species using PCK metrics. In Fig. 10, we plot the accu-

Table 3 Model comparison with PCK@0.2 metric of each landmark with top-down and bottom-up methods on the OpenMonkeyChallenge test set

Method	Eye	Nose	Head	Neck	Shoulder	Elbow	Wrist	Hip	Knee	Ankle	Tail	Mean
Top-down												
DeepLabCut	0.938	0.936	0.926	0.922	**0.907**	**0.875**	0.812	0.855	0.865	0.818	0.747	0.871
CPM	0.995	0.994	0.960	**0.945**	0.887	0.847	0.800	**0.892**	0.909	0.878	0.809	0.896
Hourglass	0.997	0.996	0.960	0.925	0.852	0.830	0.836	0.869	0.896	0.872	0.814	0.890
HRNet-W48	0.997	0.996	0.951	0.940	0.872	0.857	**0.855**	0.885	0.908	**0.885**	**0.842**	**0.903**
HRNet-W32	0.997	0.996	0.958	0.934	0.867	0.851	0.836	0.867	**0.910**	0.876	0.830	0.897
SimpleBaseline(ResNet152)	**0.997**	**0.996**	0.954	0.942	0.866	0.857	0.849	0.883	0.907	0.881	0.838	0.901
SimpleBaseline (ResNet101)	0.995	0.994	**0.983**	0.877	0.829	0.787	0.776	0.827	0.884	0.848	0.756	0.863
Bottom-up												
HigherHRNet-W32	0.971	0.981	0.978	0.856	0.730	0.672	0.793	0.710	0.898	0.812	0.633	0.818
HigherHRNet-W48	0.986	0.985	0.965	0.860	0.759	0.754	0.826	0.779	0.869	0.831	0.715	0.844

The values represented in bold signify the performance of the best model for a given metric/landmark

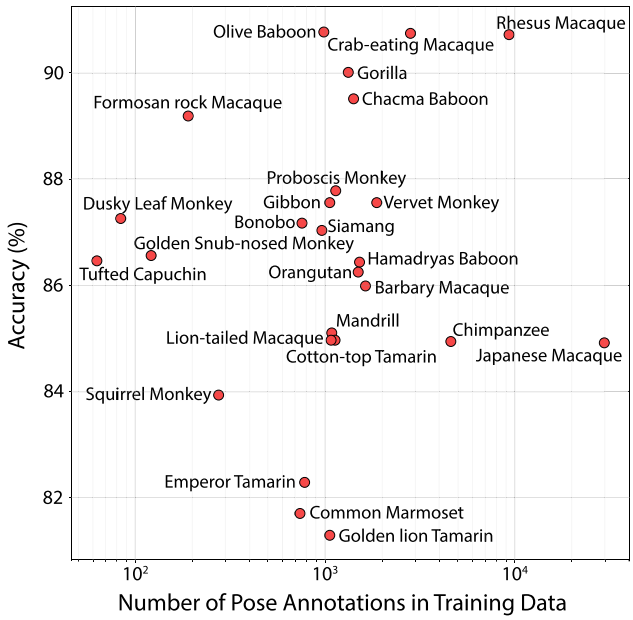


Fig. 10 We show the accuracy for different species with respect to the number of pose annotations in the training set

racy for each species, observed using HRNet, with respect to the number of pose annotations in the training set. We see that our dataset is able to predict the different species with an accuracy greater than 80%. The variety in species and annotations observed in the training dataset does aid in improving pose accuracy across species.

6 Discussion

Here we present a new resource, a very large (111,529 images of 26 species) and fully annotated database of photographs of non-human primates. The primates come in a range of species and poses, and with a range of backgrounds. The primary goal of this resource is to serve as a training tool for scholars interested in developing computer vision approaches to identifying pose in the primate order. This resource can be found on our new website (http://openmonkeychallenge.com). The website also presents a new benchmark challenge for primate landmark detection. In parallel with our resource and the challenge, and as a baseline for modeling efforts, we provide some analyses of existing models. These analyses reveal that non-human primate detectors have substantially worse performance than human ones. We propose that our large dataset will be a critical tool in closing that performance gap.

We know of only two existing large datasets of annotated primate images, OpenMonkeyPose (Bala et al., 2020) and MacaquePose (Labuguen et al., 2021). OpenMonkeyPose, which our group developed, consists of nearly 200,000 annotated (13 landmarks) multiview (62 cameras) images of rhesus macaques in a specific carefully controlled laboratory environment. That dataset has a very different purpose than the present one—its chief virtue is its robust characterization of a single environment and species, and its multiview aspect for 3D motion capture. However, it is highly limited for the general purpose of pose identification because of its narrow number of backgrounds, species, individuals, and poses. The MacaquePose dataset, which consists of 13,000 images, is likewise limited to a single species and is also substantially smaller. Our analyses confirm that these datasets cannot be used to train robust models that can identify pose in general contexts nearly as well as this one can. These results, then, argue for the value of large variegated datasets like the one we present here. More generally, they demonstrate the critical importance of variety when training robust detection networks.

Table 4 Model comparison with AP metric based on OKS of each landmark with top-down and bottom-up methods on the OpenMonkeyChallenge test set

Method	# Params	AP@0.5	AP@0.6	AP@0.7	AP@0.8	AP@0.9	AP
Top-down							
DeepLabCut	24.4M	92.3	**89.7**	**83.9**	74.1	52.6	73.2
CPM	31.4M	91.8	86.1	78.9	69.6	54.2	72.9
Hourglass	21M	91.3	85.7	80.8	74.7	63.8	74.5
HRNet-W32	28.5M	89.4	80.6	71.0	64.6	65.7	70.7
HRNet-W48	63.6M	90.2	85.7	80.8	74.7	63.8	76.5
SimpleBaseline (ResNet152)	68.0M	89.5	84.9	81.2	**76.9**	**67.8**	**78.5**
SimpleBaseline (ResNet101)	53.0M	**97.2**	82.6	65.9	46.9	31.4	65.3
Bottom-up							
HigherHRNet-W32	28.6M	88	79.5	57.3	32.0	20.0	59.1
HigherHRNet-W48	63.8M	91.5	82.6	65.9	46.9	31.4	65.3

The values represented in bold signify the performance of the best model for a given metric/landmark

Table 5 Model comparison with different metrics of each landmark obtained from models trained on varying training sets

Method	t	Nose	Shoulder	Wrist	Knee	Mean
Train: 100%OMC; Test: OMC	PCK@0.1	0.971	0.758	0.786	0.736	0.778
Train: 100%OMC+pretrained; Test: OMC	PCK@0.1	0.975	0.769	0.806	0.754	0.793
Train: 100%OMC+OMS+MP; Test: OMC	PCK@0.1	0.954	0.713	0.757	0.683	0.739
Train: 100%COCO; Test: COCO	PCK@0.1	0.975	0.892	0.888	0.859	0.894
Train: 100%COCO; Test: OMC	PCK@0.1	0.635	0.212	0.199	0.180	0.239
Train: 100%OMC; Test: OMC	PCKh@0.1	0.915	0.568	0.661	0.576	0.633
Train: 100%OMC+pretrained; Test: OMC	PCKh@0.1	0.913	0.553	0.672	0.574	0.651
Train: 100%OMC+OMS+MP; Test: OMC	PCKh@0.1	0.830	0.529	0.599	0.517	0.575
Train: 100%COCO; Test: COCO	PCKh@0.1	0.884	0.652	0.653	0.595	0.685
Train: 100%COCO; Test: OMC	PCKh@0.1	0.491	0.168	0.133	0.136	0.169
Train: 100%OMC; Test: OMC	PCKd@0.1	0.989	0.911	0.878	0.877	0.895
Train: 100%OMC+pretrained; Test: OMC	PCKd@0.1	0.990	0.918	0.888	0.888	0.904
Train: 100%OMC+OMS+MP; Test: OMC	PCKd@0.1	0.984	0.898	0.867	0.847	0.878
Train: 100%COCO; Test: COCO	PCKd@0.1	0.994	0.971	0.956	0.956	0.967
Train: 100%COCO; Test: OMC	PCKd@0.1	0.502	0.586	0.316	0.290	0.364

A key finding from our comparative study is that the state-of-the-art designs of convolutional neural networks (CNNs), including DeepLabCut, perform, by large, on a par with each other. These CNNs effectively learn a visual representation of primates from sufficiently large and diverse image data in a fully supervised manner where generalizable image features can be learned. This closes the gap between models. On the other hand, this finding implies that there is a fundamental limitation to the supervised learning paradigm. That is, our results indicate that the CNN models overfit to the training data; the distribution of the training data differs considerably from that of the testing data. As a consequence, the generalization is strictly bounded, which leaves a large performance gap between human and primate landmark detections. This requires employing the new semi- or unsupervised learning paradigm, which allows utilizing a potentially unlimited amount of unlabeled, or weakly labeled primate images, which can close the domain difference.

Through the OpenMonkeyChallenge, we aim to derive two major innovations to solve challenging computer vision problems. First, algorithmic innovation can lead to substantial performance gain by learning an efficient representation from limited annotated data. Transfer learning, or domain adaptation, used in DeepLabCut is one of such kinds that leverage a pre-trained generic model learned from a large dataset (e.g., ImageNet). Such approaches have shown a remarkable generalization over frames within a target video while showing limited performance when applying to new videos with different viewpoints, poses, illumination, background, and identities. Second, data innovation can lead

to great advances in generalization by being agnostic to algorithms and representations. For example, the field has witnessed such gains from the object detection community, e.g., from a few hundreds of images in Caltech-101 and Pascal VOC datasets to millions of images in ImageNet and COCO datasets (Torralba & Efros, 2011). OpenMonkeyChallenge facilitates these two indispensable innovations for developing a generalizable primate detector through community effort.

Acknowledgements We thank Lin Huynh, Peeyush Samba, Justin Aronson, and Jen Holmberg for help on image acquisition. We thank the staff at the Minnesota Zoo for copious help, especially Tom Ness, Kathy Schlegel, Jamie Toste, Laurie Trechsel, and Kelli Gabrielson.

Funding This work is partially supported by NSF IIS 2024581 (HSP, JZ, and BYH), NIH P51 OD011092 (ONPRC), NIH P51 OD011132 (YNPRC), NIH R01-NS120182 (JR), and K99-MH083883 (CJM).

Declarations

Conflict of interest The authors declare no conflicts of interest. All procedures were performed in compliance with the guidelines of the IACUC of the University of Minnesota.

Ethical approval All procedures were performed in compliance with the guidelines of the IACUC of the University of Minnesota.

Informed Consent Informed consent is not relevant because there were no human subjects.

References

Andriluka, M., Pishchulin, L., Gehler, P., & Schiele, B. (2014). 2D human pose estimation: New benchmark and state of the art analysis. In Computer Vision and Pattern Recognition.

Andriluka, M., Iqbal, U., Milan, A., Insafutdinov, E., Pishchulin, L., Gall, J., & Schiele, B. (2018). Posetrack: A benchmark for human pose estimation and tracking. In Computer Vision and Pattern Recognition.

Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Zitnick, C. L., & Parikh, D. (2015). VQA: Visual question answering. In International Conference on Computer Vision.

Bala, P., Eisenreich, B., Yoo, S. B., Hayden, B., Park, H., & Zimmermann, J. (2020). Automated markerless pose estimation in freely moving macaques with openmonkeystudio. *Nature Communications*.

Belagiannis, V., & Zisserman, A. (2017). Recurrent human pose estimation. In International Conference on Automatic Face & Gesture Recognition.

Bliss-Moreau, E., Machado, C. J., & Amaral, D. G. (2013). Macaque cardiac physiology is sensitive to the valence of passively viewed sensory stimuli. *PLoS One*.

Cao, Z., Martinez, G. H., Simon, T., Wei, S.-E., & Sheikh, Y. A. (2019). Openpose: Realtime multi-person 2d pose estimation using part affinity fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Cheng, B., Xiao, B., Wang, J., Shi, H., Huang, T. S., & Zhang, L. (2020). Higherhrnet: Scale-aware representation learning for bottom-up human pose estimation. In Computer vision and pattern recognition.

de Bem, R., Arnab, A., Golodetz, S., Sapienza, M., & Torr, P. H. S. (2018). Deep fully-connected part-based models for human pose estimation. In Asian conference on machine learning.

Dunn, T., Marshall, J., Severson, K., Aldarondo, D., Hildebrand, D., Chettih, S., Wang, W., Gellis, A., Carlson, D., Aronov, D., Freiwald, W., Wang, F., & Olveczky, B. (2021). Geometric deep learning enables 3D kinematic profiling across species and environments. *Nature Methods*.

Eichner, M., & Ferrari, V. (2010). We are family: Joint pose estimation of multiple persons. In European Conference on Computer Vision.

Fang, H.-S., Xie, S., Tai, Y.-W., & Lu, C. (2017). RMPE: Regional multi-person pose estimation. In International conference on computer vision.

Güler, R. A., Neverova, N., & Kokkinos, I. (2018). Densepose: Dense human pose estimation in the wild. In Computer vision and pattern recognition.

Günel, S., Rhodin, H., Morales, D., Campagnolo, J., Ramdya, P., & Fua, P. (2019). Deepfly3d, a deep learning-based approach for 3d limb and appendage tracking in tethered, adult drosophila. *eLife*.

Hayden, B. Y., Park, H. S., & Zimmermann, J. (2021). Automated pose estimation in primates. *American Journal of Primatology*.

Insafutdinov, E., Pishchulin, L., Andres, B., Andriluka, M., & Schiele, B. (2016). Deepercut: A deeper, stronger, and faster multi-person pose estimation model. In European conference on computer vision.

Iqbal, U., Milan, A., & Gall, J. (2017). Posetrack: Joint multi-person pose estimation and tracking. In Computer vision and pattern recognition.

Iskakov, K., Burkov, E., Lempitsky, V., & Malkov, Y. (2019). Learnable triangulation of human pose. In International conference on computer vision.

Jakab, T., Gupta, A., Bilen, H., & Vedaldi, A. (2020). Self-supervised learning of interpretable keypoints from unlabelled videos. In Computer vision and pattern recognition.

Karashchuk, P., Rupp, K., Dickinson, E., Sanders, E., Azim, E., Brunton, B., & Tuthill, J. (2020). Anipose: A toolkit for robust markerless 3D pose estimation. In BioRxiv.

Knaebe, B., Weiss, C., Zimmermann, J., & Hayden, B. (2022). The promise of behavioral tracking systems for advancing primate animal welfare. *Animals*.

Labuguen, R., Matsumoto, J., Negrete, S., Nishimaru, H., Nishijo, H., Takada, M., Go, Y., Inoue, K.-I., & Shibata, T. (2021). Macaquepose: A novel "in the wild" macaque monkey pose dataset for markerless motion capture. *Frontiers in Behavioral Neuroscience*.

Li, S., Li, J., Tang, H., Qian, R., & Lin, W. (2020). ATRW: A benchmark for amur tiger re-identification in the wild. In ACM International Conference on Multimedia.

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 European conference on computer vision.

Lin, W., Liu, H., Liu, S., Li, Y., Qian, R., Wang, T., Xu, N., Xiong, H., Qi, G.-J., & Sebe, N. (2020). Human in events: A large-scale benchmark for human-centric video analysis in complex events. *arXiv preprint* arXiv:2005.04490.

Ludwig, K., Scherer, S., Einfalt, M., & Lienhart, R. (2021). Self-supervised learning for human pose estimation in sports. In IEEE International Conference on Multimedia Expo Workshops.

Machado, C. J., Bliss-Moreau, E., Platt, M. L., & Amaral, D. G. (2011). Social and nonsocial content differentially modulates visual attention and autonomic arousal in rhesus macaques. *PLoS One*.

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*.

Mathis, M. W., & Mathis, A. (2020). Deep learning tools for the measurement of animal behavior in neuroscience. *Current Opinion in Neurobiology*.

McInnes, L., Healy, J., & Melville, J. (2018). Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv*.

Newell, A., Yang, K., & Deng, J. (2016). Stacked hourglass networks for human pose estimation. In European conference on computer vision.

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 Computer vision and pattern recognition.

Pereira, T. D., Aldarondo, D. E., Willmore, L., Kislin, M., Wang, S. S. H., Murthy, M., & Shaevitz, J. W. (2018). Fast animal pose estimation using deep neural networks. *Nature Methods*.

Pishchulin, L., Insafutdinov, E., Tang, S., Andres, B., Andriluka, M., Gehler, P., & Schiele, B. (2016). Deepcut: Joint subset partition and labeling for multi person pose estimation. In Computer vision and pattern recognition.

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

Ren, Z., & Lee, Y. J. (2018). Cross-domain self-supervised multi-task feature learning using synthetic imagery. In computer vision and pattern recognition.

Sade, D. S. (1973). An ethogram for rhesus monkeys i. Antithetical contrasts in posture and movement. *American Journal of Physical Anthropology*.

Sapp, B., & Taskar, B. (2013). Modec: Multimodal decomposable models for human pose estimation. In Computer vision and pattern recognition.

Sumer, O., Dencker, T., & Ommer, B. (2017). Self-supervised learning of pose embeddings from spatiotemporal relations in videos. In International conference on computer vision.

Sun, K., Xiao, B., Liu, D., & Wang, J. (2019). Deep high-resolution representation learning for human pose estimation. In Computer vision and pattern recognition.

Torralba, A., & Efros, A. A. (2011). Unbiased look at dataset bias. In Computer vision and pattern recognition.

Toshev, A., & Szegedy, C. (2014). Deeppose: Human pose estimation via deep neural networks. In Computer vision and pattern recognition.

von Marcard, T., Henschel, R., Black, M., Rosenhahn, B., & Pons-Moll, G. (2018). Recovering accurate 3D human pose in the wild using imus and a moving camera. In European conference on computer vision.

Wan, C., Probst, T., Gool, L. V., & Yao, A. (2019). Self-supervised 3D hand pose estimation through training by fitting. In Computer vision and pattern recognition.

Wei, S.-E., Ramakrishna, V., Kanade, T., & Sheikh, Y. (2016). Convolutional pose machines. In Computer vision and pattern recognition.

Wiltschko, A., Johnson, M., Iurilli, G., Peterson, R., Katon, J., Pashkovski, S., Abraira, V., Adams, R., & Datta, S. (2015). Mapping sub-second structure in mouse behavior. *Neuron*.

Xiao, B., Wu, H., & Wei, Y. (2018). Simple baselines for human pose estimation and tracking. In European conference on computer vision.

Yang, H., Dong, W., Carlone, L., & Koltun, V. (2021). Self-supervised geometric perception. In Computer vision and pattern recognition.

Yao, Y., Jafarian, Y., & Park, H. S. (2019). Monet: Multiview semi-supervised keypoint via epipolar divergence. In International Conference on Computer Vision.