Received: 11 April 2022 | Accepted: 22 March 2023

DOI: 10.1111/1365-2656.13932

Journal of Animal Ecology | BRITISH ECOLOGICAL SOCIETY

RESEARCH ARTICLE

Contemporary Methods for Studying Animal Sociality in the Wild

DeepWild: Application of the pose estimation tool DeepLabCut for behaviour tracking in wild chimpanzees and bonobos

Charlotte Wiltshire[1] | James Lewis-Cheetham[1] | Viola Komedová[1] | Tetsuro Matsuzawa[2,3] | Kirsty E. Graham[1] | Catherine Hobaiter[1]

[1]Wild Minds Lab, School of Psychology and Neuroscience, University of St Andrews, St Andrews, UK

[2]Department of Pedagogy, Chubu Gakuin University, Gifu, Japan

[3]Division of the Humanities and Social Sciences, California Institute of Technology, Pasadena, California, USA

Correspondence
Catherine Hobaiter
Email: clh42@st-andrews.ac.uk

Funding information
Horizon 2020 Framework Programme, Grant/Award Number: 802719; St Andrews Restarting Research Funding Scheme

Handling Editor: Thibaud Gruber

Abstract

1. Studying animal behaviour allows us to understand how different species and individuals navigate their physical and social worlds. Video coding of behaviour is considered a gold standard: allowing researchers to extract rich nuanced behavioural datasets, validate their reliability, and for research to be replicated. However, in practice, videos are only useful if data can be efficiently extracted. Manually locating relevant footage in 10,000s of hours is extremely time-consuming, as is the manual coding of animal behaviour, which requires extensive training to achieve reliability.
2. Machine learning approaches are used to automate the recognition of patterns within data, considerably reducing the time taken to extract data and improving reliability. However, tracking visual information to recognise nuanced behaviour is a challenging problem and, to date, the tracking and pose-estimation tools used to detect behaviour are typically applied where the visual environment is highly controlled.
3. Animal behaviour researchers are interested in applying these tools to the study of wild animals, but it is not clear to what extent doing so is currently possible, or which tools are most suited to particular problems. To address this gap in knowledge, we describe the new tools available in this rapidly evolving landscape, suggest guidance for tool selection, provide a worked demonstration of the use of machine learning to track movement in video data of wild apes, and make our base models available for use.
4. We use a pose-estimation tool, DeepLabCut, to demonstrate successful training of two pilot models of an extremely challenging pose estimate and tracking problem: multi-animal wild forest-living chimpanzees and bonobos across behavioural contexts from hand-held video footage.
5. With DeepWild we show that, without requiring specific expertise in machine learning, pose estimation and movement tracking of free-living wild primates in visually complex environments is an attainable goal for behavioural researchers.

KEYWORDS
artificial intelligence, automation, behaviour, deep learning, machine learning, primate

1 | INTRODUCTION

Studying animal behaviour allows us to understand how different individuals navigate their physical and social worlds, and cross-species comparisons can provide insight into the evolutionary trajectory of behavioural capacities. Video recordings provide particularly abundant and robust data collection, allowing the extraction of many kinds of behaviour: from social organisation, to communication, to movement and so forth. Unlike direct observation, there is the opportunity to repeatedly revisit the same events—allowing researchers to explore new questions, and to improve or validate data collection on existing ones. As a result, video coding is considered a gold standard: allowing rich nuanced behavioural datasets, on which research can be conducted and replicated by others now and in the future. As well as targeted novel video data collection, many research groups have established large video archives from which we can extract data (e.g. Arandjelovic et al., 2016; Bain et al., 2021; Burton et al., 2015; Hobaiter et al., 2021; Schofield et al., 2019). These archives represent data arks: stable, long-term resources that help us continue to address scientific questions in taxon, such as primates, who are in catastrophic population decline (Estrada et al., 2017). In addition, these digital resources help address the systemic financial and physical barriers related to collecting behavioural data in the wild, opening up scientific research to a more diverse pool of researchers.

However, in practice, these videos are only useful if data can be efficiently extracted from them. Manually locating relevant footage in hundreds or thousands of hours of archival material is extremely time-consuming, as is the subsequent manual coding of animal behaviour, which requires extensive training to achieve reliability and limit coder error and bias (Munch et al., 2019; Pathak et al., 2003). Where the time-burden and its associated financial costs outweighs those of novel data collection, these potentially invaluable archives sit unused.

Machine learning approaches are used to automate the recognition of patterns within data (Hastie et al., 2001) and can considerably reduce the time taken to extract data, while improving the reliability of results (Schofield et al., 2019). They have been successfully employed across diverse behavioural datasets from acoustics (Bianco et al., 2019), to taxonomy (Wäldchen & Mäder, 2018), to movement (e.g. flies: Günel et al., 2019; robots and humans: Islam et al., 2021; fish: Mei et al., 2021; mice: Sheppard et al., 2022). More recently, there has been considerable success extending these to visual data, with a sudden explosion of tools for automating pattern recognition within photographic and video data. While early algorithms focused on photographic data (for a thorough review, see Weinstein, 2017), their extension to video data is particularly relevant for behavioural research, allowing the capture of information over time.

One widespread use of machine learning with video data is for species recognition, for example from camera-trap data (see Table S1 for examples). Camera-traps can be deployed in large numbers, across wide areas, and left to capture data 24 h a day. With the appropriate considerations (Swann et al., 2011), they allow for the monitoring of species and individuals who are typically not easily observable in-person: populations not habituated to direct observation, or who are sparse or nocturnal. Camera-traps are an effective means to address questions investigating species presence, abundance, and diversity, as well as to monitor distribution and density over time within and between locations (Steenweg et al., 2016). However, when used at scale, they create vast amounts of video data that can be extremely time-consuming to decode. In one example, manually sorting camera trap data on wolf monitoring had a lag of approximately 5 years (Tuia et al., 2022). With the use of machine learning (Microsoft AI4Earth MegaDetector: Beery et al., 2019) all data were labelled within 12 months, allowing data to be reviewed before the start of the next monitoring season (Tuia et al., 2022). In another model, identification of individual primates across species could be processed with 94% accuracy at over 30 images per second (Guo et al., 2020).

Automated species identification tools split videos into discrete frames and then examine each one to perform initial triage by filtering out blank images (AIDE: Kellenberger et al., 2020; Wildlife Insights: Ahumada et al., 2019; Microsoft AI4Earth MegaDetector: Beery et al., 2019), followed by the identification of species (e.g. Narouzzadeh et al., 2018; Willi et al., 2018; Yu et al., 2013; AIDE: Kellenberger et al., 2020; Wildlife Insights: Ahumada et al., 2019; Whytock et al., 2021) and even individuals, known as 'individual reidentification' (Wildbook: Berger-Wolf et al., 2017; Guo et al., 2020; Schofield et al., 2019) that were captured. Current automated tools perform this work so quickly that they are used to send out real-time alerts if humans or unknown vehicles are unexpectedly present in protected areas, offering rapid-response opportunities for conservation teams on the ground (wpsWatch: Tuia et al., 2022). Similar systems inform local communities of the approach of potentially dangerous wildlife, such as elephants (Premarathna et al., 2020). Individual identification can present a particularly challenging problem for human coders. In a study of 23 chimpanzees containing approximately a million images, human annotators given 1–2 h of exposure reached only 20% (novices) and 42% (experts) accuracy. A model trained on the same data took only a matter of seconds to achieve 84% accuracy (Schofield et al., 2019), let alone subsequent savings in processing of the vast data set. Similarly accurate models for individual identification are now available for a growing number of taxa (tigers: Li et al., 2020, elephants: Körschens & Denzler, 2019, cattle: Bergamini et al., 2018, primates: Guo et al., 2020; for a review see: Schneider et al., 2018). Nevertheless, there is typically a need for substantial upfront investment in the development of training sets, and subsequent tools are typically population-, and even site-, specific limiting their generalisability.

While analysing photographs (or discrete frames from a video) already provides a powerful approach, the study of many behaviours requires the extraction of data across video frames—restoring the component of time. One recent application of this in wild chimpanzees employs a pipeline that moves from detection and tracking of chimpanzees, to individual identification, to identification of behavioural categories, such as feeding (Bain et al., 2021). However,

doing so requires a choice, in advance, of the behaviour of interest, and—to date—has only been demonstrated in two behaviours with distinctive auditory, as well as visual, signatures (nut cracking and drumming). An alternative approach to behaviour categorisation is pose estimation and movement tracking: here, individual points are marked on the body and their relative location is tracked across frames. These too require a decision in advance, here the number and placement of the landmarks on the body, but the same base model can be used to generate coordinate data for kinematic analyses of tool use movements and gestural actions, although doing so would likely require separate behavioural segmentation tools in a subsequent step. A potentially powerful approach is to combine them using spatiotemporal action CNNs (Convolutional Neural Networks; Achour et al., 2020), which retain some information on the broader visual context in which behaviour is situated, with pose-estimation approaches that provide refined kinematic analysis of particular actions. A full list of machine learning based tracking tools, with information on their uses and functionality, is available in Table S2.

In some cases, these tools track the location of individual animals relative to each other and their environment, allowing, for example, the detailed study of group movements of hundreds of individuals in synchrony (Walter & Couzin, 2021). The most recent generation of tracking tools provides pose estimation by tracking multiple points on an individual (for a full list of pose estimation tools, their usability, and their functionality, see Table S3). Doing so offers flexibility in which behaviour are tracked, and the opportunity to analyse movement within behaviour in substantial detail (see Panadeiro et al., 2021 for an in-depth summary). For example, allowing the study of facial expressions (Wang & Lien, 2009) or gait analysis (Rohan et al., 2020) in humans (Khan & Wan, 2018; Sarafianos et al., 2016). But with the recent arrival of 'plug and play' software that incorporates user-friendly non-coding-based interfaces, there has been an explosion of interest in the wider field of animal behaviour (Panadeiro et al., 2021; Tuia et al., 2022).

There are obvious reasons why—while the use of video-based data extraction is a powerful method for robust studies of animal behaviour, manual coding is extremely time consuming, and—even with substantial training periods—experienced researchers are still subject to some human error. Even relatively 'simple' problems, such as the marking of two points (e.g. in lip-smacking; Pereira, Kavanagh, et al., 2020) still requires that these points are manually marked on every frame, and with typical frame rates of 25fps and behaviour that are measured in minutes, doing so is a substantial time investment—often months of work. With an appropriate model, machine learning tools can extract the same data in a matter of minutes or seconds. There are of course substantial caveats—appropriate models are rarely available 'off the shelf' (although cf. the DeepLabCut 'model zoo', Kane et al., 2020). And, as in the case of manual coding, these models typically provide you with raw data output (*x*-*y* coordinates for each marked point within the frame) that need substantial further processing to be translated into behavioural categories or measures. For example: performing gait analysis on the co-ordinates to extract walking rhythm (Prakash et al., 2018). However, open-source machine learning tools that classify behaviour from coordinates are emerging from laboratory-work (e.g. Hsu & Yttri, 2021), and may soon have the potential to be expanded to wild data.

Tracking visual information in such detail is a challenging problem, and, to date, tracking algorithm tools are typically applied within laboratory studies where the environment is fixed and/or controlled, and tend to have been developed for model animal species in widespread use such as mice (e.g. drosophila: Yu et al., 2011; rodents: Geuther et al., 2019; ants: Gal et al., 2020; worms: Kiel et al., 2018; fish: Xu & Cheng, 2017; for a summary of currently available software tools see Table S2). Recent advances include 3-dimensional descriptions of an individual's movements in its environment. Doing so requires at least two static camera angles that can be used to provide the depth estimation needed to re-create objective distances (without this the distance between any two points in a single frame is arbitrary; are they small or are they far away? Although, cf. Haucke et al., 2021). The range of species has also started to expand, moving from laboratory model species, e.g. mice (Gosztolai et al., 2021; Karashchuk et al., 2021), flies (Gosztolai et al., 2021; Günel et al., 2019; Karashchuk et al., 2021) to include primates and larger mammals (macaques: Bala et al., 2020; Gosztolai et al., 2021; Marks et al., 2022; cheetahs: Nath et al., 2019).

Studying social behaviour requires tracking of more than one individual—doing so requires more than a simple extension of the single individual method. The model must be able to not only track body parts, but also keep track of to whom those body parts belong to (i.e. elbow A belongs to individual A, even when they swap places with individual B or C). As a result, additional time investments are needed in training to manually correct accidental body part swaps (Mathis et al., 2018; Pereira et al., 2019; Pereira, Tabris, et al., 2020), however, these may be more than offset by the subsequent ability to automate rapid data generation across many individuals. Some tools, such as TRex (Walter & Couzin, 2021), focus on tracking across very large numbers of individuals, for example looking at flock, herd, or school movements; others, such as the multi-animal tracking options in SLEAP, can track discrete body parts across moderate numbers of individuals (i.e. <100; Pereira et al., 2019; Pereira, Tabris, et al., 2020).

Until recently, pose estimation software was limited to laboratory and, increasingly, for domestic animals and pets (Kane et al., 2020), and in captive environments such as zoos (Hayden et al., 2021; Marks et al., 2022). In these environments the 'visual noise' is both relatively low and stable across videos. Just as for humans, it is much easier for machine learning tools to detect an animal moving when nothing else in the frame is moving, or an animal on a plain background with good lighting. However, with increasingly sophisticated software capable of learning across multiple individuals and in more variable conditions, pose-estimation tools could finally be extended to wild populations. Doing so offers substantial power to researchers exploring behavioural variation across a wide variety of disciplines: from ecology to cognition, from conservation to culture.

While there appears to be significant interest in trying to do so, with so many different machine learning tools available, it can be overwhelming to know which are suitable for different types of data and questions. Recent summaries are available for laboratory

(Panadeiro et al., 2021) and conservation (Tuia et al., 2022) applications, but less information is available for behavioural scientists who work with wild populations. The decision on which tool to employ can be approached by considering a few key questions (Figure 1; for an up-to-date list of available software see Tables S2 and S3).

In this paper we take one of the leading tools currently available, DeepLabCut (Mathis et al., 2018), and provide a worked example of its functionality with a particularly challenging dataset: that of wild chimpanzee and bonobo video. We do so from the perspective of a group of animal behaviour researchers, with substantial experience in working with manual coding of nuanced behaviour extraction from video, but only basic skills in machine learning.

Initially developed for use in mouse and drosophila behavioural tracking (Mathis et al., 2018), DeepLabCut has since been applied to a wide range of other species (rats, Clemensson et al., 2020; fish, Habe et al., 2021; cheetahs, Joska et al., 2021; horses, Tsuruo et al., 2020). DeepLabCut offers multi-animal pose estimation, a straightforward user-friendly graphical interface, and example tracking videos. Extracting visual data from video of wild, forest-living apes may be among the most challenging of tasks for machine learning: the apes move freely in all three dimensions of their environment, we move as we follow the apes, our hand-held cameras move, the lighting is often dark but can include dramatic contrast—with dark apes, in a dark forest, backlit against bright skies. And finally, the forests themselves are visually dense—with many visual obstacles (branches, trees, leaves, other apes) that themselves move. In addition, here we train a model that includes individuals from two closely related but physically distinct species: bonobos and chimpanzees, including two subspecies of chimpanzee (East and West African) and individuals of all age sex classes, as well as populations living in different habitat types. A typical decision researchers must make is whether to increase the training set size in a set time frame by having multiple people mark frames. While doing this increases the size of the training set, it introduces a new aspect of noise in the data: inter-marker variation. We provide a basic example of this trade-off by training a second model. Model 2 replicates Model 1, but includes additional frames in the training set (27% increase) marked by a second marker. We provided what would be considered minimalistic training sets (<2000 frames; cf with, for example, 195,228 frames used to create OpenMonkeyStudio, a pose estimator for captive primates; Bala et al., 2020), representing ~100–140 human coder hours to produce. As a result, our data and findings likely represent an outlier in terms of task difficulty: in essence, if our model, trained on a minimalist set of frames, can perform basic tracking despite the high level of diverse forms of visual noise in these data, it suggests similar models could work for most other primate behavioural video datasets.

2 | MATERIALS AND METHODS

2.1 | DeepLabCut use

Multiple user guides are available for DeepLabCut, including those from the developers (see: DeepLabCut Github, 2021a, 2021b), as well as from users (e.g. Gadea, 2021). Download and installation of DeepLabCut and initial use of the Graphical User Interface (GUI)

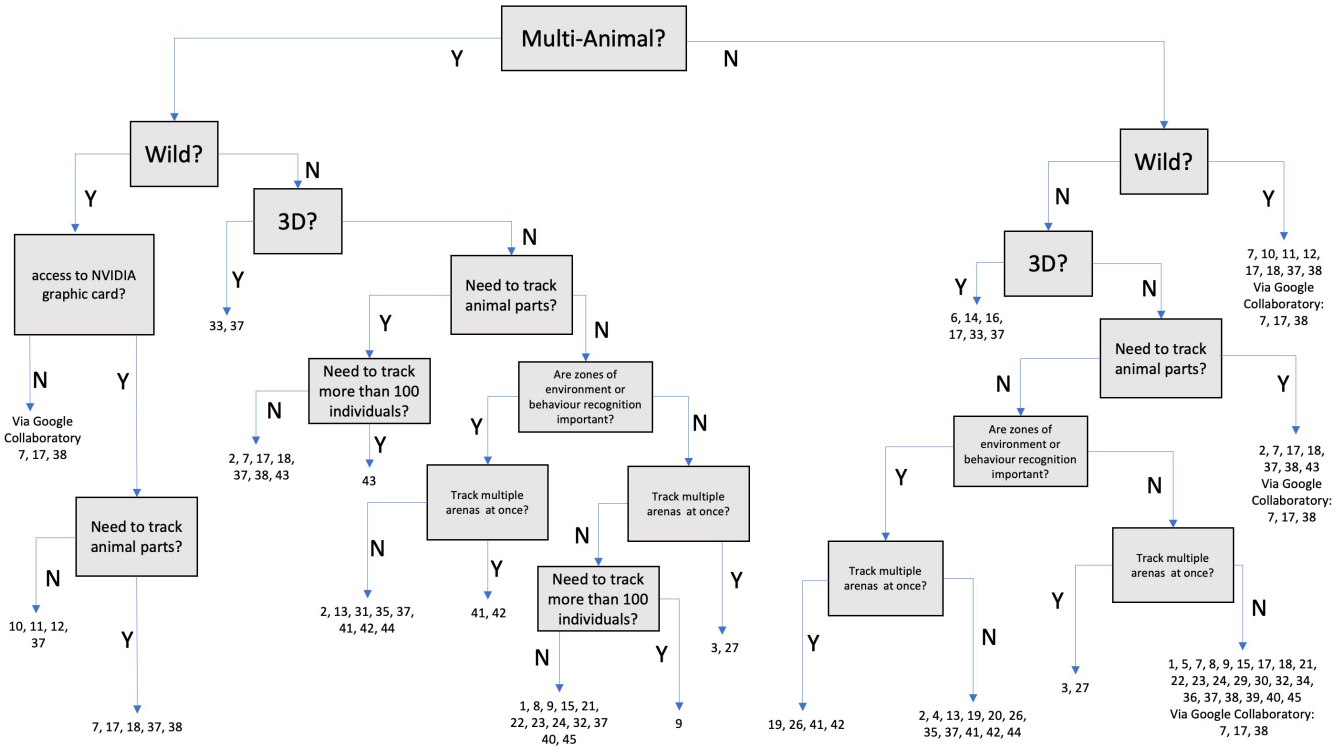

FIGURE 1 Decision tree for software selection. Software are numbered and linked to Table S2, which provides a description of each tool, its previous uses and functionality. Further detailed assessment of the subset of tracking tools that provide pose estimation can be found in Table S3.

requires that users first instal Python. The Anaconda distribution of Python is recommended as it includes useful pre-installed packages. While the DeepLabCut developers do provide instructions for installing Python and DeepLabCut from scratch (DeepLabCut Github, 2021c), it is useful to have a basic understanding of Python or the command terminal of your chosen operating system.

DeepLabCut can be used with or without specialised hardware. A Graphics Processing Unit (GPU) is recommended and decreases training time. However, standard modern computing hardware can be used. Alternatively, Google Colaboratory can be used to access a free cloud-hosted GPU. The DeepLabCut GUI is not available if using Google Colaboratory, and users will need a more substantive knowledge of Python, but example workbooks and tutorials are available on the DeepLabCut Youtube (https://www.youtube.com/channel/UC2HEbWpC_1v6i9RnDMy-dfA).

Once installed and opened, the GUI uses tabs to guide users through the process of creating new projects or opening existing ones. It does not, as yet, offer features such as a 'loss graph' that allows users to track model training progress (cf. SLEAP, Pereira, Tabris, et al., 2020). But a simple bespoke version can be easily generated to assess loss once training is completed (example code available here: https://github.com/Wild-Minds/DeepWild). Loss graphs help users to understand when to terminate model training, as overtraining can lead to overfitting, reducing model performance. DeepLabCut recommends terminating training when the loss plateaus, thus the visual aid of a graph is useful.

2.2 | Data and study subjects

We extracted video data from the Great Ape Dictionary Database (Hobaiter et al., 2021). Videos were recorded between 2013 and 2020, were all originally recorded as either high-definition or 4K footage using handheld Panasonic video camcorders with a frame rate of 25fps (e.g. HCV770 or HCVX-F1). Original video data were collected from one bonobo *Pan paniscus* and four chimpanzee *Pan troglodytes* communities from two subspecies (East African chimpanzees: *Pan troglodytes schweinfurthi*, West African chimpanzees: *Pan troglodytes verus*). While very closely related, the different *Pan* species nevertheless show characteristic differences in morphology and movement (Doran, 1993; Jungers & Susman, 1984).

The bonobo population included was Wamba in the Luo Scientific Reserve in the Democratic Republic of Congo, from which we included two neighbouring groups of bonobos E1- and P-group, who have overlapping ranges and encounter each other frequently. The Wamba communities' habitat is characterised by dry first and secondary forest (Hashimoto et al., 1998; Terada et al., 2015) within anthropogenic habitat (Terada et al., 2015). Three of the four chimpanzee communities were East African chimpanzee communities: Sonso and Waibira are both in the Budongo Forest Reserve, Uganda, and the M-group in the Kalinzu Central Forest Reserve, Uganda. Their habitats are characterised by dense medium-altitude, semi-deciduous, secondary-rainforest growth (Eggeling, 1947). The fourth chimpanzee community was Bossou, in Guinea, a West African chimpanzee community living in forest fragments within anthropogenic habitat and are filmed at an open cleaning that they regularly visit to crack open nuts (Matsuzawa et al., 2011).

2.3 | Ethics

Ethical approval for original data collection and use of the Great Ape Dictionary Database (Hobaiter et al., 2021) was provided by the University of St Andrews Animal Welfare and Ethics Committee (Approval code: PS15842). Ethical approval was provided by both the Uganda Wildlife Authority and the Ugandan National Council for Science and Technology (NS179) for the original data collection of chimpanzee video in Uganda, by the Ministère de la Recherche Scientifique et Technologie, for original data collection of bonobo video in the Democratic Republic of the Congo, and by the Ministre de l'Enseignement Supérieur et de la Recherche Scientifique, and Direction Générale de la Recherche Scientifique et de l'Innovation Technologique for original data collection of chimpanzee video in Guinea.

2.4 | Video selection

Videos were chosen to include as much visual 'noise' as possible. 'Noise' refers to variation that increases the difficulty in discriminating the visual input available for learning, for example noise is generated by variation in behaviour, and the species, age, and sex of the individuals. Noise is also generated by variation such as: uneven lighting, poor lighting, strong contrast, shadows, similarity of colour or texture between individual of interest and the environment, overlapping individuals, occluded body parts, movement of the camera, movement of the individual, movement of the environment. All these increase difficulties in the recognising and tracking of body parts for pose estimation. Given that our data are subject to all of these, often many at once, we trained our model to incorporate representative variation in our training set.

A typical problem researchers face is how many training frames are needed. In a controlled laboratory environment, DeepLabCut can begin tracking with just a few hundred marked frames (Lauer et al., 2021; Mathis et al., 2018), with successful models being created on a few thousand frames for laboratory-based videos (e.g. 1080 frames for dark mice on a plain white background, Mathis et al., 2018). However, the number of training frames needed largely reflects the amount of visual noise in your data. Therefore, large numbers of frames are required for more visually noisy data (e.g. >13,000 frames for macaques in an open zoo-like environment: Labuguen et al., 2021; 7600 frames for multi-animal zoo-housed marmosets: Lauer et al., 2021; 7588 Frames for cheetahs in open savannah, Joska et al., 2021). However, manually marking frames requires a substantial upfront investment in developing the training sets, and there are likely—after a point—diminishing returns in the

trade-off between time invested and increased model accuracy. A second issue is whether to use multiple human coders in establishing the training set—this can substantially reduce the lead-time needed to develop the training set, but can introduce further noise in terms of differences between coders (even trained coders rarely achieve perfect inter-observer reliability, e.g. human variability in pixel RMSE on marking DeepLabCut frames in mice was 3 to 4; Mathis et al., 2018). Here, we provide the models with a minimalist training set (<2000 frames), and train two models to investigate whether increased training set size was offset by multi-coder variability (see Table 1 for summary).

Model 1 contained 1375 training frames from 55 videos. These included two species (bonobos and chimpanzees) from a total of 5 ape communities (Wamba-E1, Wamba-P, Sonso, Waibira, and Bossou), and all training frames were marked by one researcher. Model 2 was an extension of Model 1, with an additional 825 frames from 55 new videos, including marking by a second coder and an additional East African chimpanzee community, Kalinzu M-group (total training set: 2200 frames, 110 videos, 6 ape communities).

2.5 | Video preparation

All videos were limited to a maximum of 90 s to reduce any effect of video length on analysis (for test frames, marking n frames from a total 1000 gives a higher marked to novel ratio within videos than marking *n* frames within a total 10,000). Videos range from 6 to 88 s (mean = 45 s; SD = 22 s). Videos were excluded if more than 7 individuals were present to limit time investment on manual marking (note that even if trained on videos limited to a maximum of five individuals, DeepLabCut models can then track up to 100 individuals in novel videos).

2.6 | Model details

Frames were marked using 18 key-points (Figure 2), which required an average of 2 h per video (10 or 25 frames), although this varies significantly with both levels of visual noise and number of individuals present in the frame. If a key-point was not in view of the camera, it was not marked for that frame.

Training was completed on a ZBook Create G7 with an Intel® Core™ i7-10750H (2.6 GHz base frequency, up to 5.0 GHz with Intel® Turbo Boost Technology, 12 MB L3 cache, 6 cores) and 32 GB DDR4-3200 MHz RAM. We did not deviate from the default options suggested for multi-animal model training. As an additional step, we trained a version of Model 1 on a single Nvidia Tesla V100 card on nodes with an Intel(R) Xeon(R) Gold 6130 CPU @ 2.10GHz to compare training time, given additional computational power. The trained models from this paper are publicly available in our GitHub and archived in Zenodo, see Data Availability Statement for details.

2.7 | Performance

We used mean absolute Euclidean distance to compare the model generated points and human-labelled points. These are produced by calculating the Euclidean distance between the model generated points and human-labelled points, for each detection. For performance on test frames these values are reported by DeepLabCut and are only performed where points were predicted with a probability above p-cut-off of 0.6. For performance on novel video frames these values are reported for all detections made by the model.

2.8 | Use 1: Performance on 'test' frames, DLC model and second human coder

DeepLabCut retains 5% of the manually marked frames users provide as a 'test' set. These are not included in model training and are used for model performance evaluation. Here performance is a comparison between the model derived points and those labelled by the human marker (in this case all marker 1). In addition to this comparison, we compared the performance of a second human by calculating the mean absolute Euclidean distance of the first human markers points with the second human markers points on the same test set.

Note that as frames are taken from some of the same videos as the frames that are used for model training this means the model has experience of the type of visual information it is being tested on. However, providing a model with a training set that includes a subset of frames from all videos that a researcher would like to code still represents a substantial time saving in practice (users mark a maximum of 25 frames per video in setting up training sets, the equivalent of marking ~1 s of video for each video).

2.9 | Use 2: Performance on 'novel' videos

Novel videos, where no frames from the video were included in training, represent a more challenging task. The videos may include

TABLE 1 Summary of training sets used in Model 1 and Model 2. We specify the number of annotators, number of videos from which training frames were extracted, the total number of frames marked for training, the number of species the model was trained on, and the test/train split used when training the model.

	# Annotators	# Videos	# Frames	# Species	# Communities	Train/test split
Model 1	2	55	1375	2	5	95/5
Model 2	1	110	2200	2	6	95/5

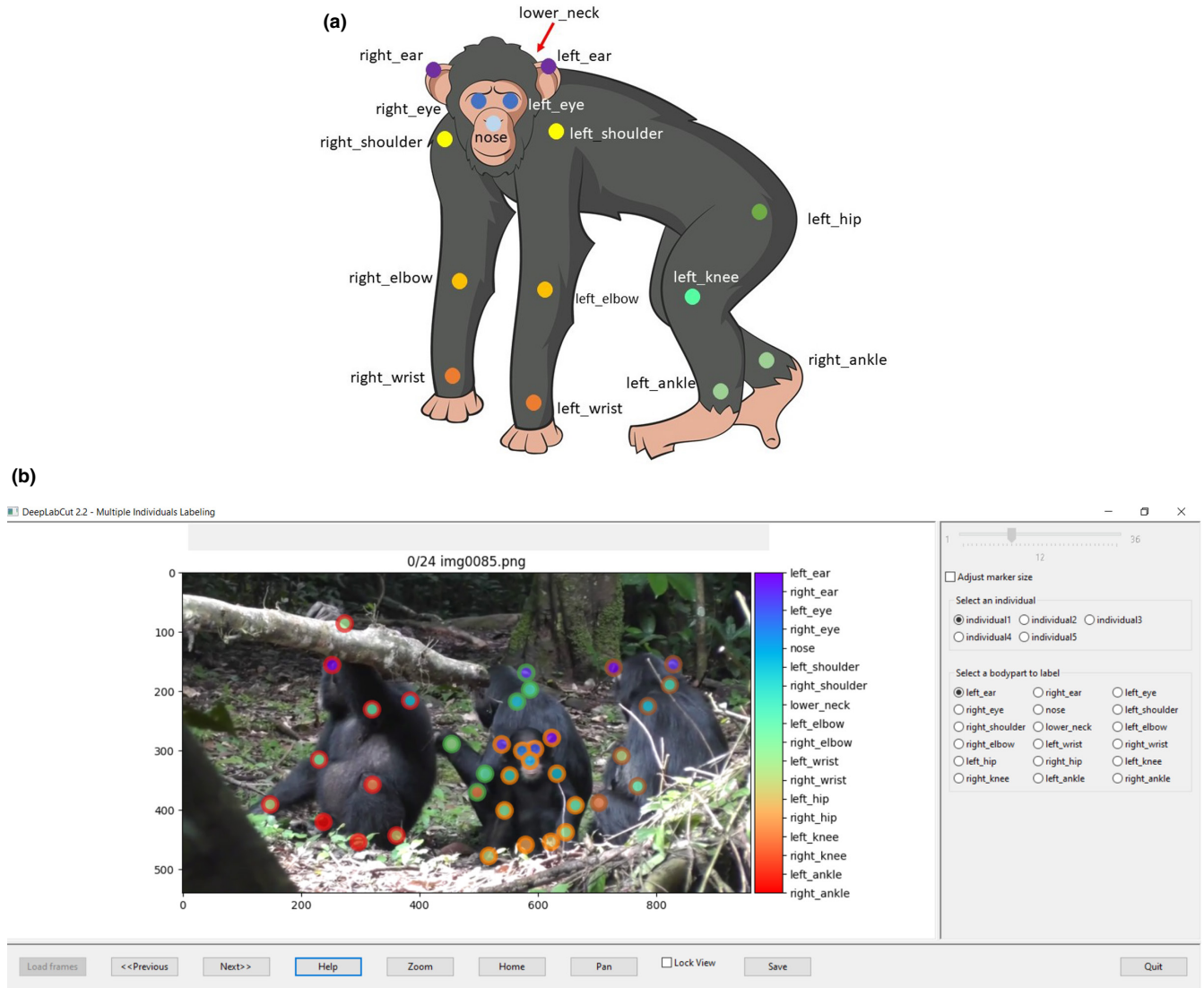

FIGURE 2 (a) The 18 key-points marked. Each of these key points is marked, whenever visible, on each individual within the frame. (b) Example frame from the DeepLabCut Graphical User Interface. Here, we show 36 key-points marked on four individual East African chimpanzees, three adults, one infant.

lighting, angles, movement, distances, environments and individuals that the model has not encountered in training. To test performance on novel videos, 25 frames from 9 videos (total = 250 frames) were manually marked by an experienced coder (CW). This gave *x*-*y* coordinates for each body part for each individual for each marked frame. We then introduced these videos to the model to generate the model predicted coordinates and compared performance with the manually marked frames. Novel videos were further categorised as: easy, medium and hard, depending on the amount of visual noise present in the video (for details see Table S4).

3 | RESULTS

Model 1 took 28 h to train to 200,000 iterations, at which point the optimizer reported a loss of 0.001 on the ZBook. Training error was reported at 5.96 pixels, test error at 18.46 pixels (where a p cut off of 0.6 was applied: training error: 4.38 pixels, test error 10.12 pixels). A matching version of Model 1 was trained on the more computationally powerful Tesla V100 on nodes with an Intel(R) Xeon(R) Gold 6130 CPU @ 2.10GHz. Training took a similar time frame (26.5 h) to train to a loss of 0.002, which occurred at 100,000 iterations.

Model 2 took 26 h to train to 200,000 iterations and a loss of 0.001 on the ZBook. Training error was reported at 7.31 pixels, test error at 18.63 pixels (where a p cut-off of 0.6 was applied: training error: 4.6 pixels, test error 9.64 pixels).

3.1 | Human performance

The mean absolute Euclidean distance of the second human marker (as compared to the original human marker) was 26.09 (SD = 14.31) across points. As with model performance this varied with body part (Table 2).

TABLE 2 Mean absolute Euclidean distance between human coders per body part in $n = 570$ frames. Note that as not all body parts are visible in all frames the number of points per body part varied and is indicated by *N*.

Body part	Mean absolute Euclidean distance (SD)	*N*
Ankle	33.44 (27.65)	41
Ear	13.36 (9.43)	63
Elbow	43.06 (35.23)	85
Eye	7.05 (9.06)	40
Hip	48.83 (27.05)	48
Knee	23.95 (25.89)	88
Neck	27.12 (13.21)	43
Nose	5.31 (2.97)	43
Shoulder	29.81 (16.54)	66
Wrist	29.06 (22.44)	53
all	26.10 (14.31)	570

3.2 | Model performance on test frames

Despite additional variability in coder input (2 coders) and the addition of a new population (Kalinzu), Model 2 ($n = 2200$ training frames, Figure 3a) out-performed Model 1 (n = 1375 training frames) across body parts (Figure 3b). Examples of Model 1 and 2 tracking on test videos can be seen in Videos S1 and S2. Variation from the original human-marked training frames was substantially lower in both models, than that of the second human-marker, with mean absolute Euclidean distance values up to ~10 times smaller (e.g. Hip).

3.3 | Model performance on novel videos

Examples of Model 1 and 2 tracking on novel videos can be seen in Videos S5–S10 (Videos S5–S8) represent good performance, videos S9 and S10 represent poor performance. Model 2 achieved a consistently larger number of detections across all videos than Model 1 (Table 3), but only achieved a smaller mean absolute Euclidean distance in eight of the 17 videos, although these values were typically generated from a larger number of detections (Table 3).

Both models experienced some difficulty stitching detections (e.g. detecting an elbow) into assemblies (eg- detecting it as individual A's elbow), causing tracking to fall short on nine videos for model 1 (kalinzu10, kalinzu18, sonso17, sonso3, sonso4, waibira1, waibira17, wamba11 and no assembly for waibira15) and eight videos for model 2 (kalinzu10, kalinzu18, kalinzu20, sonso3, sonso4, waibira1, waibira17, wamba11).

In contrast, when considering specific body parts, Model 2 out-performed Model 1 more consistently, with the exception of the ear and shoulder (8 of 10 body parts, Table 4).

4 | DISCUSSION

Using DeepLabCut we successfully trained two models on an extremely challenging pose estimate problem: multi-animal wild forest-living chimpanzees and bonobos across behavioural contexts from hand-held video footage. We provide the first successful demonstration of multi-animal full body pose estimation and tracking in wild apes, our models are robust across the two closely related *Pan* species, across individuals of diverse ages and sexes, and across a wide range of socio-ecological environments—including from open clearings to dense forest, and from static behaviour such as grooming to highly dynamic behaviour such as play.

Tracking performance on test videos, videos from which some frames had been used in training, was substantially better than inter-human coder variation on similar video frames of wild chimpanzees. Direct comparison of performance on test and novel videos is challenging because the videos are themselves highly variable—thus, whether a body part is visible (therefore potentially detectable) or occluded also varies. Performance on entirely novel videos was lower, but tracking was still frequently successful, with accuracy in the easier body parts (ears, eyes, nose) reaching similar levels to that of inter-human marker variation on harder body parts (hips, shoulders).

Model 2 showed an improvement in detecting body points over Model 1 (approximately 10% more detections). Model 2 accuracy within this larger set of detections showed a consistent improvement across 8 of 10 body parts, with mean absolute Euclidean distance typically half to three quarters that of Model 1. In addition to a larger training set, Model 2 included training frames marked by a second human coder, and an additional chimpanzee community. DeepLabCut prioritises precision (how accurately it detected points) and as a result requires relatively high confidence to indicate a point, which can lead to lower recall (how many points it successfully detected) in models with homogeneous training sets. Building in diversity into training sets is an essential step in developing robust models and requires careful selection of video material and an understanding of the content of the video sets to which the model will be applied.

While a significant first step, there remain several limitations to the use of automated tracking of wild primate pose and behaviour. Perhaps the most significant is the time investment needed for model development—our models represent a first step, but still require further development before they are sufficiently robust to no longer require post-hoc human coder correction. However, larger training sets require further substantial time investment. Both current DeepWild models employed minimal training sets (<2000 frames), representing approximately 110–200 person-hours of investment to produce. Model training required an additional 26–28 h, with additional time invested in other work (of interest, there was no gain in training time—to a similar loss—in using greater computational power). Given an estimated investment of 200 h, and a human mark-up rate of approximately 2 h for 25 frames, use of Model 2

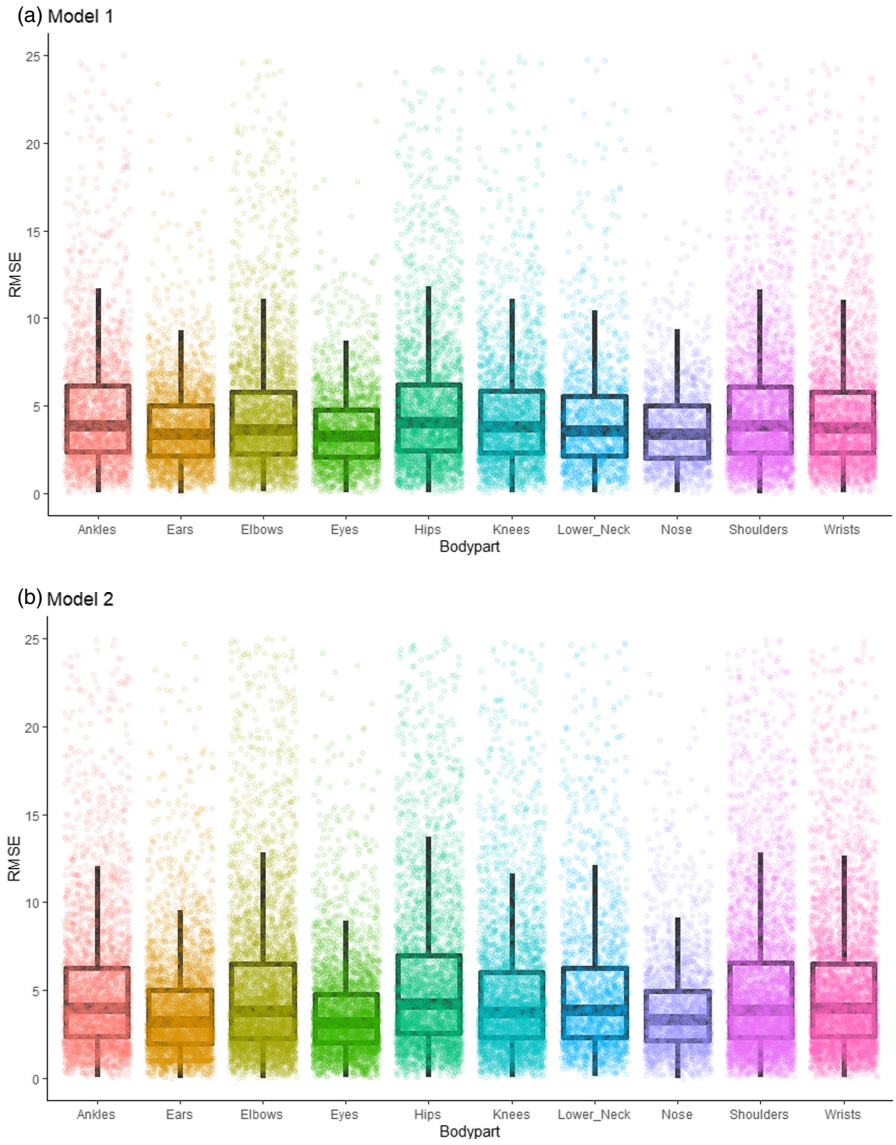

FIGURE 3 Model performance across body parts. Mean absolute Euclidean distance values for Model 1 (a), and Model 2 (b) across body parts (labelled as RMSE by DeepLabCut). Figure shows values between 0 and 25, full range was larger.

would pay-off in terms of time investment after just 40–50 min of video, a fraction of that coded in many studies of animal behaviour. Nevertheless, training set development costs may still represent a barrier to initial access. Given the strong model performance on the test frames, one approach for behavioural researchers who wish to start to employ tracking in very large video datasets is to manually mark a small subset of frames for each video they plan to track. Doing so offers a relatively low-cost easy point of entry to the use of machine learning tools for rapidly generating highly accurate pose estimation and behaviour tracking data. The enhanced flexibility needed to track fully novel videos requires investment in a more substantial training set. One approach here is to consider carefully what body parts may be of interest across projects, and then marking a full set, even where only a few are required for any one project. While doing so increases marking time in developing any individual training set, subsequent training sets can be stitched together with existing ones to produce consecutively more powerful models. At a certain point model performance is likely to be such that no further training frames are needed. An additional benefit of this multi-set building approach is that the training sets can be combined in different ways to tailor a particular model to a specific need—for example, particular individuals, species, or socio-ecological contexts. Multi-set building may be particularly effective if it can be adopted collaboratively across research groups—for example in research consortium such as ManyPrimates (Altschul et al., 2019), mitigating the cost for any one individual researcher or group, while rapidly producing large training sets and highly flexible models. A similar approach is taken with DeepLabCut's model zoo (Kane et al., 2020), where base models can be contributed and downloaded by users, who can then further refine them to their specific needs. Another means to offsetting development costs is to employ a community-science approach—here friendly online graphical user interfaces allow members of the public to contribute their time to research projects. Already used intensively with camera-trapping work (e.g. Chimp&See; Arandjelovic et al., 2016) for species identification and behaviour classification, platforms such as Zooniverse (www.zooniverse.org) provide scientists with an easy way to host online community science projects, including built in tools to assess the

TABLE 3 Model performance across novel videos. Model 1 contained 1375 frames from 55 videos across 5 *Pan* communities labelled by a single coder, and Model 2 contains 2200 frames from 110 videos across 6 *Pan* communities labelled by two coders. Videos were classified for difficulty on the basis of visual noise factors present. MAED = mean absolute Euclidean distance.

		Model 1		Model 2	
Video	Difficulty	MAED (SD)	n detections	MAED (SD)	*n* detections
Bossou7	Easy	109.3 (212.1)	470	27.4 (45.0)	541
Bossou8	Easy	44.6 (74.8)	175	31.9 (55.4)	279
Kalinzu19	Easy	80.3 (71.3)	148	78.8 (71.3)	214
Sonso17	Easy	19.1 (21.2)	59	15.6 (13.1)	65
Waibira7	Easy	26.3 (40.6)	55	49.9 (123.7)	83
Wamba16	Easy	27.9 (42.3)	136	69.9 (76.6)	93
Kalinzu18	Medium	94.2 (13.1)	3	161.7 (108.7)	4
Kalinzu20	Medium	56.5 (62.8)	62	48.7 (60.4)	63
Sonso3	Medium	19.4 (23.0)	33	172.3 (258.8)	17
Sonso6	Medium	23.2 (88.3)	227	85.5 (172.6)	189
Waibira17	Medium	16.5 (22.0)	152	43.4 (130.3)	154
Kalinzu10	Hard	62.6 (47.6)	4	31.7 (57.4)	7
Sonso4	Hard	11.1 (14.4)	24	26.6 (30.1)	9
Sonso9	Hard	31.3 (80.3)	109	88.5 (93.6)	64
Waibira1	Hard	134.7 (246.1)	15	131.4 (143.3)	15
Waibira18	Hard	74.4 (142.0)	91	101.7 (165.5)	159
Wamba11	Hard	170.8 (408.0)	38	82.7 (126.1)	24
All		60.4 (144.8)	1801	54.5 (104.8)	1980

	Model 1		Model 2	
Body part	MAED (SD)	*n* detections	MAED (SD)	*n* detections
Ankle	86.9 (131.6)	66	53.8 (99.7)	51
Ear	40.9 (117.1)	219	58.9 (118.5)	357
Elbow	109.5 (199.4)	143	53.0 (96.6)	247
Eye	32.8 (128.8)	389	26.9 (84.1)	247
Hip	85.4 (89.5)	38	67.0 (81.7)	105
Knee	87.7 (147.2)	103	69.8 (118.8)	77
Neck	78.2 (162.8)	154	53.6 (84.4)	170
Nose	48.7 (166.5)	222	37.4 (103.5)	140
Shoulder	47.4 (80.7)	345	70.0 (127.9)	362
Wrist	116.7 (207.2)	122	55.3 (81.6)	224
All	60.4 (144.7)	1801	54.5 (104.8)	1980

TABLE 4 Model performance across body parts within novel videos. Model 1 contained 1375 frames from 55 videos across 5 *Pan* communities labelled by a single coder, and Model 2 contains 2200 frames from 110 videos across 6 *Pan* communities labelled by two coders. Videos were classified for difficulty on the basis of visual noise factors present. MAED = mean absolute Euclidean distance.

reliability of the data contributed. To aid in these efforts we have made the base-models used in this paper available open-access and have collaborated with the DeepLabCut developers to allow open-access online marking of frames from our dataset, which will be regularly added into the base-model to improve performance (see Data Availability Statement for further details).

The use of pose-estimation tools for tracking movement in animal behaviour represents just a first step in analysis, generating large quantities of positional data that then need further analysis to detect particular patterns of movement, for example: a reach gesture, or dipping a water-tool. Several tracking and pose estimation tools now offer simple behavioural analyses options (see Tables S2 and S3). Once again, pre-existing options are typically available only for frequently used behaviour in model lab species (e.g. gait analysis in rodents; Adonias et al., 2019), but some tools now incorporating labelling of behaviour during key-point marking of training sets to allow for more bespoke behavioural analyses (e.g. Junior et al., 2012).

The automated coding of pose and movement offers faster, more accurate and robust ways to support current approaches to coding behaviour in wild primates. Moreover, the generation of 'big data'

on previously unattainable timescales together with the availability of collaborative large-scale primate behaviour video data-arks (e.g. Arandjelovic et al., 2016; or Hobaiter et al., 2021) allows us to ask new questions and model new processes, for example exploring variation in both space, across populations and species, and in time, across generations. For example, analyses of rhythmic movements, such as gait, lip-smacking, or drumming (cf. Eleuteri et al., 2022; Pereira, Kavanagh, et al., 2020; Schweinfurth et al., 2022); analyses of gestural signals would benefit from systematic descriptions of variation in movement patterns within and between action types, or features such as emphasis and arousal (cf. Graham et al., 2022; Grund et al., 2023); and analyses of variation in movement and the efficiency of motion paths could be applied to questions about the ontogenetic development and acquisition of tool-use. We describe the new tools available in this rapidly evolving landscape and suggest guidance for tool selection. With DeepWild we show that, without requiring specific expertise in machine learning, pose estimation and movement tracking of free-living wild primates in visually complex environments is now an attainable goal for behavioural researchers.

AUTHOR CONTRIBUTIONS

Catherine Hobaiter and Charlotte Wiltshire conceived the ideas and designed methodology; Catherine Hobaiter, Tetsuro Matsuzawa and Kirsty E Graham collected the original video data; Charlotte Wiltshire and Viola Komedová coded the data; Charlotte Wiltshire, James Lewis-Cheetham and Catherine Hobaiter analysed the data; Catherine Hobaiter and Charlotte Wiltshire led the writing of the manuscript. All authors contributed critically to the drafts and gave final approval for publication.

ACKNOWLEDGEMENTS

We thank the Guest Editors, Dr Thibaud Gruber and Dr Erica van de Waal for their invitation to contribute this paper, and we thank the journal editor and two anonymous reviewers for their constructive guidance on how to improve it. We thank the staff and communities at the field stations in which we collected our data in Uganda, Democratic Republic of Congo, and Guinea. In particular, we thank the staff of the Budongo Conservation Field Station for the years of support that has facilitated video data collection across projects from 2004 to 2022. We thank Professors Hashimoto and Furuichi for permission to collect video data at the Kalinzu and Wamba field sites. We thank Dr. Aly Garpard Soumah, the Institut de Recherche Environmentale de Bossou (IREB), for permission to collect data at the Bossou field site, which has run continuously through collaboration between the scholars of Kyoto University Primate Research Insitute, led by Yukimaru Sugiyama and Guinean scholars including: Jeremie Koman, Soh Pletah Bonimy, Bakary Coulibary, Tamba Tagbino, Makan Kourouma, Mamadou Diakite, Cécé Kolié, Iba Conde and Sekou Moussa Keita. We also thank the Guinean authorities who provided permission for the long-term researcher including: Ministre de l'Enseignement Supérieur et de la Recherche Scientifique, and Direction Générale de la Recherche Scientifique et de l'Innovation Technologique. We thank Alexander Mielke for his help with the R-code for the figures. All research projects within Uganda were conducted with permission from the Uganda Wildlife Authority, the Ugandan National Council for Science and Technology. All research projects were conducted under ethical permissions from the Animal Welfare and Ethics Committee of the University of St Andrews. We thank the developers of the programs we evaluated, and the machine learning communities for their work and their patience in answering many of our questions, as well as our lab group for their constructive discussions. This project received funding from the European Union's 8th Framework Programme, Horizon 2020 (grant agreement number: 802719) and the St Andrews Restarting Research Funding Scheme (2020).

CONFLICT OF INTEREST STATEMENT

The authors declare they have no conflicts of interest.

DATA AVAILABILITY STATEMENT

The DeepWild models as well as all data and code used in this paper are available for download in our Github repository https://github.com/Wild-Minds/DeepWild, which is archived in Zenodo at https://doi.org/10.5281/zenodo.7414432 (Wiltshire et al., 2022) Information on use of data from the Great Ape Video Database are available at https://doi.org/10.5281/zenodo.5600472 (Hobaiter et al., 2021). An online open-access interface for marking additional frames is available here: https://contrib.deeplabcut.org/label Frames received in through this process will be used to regularly update the base-model in our Github and shared with the DeepLabCut model zoo.

ORCID

Viola Komedová https://orcid.org/0000-0001-5554-7271
Tetsuro Matsuzawa https://orcid.org/0000-0002-8147-2725
Kirsty E. Graham https://orcid.org/0000-0002-7422-7676
Catherine Hobaiter https://orcid.org/0000-0002-3893-0524

REFERENCES

Achour, B., Belkadi, B., Filali, I., Laghrouche, M., & Lahdir, M. (2020). Image analysis for individual identification and feeding behaviour monitoring of dairy cows based on convolutional neural networks (CNN). *Biosystems Engineering*, *198*(1), 31–49. https://doi.org/10.1016/j.biosystemseng.2020.07.019

Adonias, A. F., Ferreira-Gomes, F., Allonso, R., Neto, F., & Cardoso, J. S. (2019). Towards automatic rat's gait analysis under suboptimal illumination conditions. *Pattern Recognition and Image Analysis*, 247–259. https://doi.org/10.1007/978-3-030-31321-0_22

Ahumada, J. A., Fegraus, E., Birch, T., Flores, N., Kays, R., O'Brien, T. G., Palmer, J., Schuttler, S., Zhao, J. Y., & Jetz, W. (2019). Wildlife insights: A platform to maximize the potential of camera trap and other passive sensor wildlife data for the planet. *Environmental Conservation*, *47*(1), 1–6. https://doi.org/10.1017/S0376892919000298

Altschul, D. M., Beran, M. J., Bohn, M., Call, J., De Troy, S., Duguid, S. J., Egelkamp, C. L., Fichtel, C., Fischer, J., Flessert, M., Hanus, D., Haun, D. B. M., Haux, L. M., Hernandez-Aguilar, R. A., Herrmann, E., Hopper, L. M., Joly, M., Kano, F., Keupp, S., … Watzek, J. (2019). Establishing an infrastructure for collaboration in primate cognition research. *PLoS ONE*, *14*, e0223675. https://doi.org/10.1371/journal.pone.0223675

Arandjelovic, M., Stevens, C. R., McCarthy, M. S., Dieguez, P., Kalan, A. K., Maldonado, N., Boesch, C., & Kuehl, H. S. (2016). Chimp&See: An online citizen science platform for large-scale, remote video camera trap annotation of chimpanzee behaviour, demography and individual identification. *PeerJ Preprints*. https://doi.org/10.7287/peerj.preprints.1792v1

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

Bala, P. C., Eisenreich, B. R., Yoo, S. B. M., Hayden, B. Y., Park, H. S., & Zimmerman, J. (2020). Automated markerless pose estimation in freely moving macaques with OpenMonkeyStudio. *Nature Communications*, *11*, 4560. https://doi.org/10.1038/s41467-020-18441-

Beery, S., Morris, D., & Yang, S. (2019). Efficient pipeline for camera trap image review. *arXiv*, 1907.06772. https://doi.org/10.48550/arXiv.1907.06772

Bergamini, L., Porrello, A., Capobianco Dondona, A., Del Negro, E., Mattioli, M., D'Alterio, A., & Calderara, S. (2018). Multi-views embedding for cattle re-identification. In *4th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS)*. https://doi.org/10.1109/SITIS.2018.00036

Berger-Wolf, T. Y., Rubenstein, D. I., Stewart, C. V., Holmberg, J. A., Parham, J., Menon, S., Crall, J., Van Oast, J., Kiciman, E., & Joppa, L. (2017). Wildbook: Crowdsourcing, computer vision, and data science for conservation. *arXiv*. https://doi.org/10.48550/arXiv.1710.08880

Bianco, M. J., Gerstoft, P., Traer, J., Ozanich, E., Roch, M. A., Gannot, S., & Deledalle, C.-A. (2019). Machine learning in acoustics: Theory and applications. *The Journal of the Acoustical Society of America*, *146*(1), 3590–3628. https://doi.org/10.1121/1.5133944

Burton, A. C., Neilson, E., Moreira, D., Ladle, A., Steeweg, R., Fisher, J. T., Bayne, E., & Boutin, S. (2015). Wildlife camera trapping: A review and recommendations for linking surveys to ecological processes. *The Journal of Applied Ecology*, *52*(3), 675–685.

Clemensson, E., Abbaszadeh, M., Fanni, S., Espa, E., & Cenci, M. A. (2020). Tracking rats in operant conditioning chambers using a versatile homemade video camera and DeepLabCut. *Journal of Visualized Experiments*, *160*, e61409. https://doi.org/10.3791/61409

DeepLabCut Github. (2021a). https://github.com/DeepLabCut/DeepLabCut/blob/master/docs/UseOverviewGuide.md

DeepLabCut Github. (2021b). https://deeplabcut.github.io/DeepLabCut/docs/intro.html

DeepLabCut Github. (2021c). https://github.com/DeepLabCut/DeepLabCut/blob/master/docs/installation.md

Doran, D. M. (1993). Comparative locomotor behavior of chimpanzees and bonobos: The influence of morphology on locomotion. *American Journal of Physical Anthropology*, *91*(1), 83–98.

Eggeling, W. J. (1947). Observations on the ecology of the Budongo rain forest, Uganda. *The Journal of Ecology*, *34*(1), 20–87.

Eleuteri, V., Henderson, M., Soldati, A., Badihi, B., Zuberbühler, K., & Hobaiter, C. (2022). The form and function of chimpanzee buttress drumming. *Animal Behaviour*, *192*, 189–205. https://doi.org/10.1016/j.anbehav.2022.07.013

Estrada, A., Garber, P. A., Rylands, A. B., Roos, C., Fernandez-Duque, E., Di Fiore, A., Nekaris, K. A. I., Nijman, V., Heymann, E. W., Lambert, J. E., Rovero, F., Barelli, C., Setchell, J. M., Gilllespie, T. R., Mittermeier, R. A., Arregoitia, L. V., de Guinea, M., Gouveia, S., Dobrovolski, R., … Li, B. (2017). Impending extinction crisis of the world's primates: Why primates matter. *Science Advances*, *3*(1), e1600946. https://doi.org/10.1126/sciadv.1600946

Gadea, G. H. (2021). https://guillermohidalgogadea.com/openlabnotebook/

Gal, A., Saragosti, J., & Kronauer, D. J. C. (2020). anTraX, a software package for high-throughput video tracking of color-tagged insects. *eLife*, e58145. https://doi.org/10.7554/eLife.58145

Geuther, B. Q., Deats, S. P., Fox, K. J., Murray, S. A., Braun, R. E., White, J. K., Chesler, E. J., Lutz, C. M., & Kumar, V. (2019). Robust mouse tracking in complex environments using neural networks. *Communications Biology*, *2*, 124. https://doi.org/10.1038/s42003-019-0362-1

Gosztolai, A., Günel, S., Lobato-Ríos, V., Abrate, M. P., Morales, D., Rhodin, H., Fua, P., & Ramdya, P. (2021). LiftPose3D, a deep learning-based approach for transforming two-dimensional to three-dimensional poses in laboratory animals. *Nature Methods*, *18*, 975–981. https://doi.org/10.1038/s41592-021-01226-z

Graham, K. E., Badihi, G., Safryghin, A., Grund, C., & Hobaiter, C. (2022). A socio-ecological perspective on the gestural communication of great ape species, individuals, and social units. *Ethology, Ecology, & Evolution*, *34*(2), 235–259.

Grund, C., Badihi, G., Graham, K. E., Safryghin, A., & Hobaiter, C. (2023). GesturalOrigins: A bottom-up framework for establishing systematic gesture data across ape species. *Behaviour Research Methods*. https://doi.org/10.3758/s13428-023-02082-9

Günel, S., Rhodin, H., Morales, D., Campagnolo, J., Ramdya, P., & Fua, R. (2019). DeepFly3D, a deep learning-based approach for 3D limb and appendage tracking in tethered, adult *Drosophila*. *eLife*, *8*, e48571. https://doi.org/10.7554/eLife.48571

Guo, S., Xu, P., Miao, Q., Shao, G., Chapman, C. A., Chen, X., He, G., Fang, D., Zhang, H., Sun, Y., Shi, Z., & Li, B. (2020). Automatic identification of individual primates with deep learning techniques. *iScience*, *23*, e101412. https://doi.org/10.1016/j.isci.2020.101412

Habe, H., Takeuchi, Y., Terayama, K., & Sakagami, M. (2021). Pose estimation of swimming fish using NACA airfoil model for collective behavior analysis. *Journal of Robotics and Mechatronics*, *33*(3), 547–555. https://doi.org/10.20965/jrm.2021.p0547

Hashimoto, C., Tashiro, Y., Kimura, D., Enomoto, T., Ingmanson, E. J., Idani, G., & Furuichi, T. (1998). Habitat use and ranging of wild bonobos (*Pan paniscus*) at Wamba. *International Journal of Primatology*, *19*(1), 1045–1060.

Hastie, T., Tibshirani, R., & Friedman, J. H. (2001). *The elements of statistical learning: Data mining, inference, and prediction*. Springer.

Haucke, T., Kühl, H. S., Hoyer, J., & Steinhage, V. (2021). Overcoming the distance estimation bottleneck in estimating animal abundance with camera traps. *arXiv*. https://doi.org/10.48550/arXiv.2105.04244

Hayden, B. Y., Park, H. S., & Zimmermann, J. (2021). Automated pose estimation in primates. *American Journal of Primatology*, *84*(10), e23348. https://doi.org/10.1002/ajp.23348

Hobaiter, C., Gal Badihi, G., de Melo Daly, G. B., Eleuteri, V., Graham, K. E., Grund, C., Henderson, M., Rodrigues, E. D., Safryghin, A., Soldati, A., & Wiltshire, C. (2021). The great ape dictionary video database. *Zenodo*. https://doi.org/10.5281/zenodo.5600471

Hsu, A. I., & Yttri, E. A. (2021). B-SOiD, an open-source unsupervised algorithm for identification and fast prediction of behaviors. *Nature*, *12*, 5188. https://doi.org/10.1038/s41467-021-25420-x

Islam, M. D., Mo, J., & Sattar, J. (2021). Robot-to-robot relative pose estimation using humans as markers. *Autonomous Robots*, *45*(1), 579–593. https://doi.org/10.1007/s10514-021-09985-6

Joska, D., Clark, L., Muramatsu, N., Jericevich, R., Nicholls, F., Mathis, A., Mathis, M. W., & Patel, A. (2021). AcinoSet: A 3D pose estimation dataset and baseline models for cheetahs in the wild. In *IEEE International Conference on Robotics and Automation (ICRA)* (pp. 13901–13908). https://doi.org/10.1109/ICRA48506.2021.9561338

Jungers, W. L., & Susman, R. L. (1984). Body size and skeletal allometry in African apes. In R. L. Sussman (Ed.), *The pygmy chimpanzees* (pp. 137–177). Springer.

Junior, C. F. C., Pederiva, C. N., Bose, R. C., Garcia, V. A., Lino-de-Oliveira, C., & Marino-Neto, J. (2012). ETHOWATCHER: Validation of a tool for behavioral and video-tracking analysis in laboratory animals. *Computers in Biology and Medicine*, *42*(2), 257–264.

Kane, G. A., Lopes, G., Saunders, J. L., Mathis, A., & Mathis, W. M. (2020). Real-time, low-latency closed-loop feedback using

markerless posture tracking. *eLife*, *9*, e61909. https://doi.org/10.7554/eLife.61909

Karashchuk, P., Rupp, K. L., Dickinson, E. S., Walling-Bell, S., Sanders, E., Bingni, E. A., Brunton, W., & Tuthill, J. C. (2021). Anipose: A toolkit for robust markerless 3D pose estimation. *Cell Reports*, *36*, 13. https://doi.org/10.1016/j.celrep.2021.109730

Kellenberger, B., Tuia, D., & Morris, D. (2020). AIDE: Accelerating image-based ecological surveys with interactive machine learning. *Methods in Ecology and Evolution*, *11*(12), 1716–1727. https://doi.org/10.1111/2041-210X.13489

Khan, N. U., & Wan, W. (2018). A review of human pose estimation from single image. In *2018 International Conference on Audio, Language and Image Processing (ICALIP)*. https://doi.org/10.1109/ICALIP.2018.8455796

Kiel, M., Berh, D., Daniel, J., Otto, N., Steege, A. T., Jiang, X., Liebau, E., & Risse, B. (2018). A multi-purpose worm tracker based on FIM. *bioRxiv*. https://doi.org/10.1101/352948

Körschens, M., & Denzler, J. (2019). ELPephants: A fine-grained dataset for elephant Re-identification. In *2019 IEEE/CVF International Conference on Computer Vision Workshop, (ICCVW)*. https://doi.org/10.1109/ICCVW.2019.00035

Labuguen, R., Matsumoto, J., Negrete, S. B., Nishimaru, H., Nishijo, H., Takada, M., Go, Y., Inoue, K., & Shibata, T. (2021). MacaquePose: A novel "in the wild" macaque monkey pose dataset for markerless motion capture. *Frontiers in Behavioural Neuroscience*, *14*, 581154. https://doi.org/10.3389/fnbeh.2020.581154

Lauer, J., Zhou, M., Ye, S., Menegas, W., Nath, T., Rahman, M. M., Santo, V. D., Soberanes, D., Feng, G., Murthy, V. N., & Lauder, G. (2021). Multi-animal pose estimation and tracking with DeepLabCut. *bioRxiv*. https://doi.org/10.1101/2021.04.30.442096

Li, S., Li, J., Tang, H., Qian, R., & Lin, W. (2020). ATRW: A benchmark for Amur tiger re-identification in the wild. *arXiv*. https://doi.org/10.1145/3394171.3413569

Marks, M., Qiuhan, J., Sturman, O., von Ziegler, L., Kollmorgen, S., von der Behrens, W., Mante, V., Bohacek, J., & Yanik, M. F. (2022). Deep-learning based identification, tracking, pose estimation, and behavior classification of interacting primates and mice in complex environments. *bioRxiv*. https://doi.org/10.1101/2020.10.26.355115

Mathis, A., Mamidanna, P., Curry, 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*(1), 1281–1289.

Matsuzawa, T., Humle, T., & Sugiyama, Y. (2011). *The chimpanzees of Bossou and Nimba*. Springer Nature.

Mei, J., Hwang, J.-N., Romain, S., Rose, C., Moore, B., & Magrane, K. (2021). Absolute 3d pose estimation and length measurement of severely deformed fish from monocular videos in longline fishing. In *ICASSP 2021–2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. https://doi.org/10.1109/ICASSP39728.2021.9414803

Munch, K. L., Wapstra, E., Thomas, S., Fisher, M., & Sinn, D. L. (2019). What are we measuring? Novices agree amongst themselves (but not always with experts) in their assessment of dog behaviour. *Ethology*, *125*(4), 203–211.

Narouzzadeh, M. S., Nguyen, A., Kosala, M., Swanson, A., Palmer, M. S., Packer, C., & Clune, J. (2018). Automatically identifying, counting, and describing wild animals in camera-trap images with deep learning. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(25), E5716–E5725. https://doi.org/10.1073/pnas.1719367115

Nath, T., Mathis, A., Chen, A. C., Patel, A., Bethge, M., & Mathis, M. W. (2019). Using DeepLabCut for 3D markerless pose estimation across species and behaviors. *Nature Protocols*, *14*, 2152–2176. https://doi.org/10.1038/s41596-019-0176-

Panadeiro, V., Rodriguez, A., Henry, J., Wlodkowic, D., & Andersson, M. (2021). A review of 28 free animal-tracking software applications: Current features and limitations. *Nature*, *50*(1), 246–254. https://doi.org/10.1038/s41684-021-00811-1

Pathak, S. D., Ng, L., Wyman, B., Fogarasi, S., Racki, S., Oelund, J. C., Spark, B., & Chalana, V. (2003). Quantitative image analysis: Software systems in drug development trails. *Drug Discovery Today*, *8*(10), 451–458.

Pereira, A. S., Kavanagh, E., Hobaiter, C., Slocombe, K. E., & Lameira, A. R. (2020). Chimpanzee lip-smacks confirm primate continuity for speech-rhythm evolution. *Biology Letters*, *16*, 20200232. https://doi.org/10.1098/rsbl.2020.0232

Pereira, T. D., Aldarondo, D. E., Willmore, L., Kislin, M., Wang, S. S.-H., Murthy, M., & Shaevitz, J. W. (2019). Fast animal post estimation using deep neural networks. *Nature Methods*, *16*, 117–125. https://doi.org/10.1038/s41592-018-0234-5

Pereira, T. D., Tabris, N., Li, J., Ravindranath, S., Papdoyannis, E. S., Wang, Z. Y., Turner, D. M., McKenzie-Smith, G., Kocker, S. D., Falkner, A. L., Shaevitz, J. W., & Murthy, M. (2020). SLEAP: Multi-animal pose tracking. *bioRxiv*. https://doi.org/10.1101/2020.08.31.276246

Prakash, C., Kumar, R., Mittal, N., & Raj, G. (2018). Vision based identification of joint coordinates for marker-less gait analysis. *Procedia Computer Science*, *132*(1), 68–75.

Premarathna, K. S. P., Rathnayaka, R. M. K. T., & Charles, J. (2020). An elephant detection system to prevent human-elephant conflict and tracking of elephant using deep learning. In *5th International Conference on Information Technology Research (ICITR)*. https://doi.org/10.1109/ICITR51448.2020.9310798

Rohan, A., Rabah, M., Hosny, T., & Kim, S.-H. (2020). Human pose estimation-based real-time gait analysis using convolutional neural network. *IEEE Access*, *8*(1), 191542–191550. https://doi.org/10.1109/ACCESS.2020.3030086

Sarafianos, N., Boteanu, B., Ionescu, B., & Kakadiaris, I. A. (2016). 3D human pose estimation: A review of the literature and analysis of covariates. *Computer Vision and Image Understanding*, *152*(1), 1–20. https://doi.org/10.1016/j.cviu.2016.09.002

Schneider, S., Taylor, G. W., Linquist, S., & Kremer, S. C. (2018). Past, present and future approaches using computer vision for animal re-identification from camera trap data. *Methods in Ecology and Evolution*, *10*(4), 461–470. https://doi.org/10.1111/2041-210X.13133

Schofield, D., Nagrani, A., Zisserman, A., Hayashi, M., Matsuzawa, T., Biro, D., & Carvalho, S. (2019). Chimpanzee face recognition from videos in the wild using deep learning. *Science Advances*, *5*(9), eaaw0736. https://doi.org/10.1126/sciadv.aaw0736

Schweinfurth, M. K., Baldridge, D. B., Finnerty, K., Call, J., & Knoblich, G. K. (2022). Inter-individual coordination in walking chimpanzees. *Current Biology*, *32*(23), 5138–5143.

Sheppard, K., Gardin, J., Sabnis, G. S., Peer, A., Darell, M., Deats, S., Geuther, B., Lutz, C. M., & Kumar, V. (2022). Stride-level analysis of mouse open field behavior using deep-learning-based pose estimation. *Cell Reports*, *38*(2), 110231. https://doi.org/10.1016/j.celrep.2021.110231

Steenweg, R., Hebblewhite, M., Kays, R., Ahumada, J., Fisher, J. T., Burton, C., Townsend, S. E., Carbone, C., Rowcliffe, J. M., Whittington, J., Brodie, J., Royle, J. A., Switalski, A., Clevenger, A. P., Heim, N., & Rich, L. N. (2016). Scaling-up camera traps: Monitoring the planet's biodiversity with networks of remote sensors. *Frontiers in Ecology and Environment*, *15*(1), 26–34. https://doi.org/10.1002/fee.1448

Swann, D. E., Kawanishi, K., & Palmer, J. (2011). Evaluating types and features of camera traps in ecological studies: A guide for researchers. In A. F. O'Connell, J. D. Nichols, & K. U. Karanth (Eds.), *Camera traps in animal ecology*. Springer. https://doi.org/10.1007/978-4-431-99495-4_3

Terada, S., Nackoney, J., Sakamaki, T., Mulavwa, M. N., Yumoto, T., & Furuichi, T. (2015). Habitat use of bonobos (*Pan paniscus*) at Wamba: Selection of vegetation types for ranging, feeding, and night-sleeping. *American Journal of Primatology*, *77*(6), 701–713. https://doi.org/10.1002/ajp.22392

Tsuruo, A., Ringhofer, M., Yamamoto, S., & Ikeda, K. (2020). Mathematical model of horse and rider interaction during horse jumping. *2020 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 939–943.

Tuia, D., Kellenberger, B., Beery, S., Costelloe, B. R., Zuffi, S., Risse, B., Mathis, A., Mathis, M. W., van Langevelde, F., Burghardt, T., Kays, R., Klinck, H., Wikelski, M., Couzin, I. D., van Horn, G., Crofoot, M. C., Stewar, C. V., & Berger-Wolf, T. (2022). Perspectives in machine learning for wildlife conservation. *Nature Communications*, *13*, 792. https://doi.org/10.1038/s41467-022-27980-y

Wäldchen, J., & Mäder, P. (2018). Machine learning for image based species identification. *Methods in Ecology and Evolution*, *9*(11), 2216–2225. https://doi.org/10.1111/2041-210X.13075

Walter, T., & Couzin, I. D. (2021). TRex, a fast multi-animal tracking system with markerless identification, and 2D estimation of posture and visual fields. *eLife*, *10*, e64000. https://doi.org/10.7554/eLife.640000

Wang, T.-H., & Lien, J.-J. (2009). Facial expression recognition system based on rigid and non-rigid motion separation and 3D pose estimation. *Pattern Recognition*, *42*(5), 962–977. https://doi.org/10.1016/j.patcog.2008.09.035

Weinstein, B. G. (2017). A computer vision for animal ecology. *Journal of Animal Ecology*, *87*(3), 533–545. https://doi.org/10.1111/1365-2656.12780

Whytock, R. C., Świeżewski, J., Zwerts, J. A., Bara-Słupski, T., Pambo, A. F. K., Rogala, M., Bahaa-el-din, L., Boekee, K., Brittain, S., Cardoso, A. W., Henschel, P., Lehmann, D., Momboua, B., Opepa, C. K., Orbell, C., Pitman, R. T., Robinson, H. S., & Abernethy, K. A. (2021). Robust ecological analysis of camera trap data labelled by a machine learning model. *Methods in Ecology and Evolution*, *12*(6), 1080–1092. https://doi.org/10.1111/2041-210X.13576

Willi, M., Pitman, R. T., Cardoso, A. W., Locke, C., Swanson, A., Boyer, A., Veldthuis, M., & Fortson, L. (2018). Identifying animal species in camera trap images using deep learning and citizen science. *Methods in Ecology and Evolution*, *10*(1), 80–91. https://doi.org/10.1111/2041-210X.13099

Wiltshire, C., Lewis-Cheetham, J., Komedová, V., Matsuzawa, T., Graham, K. E., & Hobaiter, C. (2022). WildMinds/DeepWild: Deep Wild. *Zenodo*. https://doi.org/10.5281/zenodo.7414432

Xu, Z., & Cheng, X. E. (2017). Zebrafish tracking using convolutional neural networks. *Scientific Reports*, *7*(1), 42815. https://doi.org/10.1038/srep42815

Yu, X., Wang, J., Kays, R., Jansen, P. A., Wang, T., & Huang, T. (2013). Automated identification of animal species in camera trap images. *EURASIP Journal on Image and Video Processing*, *52*(1). https://doi.org/10.1186/1687-5281-2013-52

Yu, X., Zhou, H., Wu, L., & Liu, Q. (2011). High-performance drosophila movement tracking. In *2011 Third Chinese Conference on Intelligent Visual Surveillance*. https://doi.org/10.1109/IVSurv.2011.6157027

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

Table S1. Examples of machine learning tools for automated species and individual identification. Twelve commonly used tools, we describe which species they have been used for, whether additional training by the user is required, and whether they can re-identify individuals.**Table S2.** A comprehensive list of machine learning tools currently available for automating animal tracking. We describe whether they are able to track multiple animals; whether they have been used on 'Wild' data, which refers to their use in free-living animals in non-controlled environments including large-scale natural sanctuaries as well as fully wild individuals, but not those living in zoos or smaller open enclosures. We list the species for which the tool has currently been used, the Tracking Style, and additional comments on functionality or requirements. This list is designed to be used in conjunction with Figure 1, which provides decision-making guidance on which tool(s) may be suitable for a particular use.**Table S3.** Detailed description of the usability and functionality of the tools currently available for automated animal pose estimation across species. We consider the specific function of the tool, whether it provides a graphical user interface (GUI); whether users are required to have a understanding of coding languages such as Python; whether multi-animal (MA) tracking is available; the format of data output; whether the user specified the features to be tracked; whether it requires access to a specific graphics card; whether, in the absence of this, it is compatible with Google Colaboratory (colab); whether it has prior use with 'Wild' or free-roaming individuals in visually complex environments; what documentation and tutorials are currently available; and any other particular pros and cons associated with current functionality. Note that we do not include three additional tools that are only currently suitable for pose-estimation with specific species: Open Monkey Studio (macaques), AlphaTracker (mice) and DeepFly3d (flies).**Table S4.** Classification of novel videos according to visual noise.**Video S1.** Model 1 tracking of test video Sonso10. In this video, four East African chimpanzees sit in the undergrowth, the two immature individuals wrestle and play. There are movements of the camera and undergrowth, and the individuals move back and forth across one another. Some frames from this video were included in the training set, and tracking is very good throughout. Video is available here: https://tinyurl.com/DeepWildvideos.

Video S2. Model 2 tracking of test video Sonso10. In this video four East African chimpanzees sit in the undergrowth, the two immature individuals wrestle and play. There are movements of the camera and undergrowth, and the individuals move back and forth across one another. Some frames from this video were included in the training set, tracking is excellent throughout with increased stability of points over Model 1. Video is available here: https://tinyurl.com/DeepWildvideos.

Video S3. Model 1 tracking of test video Wamba10. In this video three bonobos sit in dense undergrowth. The video was classified as 'hard' and some frames from this video were included in the training set. This is at the poor end of tracking performance, the main key points are well tracked throughout, but some body parts are occasionally missed or lost, and the model mistakes some parts of the environment for bonobos, adding out of place key-points that would require manual cleaning. Video is available here: https://tinyurl.com/DeepWildvideos.

Video S4. Model 2 tracking of test video Wamba10. As in video S3, three bonobos sit in dense undergrowth. The video was classified as 'hard' and some frames from this video were included in the training set. While there are still some problems with tracking, for example out of place key-points, tracking performance is clearly improved over that of Model 1. Video is available here: https://tinyurl.com/DeepWildvideos.

Video S5. Model 1 tracking of novel video Waibira17. Two East African chimpanzees walk near to the camera, the video is short and was entirely novel to the model, tracking is very good throughout. Video is available here: https://tinyurl.com/DeepWildvideos

Video S6. Model 2 tracking of novel video Waibira17. As in video S5, two East African chimpanzees walk near to the camera, the video is short and was entirely novel to the model, tracking is excellent throughout with an improvement of tracking of the second partially obscured individual over Model 1. Video is available here: https://tinyurl.com/DeepWildvideos.

Video S7. Model 1 tracking of novel video Sonso6. Three East African chimpanzees are in an open area near to the road. One individual runs quickly across, the camera pans and the two individuals are seen next two each other against the forest. The video was entirely novel to the model and tracking is very good throughout. Video is available here: https://tinyurl.com/DeepWildvideos.

Video S8. Model 2 tracking of novel video Sonso6. As in video S7, three East African chimpanzees are in an open area near to the road. One individual runs quickly across, the camera pans and the two individuals are seen next two each other against the forest. The video was entirely novel to the model and tracking is excellent throughout, with increased stability of detections over Model 1. Video is available here: https://tinyurl.com/DeepWildvideos.

Video S9. Model 1 tracking of novel video Sonso4. As in video S7, Three East African chimpanzees are on a branch in the canopy. The individuals are back-lit, there are obstructions to the view, overlapping individuals, camera angle changes and zoom. The video was classified as 'hard' and was entirely novel to the model. Tracking performance varied across the individuals, but was poor for the two on the right, there are very few out of place key-points, but some body parts are consistently not recognised or mis-identified. Video is available here: https://tinyurl.com/DeepWildvideos.

Video S10. Model 2 tracking of novel video Sonso4. As in video S7, three East African chimpanzees are on a branch in the canopy. The individuals are back-lit, there are obstructions to the view, overlapping individuals, camera angle changes and zoom. The video was classified as 'hard' and was entirely novel to the model. Tracking performance varied across the individuals, but was still poor for the two on the right; however, there is a clear improvement over Model 1 with fewer mis-identifications and more stable key-point tracking of the more difficult individuals. Video is available here: https://tinyurl.com/DeepWildvideos

Video S11. Model 1 tracking of test video Sonso5. Two East African chimpanzees walking through a clearing in the forest. The area is well lit with little obstruction. Some frames from this video were included in the training set, tracking is satisfactory throughout. Video is available here: https://tinyurl.com/DeepWildvideos

Video S12. Model 2 tracking of test video Sonso5. Two East African chimpanzees walking through a clearing in the forest. The area is well lit with little obstruction. Some frames from this video were included in the training set, tracking is satisfactory throughout. Video is available here: https://tinyurl.com/DeepWildvideos.

How to cite this article: Wiltshire, C., Lewis-Cheetham, J., Komedová, V., Matsuzawa, T., Graham, K. E., & Hobaiter, C. (2023). DeepWild: Application of the pose estimation tool DeepLabCut for behaviour tracking in wild chimpanzees and bonobos. *Journal of Animal Ecology*, *00*, 1–15. https://doi.org/10.1111/1365-2656.13932