



Deep learning tools for the measurement of animal behavior in neuroscience

Mackenzie Weygandt Mathis and Alexander Mathis

Recent advances in computer vision have made accurate, fast and robust measurement of animal behavior a reality. In the past years powerful tools specifically designed to aid the measurement of behavior have come to fruition. Here we discuss how capturing the postures of animals — pose estimation — has been rapidly advancing with new deep learning methods. While challenges still remain, we envision that the fast-paced development of new deep learning tools will rapidly change the landscape of realizable real-world neuroscience.

Address

Harvard University, Cambridge, MA, USA

Corresponding author: Mathis, Mackenzie Weygandt
 (mackenzie@post.harvard.edu)

Current Opinion in Neurobiology 2019, 60:1–11

This review comes from a themed issue on **Neurobiology of behavior**

Edited by **Michael Brecht** and **Richard Mooney**

<https://doi.org/10.1016/j.conb.2019.10.008>

0959-4388/Published by Elsevier Ltd.

Introduction

Behavior is the most important output of the underlying neural computations in the brain. Behavior is complex, often multi-faceted, and highly context dependent both in how con-specifics or other observers understand it, as well as how it is emitted. The study of animal behavior — ethology — has a rich history rooted in the understanding that behavior gives an observer a unique look into an animal's *umwelt* [1–3]; what are the motivations, instincts, and needs of an animal? What survival value do they provide? In order to understand the brain, we need to measure behavior in all its beauty and depth, and distill it down into meaningful metrics. Observing and efficiently describing behavior is a core tenant of modern ethology, neuroscience, medicine, and technology.

In 1973 Tinbergen, Lorenz, and von Frisch were the first ethologists awarded the Nobel Prize in Physiology or Medicine for their pioneering work on the patterns of individual and social group behavior [4]. The award heralded a coming-of-age for behavior, and how

rigorously documenting behavior can influence how we study the brain [4]. Manual methods are powerful, but also highly labor intensive and subject to the limits of our senses. Matching (and extending) the capabilities of biologists with technology is a highly non-trivial problem [5,6], yet harbors tremendous potential. How does one compress an animal's behavior over long time periods into meaningful metrics? How does one use behavioral quantification to build a better understanding of the brain and an animal's *umwelt* [1]?

In this review we discuss the advances, and challenges, in animal pose estimation and its impact on neuroscience. Pose estimation refers to methods for measuring posture, while posture denotes the geometrical configuration of body parts. While there are many ways to record behavior [7–9], videography is a non-invasive way to observe the posture of animals. Estimated poses across time can then, depending on the application, be transformed into kinematics, dynamics, and actions [3,5–7,10]. Due to the low-dimensional nature of posture, these applications are computationally tractable.

A very brief history of pose estimation

The postures and actions of animals have been documented as far back as cave paintings, illustrating the human desire to distill the essence of an animal for conveying information. As soon as it was possible to store data on a computer, researchers have built systems for automated analysis. Over time, these systems reflected all flavors of artificial intelligence from rule-based via expert systems, to machine learning [11,12*]. Traditionally posture was measured by placing markers on the subject [3], or markerlessly by using body models (i.e. cylinder-based models with edge features [13]). Other computer vision techniques, such as using texture or color to segment the person from the background to create silhouettes [14,15], or using so-called hand-crafted features with decoders [11,12*,16**] were also popular before deep learning flourished.

The deep learning revolution for posture

Pose estimation is a challenging problem, but it has been tremendously advanced in the last five years due to advances in deep learning. Deep neural networks (DNNs) are computational algorithms that consist of simple units, which are organized in layers and then serially stacked to form 'deep networks'. The connections between the units are trained on data and therefore learn to extract information from raw data in order to solve

tasks. The current deep learning revolution started with achieving human-level accuracy for object recognition on the ImageNet challenge, a popular benchmark with many categories and millions of images [16^{••},17]. A combination of large annotated data sets, sophisticated network architectures, and advances in hardware made this possible and quickly impacted many problems in computer vision (see reviews [12[•],16^{••},18]).

2D and 3D (human) pose estimation

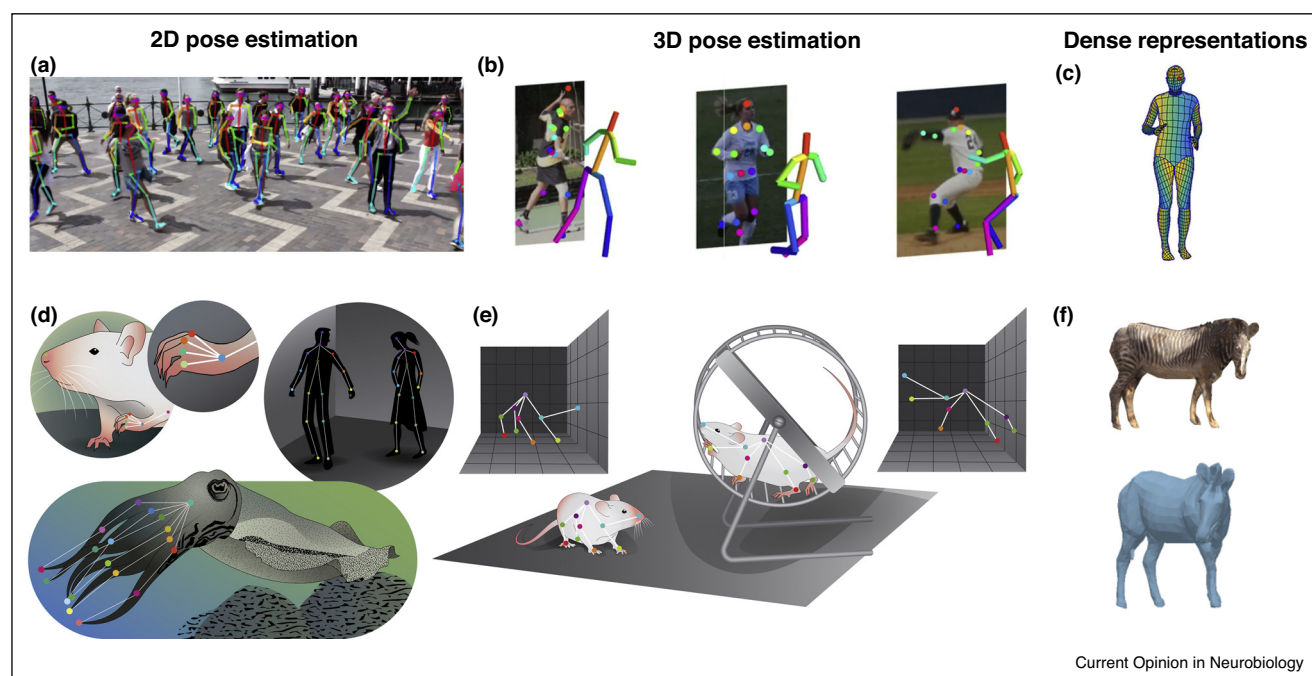
In 2014 ‘DeepPose’ was the first paper to apply deep learning to human 2D pose estimation [23], and immediately new networks were proposed that improved accuracy by introducing a translation invariant model [24], and convolutional networks plus geometric constraints [25,26]. In the few years since, numerous human pose estimation papers (approx. 4000 on Google Scholar), and new benchmarks with standardized datasets plus evaluation metrics appeared, which allow better comparisons of ‘state-of-the-art’ performance [27]. This culture has driven rapid and remarkable increases in performance: from 44% of body parts correctly labeled to nearly 94% — with the top 15 networks being within a few percentage points of each other (an example top network is shown in Figure 1a) [19^{••},28,29^{••},30–32]. The history and many advances in 2D human pose estimation are comprehensively reviewed in [11,33].

3D human pose estimation is a more challenging task and 3D labeled data is more difficult to acquire. There have been massive improvements in networks; see review [34]. Yet currently, the highest accuracy is achieved by using multiple 2D views to reconstruct a 3D estimate (Figure 1b; [35,20]), but other ways of ‘lifting’ 2D into 3D are being actively explored [36,37,20].

Dense-representations of bodies

Other video-based approaches for capturing the posture and soft tissue of humans (and other animals) also exist. Depth-cameras such as the Microsoft Kinect have been used in humans [38,39] and rodents [40,41]. Recently dense-pose representations, that is, 3D point clouds or meshes (Figure 1c), have become a popular and elegant way to capture the soft-tissue and shape of bodies, which are highly important features for person identification, fashion (i.e. clothing sales), and in medicine [21[•],42,43]. However, state-of-the-art performance currently requires body-scanning of many subjects to make body models. Typically, large datasets are collected to enable the creation of robust algorithms for inference on diverse humans (or for animals, scanning toy models has been fruitful [44]). Recently, outstanding improvements have been made to capture shapes of animals from images [22^{••},45[•],46]. However, there are no animal-specific tool-boxes geared towards neuroscience applications, although

Figure 1



2D pose estimation, 3D pose estimation & dense representations of humans and other animals: (a) Example 2D multi-human pose estimation from OpenPose [19^{••}]. (b) Example 3D human pose estimation from [20]. (c) Dense representations of humans with DensePose, adapted from Guler *et al.* [21[•]]. (d) Animals have diverse bodies and experimenter's are often interested in specific key points, making tailored network attractive solutions. DNNs open a realm of possibilities: from mice to cuttlefish. (e) 3D pose estimation requires multiple cameras views, or 2D to 3D ‘lifting’. (f) The new SMALST model which fits full 3D models to images from Zuffi *et al.* [22^{••}] applied to zebras.

we believe that this will change in the near future, as for many applications having the soft-tissue measured will be highly important, that is, in obesity or pregnancy research, etc.

Animal pose estimation

The remarkable performance when using deep learning for human 2D & 3D pose estimation plus dense-representations made this large body of work ripe for exploring its utility in neuroscience (Figure 1d–f). In the past two years, deep learning tools for laboratory experiments have arrived (Figure 2a–d).

Many of the properties of DNNs were extremely appealing: remarkable and robust performance, relatively fast inference due to GPU hardware, and efficient code due to modern packages like TensorFlow and PyTorch (reviewed in [47]). Furthermore, unlike for many previous algorithms, neither body models nor tedious manual tuning of parameters is required. Given the algorithms, the crucial ingredient for human pose estimation success was large-scale well annotated data sets of humans with the locations of the bodyparts.

Here, we identify **five** key areas that were important for making DNN-based pose estimation tools useful for neuroscience laboratories, and review the progress in the last two years:

1. Can DNNs be harnessed with small training datasets? Due to the nature of ‘small-scale’ laboratory experiments, labeling >20,000 or more frames is not a feasible approach (the typical human benchmark dataset sizes).
2. The end-result must be as accurate as a human manually-applied labels (i.e. the gold standard), and computationally tractable (fast).
3. The resulting networks should be robust to changes in experimental setups, and for long-term storage and re-analysis of video data, to video compression.
4. Animals move in 3D, thus having efficient solutions for 3D pose estimation would be highly valuable, especially in the context of studying motor learning and control.
5. Tracking multiple subjects and objects is important for many experiments studying social behaviors as well as for animal-object interactions.

1. Small training sets for lab-sized experiments

While the challenges discussed above for human pose estimation also apply for other animals, one important challenge for applying these methods to neuroscience was annotated data sets — could DNNs be harnessed for much smaller datasets, at sizes reasonable for typical labs? Thus, while it was clear that given enough annotated frames the same algorithms will be able to learn to track

the body parts of any animal, there were feasibility concerns.

Human pose estimation networks are typically trained on thousands of images, and nearly all the current state-of-the-art networks provide tailored solutions that utilize the human skeleton structure during inference [19^{••}, 30]. Thus, applying these tools to new datasets was not immediately straight-forward, and to create animal-specific networks one would need to potentially curate large datasets of the animal(s) they wanted to track. Additionally, researchers would need tailored DNNs to track their subjects (plus the ability to track unique objects, such as the corners of a box, or an implanted fiber).

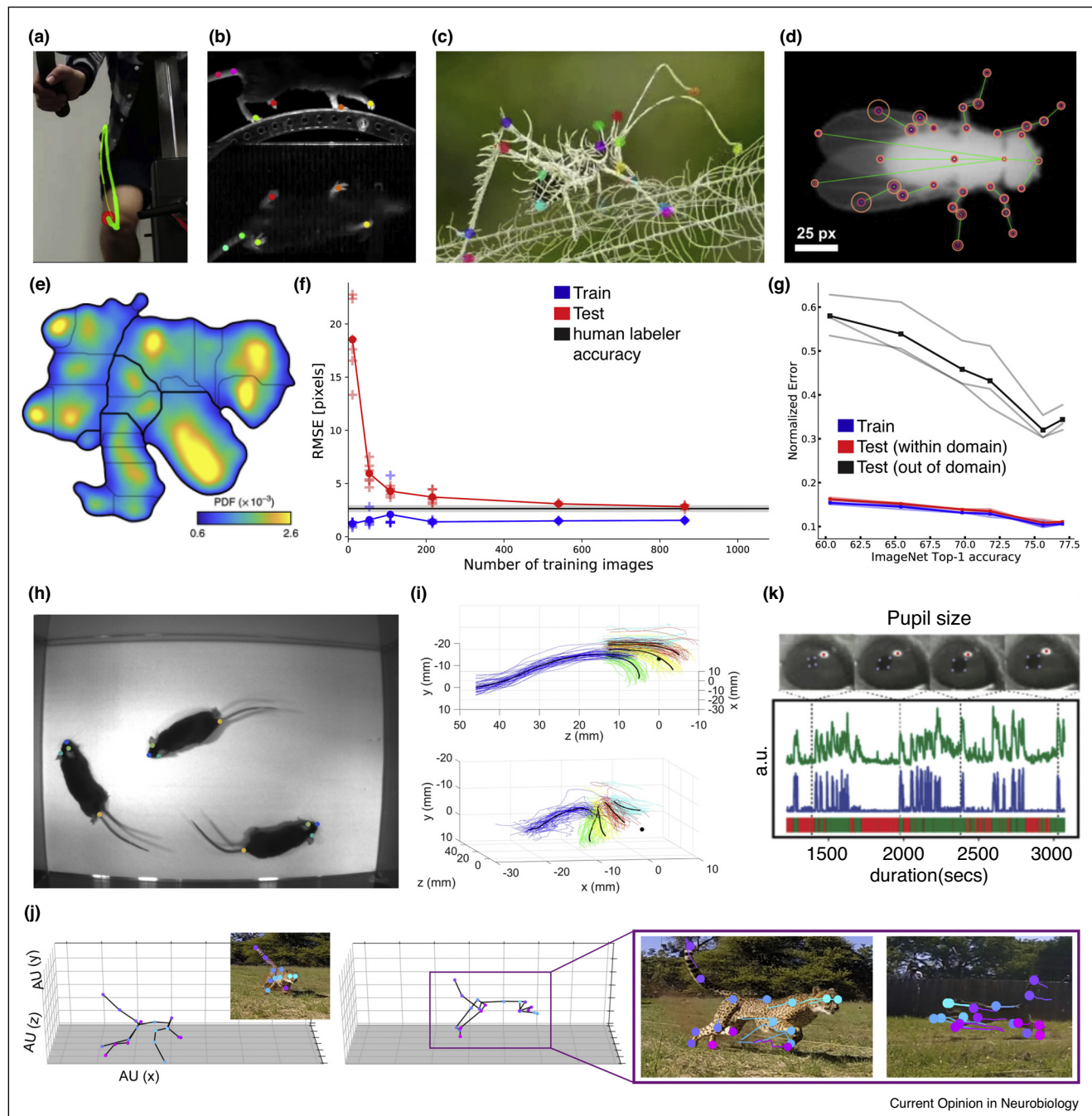
Thus, one of the most important challenges is creating tailored DNNs that are robust and generalize well with little training data. One potential solution for making networks for animal pose estimation that could generalize well, even with little data, was to use transfer learning — the ability to take a network that has been trained on one task to perform another. The advantage is that these networks are pretrained on larger datasets (for different tasks where a lot of data is available like ImageNet), therefore they are effectively imbued with good image representations. This is indeed what ‘DeepLabCut’, the first tool to leverage the advances in human pose estimation for application to animals did [48^{••}]. DeepLabCut was built on a subset of ‘DeeperCut’ [29^{••}], which was an attractive option due to its use of ResNets, which are powerful for transfer learning [49,50]. Moreover transfer learning reduces training times [49,50,51[•]], and there is a significant gain over using randomly-initialized networks in performance, especially for smaller datasets [50].

The major result from DeepLabCut was benchmarking on smaller datasets and finding that only a few hundred annotated images are enough to achieve excellent results for diverse pose estimation tasks like locomotion, reaching and trail-tracking in mice, egg-laying in flies and hunting in cheetahs, due to transfer learning (Figure 2f,g,h) [50,52,53[•]]. ‘DeepBehavior’, which utilized different DNN-packages for various pose estimation problems, also illustrated the gain of transfer learning [51[•]].

2. Accuracy & speed

To be useful, pose estimation tools need to be as good as human annotation of frames (or tracking markers, another proxy for a human-applied label). DeepLabCut was shown to reach human labeling accuracy [48^{••}], and can replace physical markers [58]. Moreover, they need to be efficient (fast) for both offline analysis and online analysis. Speed is often related to the depth of the network. Stacked-hourglass networks, which use iterative refinement [28,59] and fewer layers, are fast. Two toolboxes,

Figure 2



DNNs applied to animal pose estimation. **(a)** Knee tracking during cycling adopted from [54]. **(b)** 3D limb estimates from [52]. **(c)** A Lichen Katydid tracked with DeepLabCut, courtesy of the authors. **(d)** Fly with LEAP annotated body parts. The circles indicate the 90% fraction for which predicted positions of the particular body part are closer to the ground truth than the radii on test images (adopted from [55]). **(e)** Density plot of t-SNE plus frequency-transformed freely moving fly body-part trajectories. Patches with higher probability indicate more common movements like different types of grooming behaviors (adopted from [55]); **(f)** DeepLabCut requires little data to match human performance. Average error (RMSE) for several splits of training and test data versus number of training images compared to RMSE of a human scorer. Each split is denoted by a cross, the average by a dot. For 80% of the data, the algorithm achieves human level accuracy on the test set. As expected, test RMSE increases for fewer training images. Around 100 diverse frames are enough to provide high tracking performance (<5-pixel accuracy — adopted from [48]). **(g)** Networks that perform better on ImageNet perform better for predicting 22 body parts on horses on within-domain (similar data distribution as training set, red) and out-of-domain data (novel horses, black). The faint lines are individual splits. (adopted from [50]). **(h)** Due to the convolutional network architecture, when trained on one mouse the network generalizes to detect body parts of three mice (adopted from [48]). **(i)** 3D reaching kinematics of rat (adopted from [56]). **(j)** 3D pose estimation on a cheetah for 2 example poses from 6 cameras as well as example 2D views (adopted from [53]). **(k)** Pupil and pupil-size tracking (adopted from [57]).

‘LEAP’ [55[•]] and ‘DeepPoseKit’ [60] adopted variants of stacked-hourglass networks. LEAP allows the user to rapidly compute postures, and then perform unsupervised behavioral analysis (Figure 2d,e) [61]. This is an attractive solution for real-time applications, but it is not quite as accurate. For various datasets, DeepPoseKit reports it is about three times as accurate as LEAP, yet similar to DeepLabCut [60]. They also report about twice faster video processing compared to DeepLabCut and LEAP for batch-processing (on small frame sizes).

Deeper networks are slower, but often have more generalization ability [49]. DeepLabCut was designed for generalization and therefore utilized deeper networks (ResNets) that are inherently slower than stacked-hourglass networks, yet DeepLabCut can match human accuracy in labeling (Figure 2f) [48^{••}]. The speed has been shown to be compatible with online-feedback applications (>50 Hz for reasonable frame sizes) [62,63,52]. Other networks recently added to DeepLabCut (with a MobileNetV2 backbone) give slightly lower accuracy, with twice the speed [50]. Overall, on GPU hardware all packages are fast, and reach speeds of several hundred frames per second in offline modes.

3. Robustness

Neuroscience experiments based on video recordings produce large quantities of data and are collected over extensive periods of time. Thus, analysis pipelines should be robust to a myriad of perturbations: such as changes in setups (backgrounds, light sources, cameras, etc.), subject appearance (due to different animal strains), and compression algorithms (which allow storage of perceptually good videos with little memory demands [64]).

How can robustness be increased? Both transfer learning (discussed above) and data augmentation strategies are popular and rapidly evolving approaches to increase robustness in DNNs (see review [65]). Moreover, active learning approaches allow an experimenter to continuously build more robust and diverse datasets, for large scale projects by expanding the training set with images, where the network fails [48^{••},53[•],55[•]]. So far, the toolboxes have been tested on data from the same distribution (i.e. by splitting frames from videos into test and training data), which is important for assessing the performance [55[•], 60], but did not directly tested out-of-domain robustness.

Over the course of long-term experiments the background or even animal strain can change, which means having robust networks would be highly advantageous. We recently tested the generalization ability of DeepLabCut with different network backbones for pose estimation. We find that pretraining on ImageNet strongly improves out-of-domain performance, and that better ImageNet performing networks are more robust

(Figure 2g) [50]. There is still a gap to close in out-of-domain performance, however.

DeepLabCut is also robust to video compression, as compression by more than 1,000X only mildly affects accuracy (less than 1 pixel average error less) [52]. The International Brain Lab (IBL) independently and synergistically showed that for tracking multiple body parts in a rodent decision making task, DeepLabCut is robust to video compression [66]. Thus, in practice users can substantially compress videos, while retaining accurate posture information.

4. 3D animal pose estimation

Currently, there are several animal pose estimation toolboxes that explicitly support 2D and 3D key-point detection [53[•], 67, 68, 69]. DeepLabCut uses 2D pose estimation to train a single camera-invariant 2D network (or multiple 2D networks) that is then used to perform traditional triangulation to extract 3D key points (Figure 2i,j; [53[•],56]). A pipeline built on DeepLabCut called ‘Anipose’ allows for 3D reconstruction from multiple cameras using a wider variety of methods [67]. ‘DeepFly3D’ [68] uses the network architecture from Newell *et al.* [28] and then adds tools to compute an accurate 3D estimate of *Drosophila melanogaster* by using the fly itself versus standard calibration boards. Zhang *et al.* use epipolar geometry to train across views and thereby improve 3D pose estimation for mice, dogs, and monkeys [69].

5. Multi-animal & object tracking

Many experiments in neuroscience require measuring interactions of multiple animals or interactions with objects. Having the ability to both flexibly track user-defined objects or multiple animals therefore is highly desirable. There are many pre-deep learning algorithms that allow tracking of objects (one such modern example called ‘Tracktor’ also nicely summarizes this body of work [70]). Recently researchers have also applied deep learning to this problem. For example, the impressive ‘idTracker.ai’ [71] allows for users to track a hundred individual, unmarked animals. Arac *et al.* used YOLO, a popular and fast object localization network, for tracking two mice during a social behavior [51[•]]. These, and others, can then be combined with pose estimation packages for estimating the pose of multiple animals. Currently, two paths are possible: one is to apply pose estimation algorithms after tracking individuals (for which any package could be used); or, two, extract multiple detections for each part on each animal (Figure 2h; [48^{••},72]) and link them using part affinity fields [19^{••}], pairwise predictions [29^{••}], or geometrical constraints [72], plus combinatorics.

The impact on experimental neuroscience

In the short time period these tools have become available there has been a rather wide adoption by the neuroscience and ethology community. Beyond the original

publications, DeepLabCut has already been used for pose estimation and behavioral analysis in many contexts. For instance, knee movement quantification during cycling (Figure 2a) [54], postural analysis during underwater running [73], social behavior in bats [74], for benchmarking thermal constraints with optogenetics [75], fly leg movement analysis [76,77], for 3D rat reaching (Figure 2i) [56], hydra in a thermal stimulation assay [78] and pupillometry (Figure 2k) [57]. Also inanimate objects can be tracked, and it has indeed also been used to track metal beads when subjected to a high voltage [79], and magic tricks (i.e. coins and the magician) [80]. LEAP [55•] has been used to track ants [81] and mice [82].

Pose estimation is just the beginning; the next steps involve careful analysis of kinematics, building detailed, multi-scale ethograms of behaviors, new modeling techniques to understand large-scale brain activity and behaviors across a multitude of timescales, and beyond. We envision three branches where powerful feature tracking and extensions will be useful: motor control studies (often involving complex motor actions), naturalistic behaviors in the lab and in the wild, and better quantification of robust and seemingly simple ‘non-motor’ tasks (Figure 3).

Many paradigms in neuroscience can be loosely arranged along three branches as natural (i.e. mouse parenting behavior), simple trial-based tasks (i.e. classical conditioning), and/or complex motor actions like skilled reaching (Figure 3). For example, you can have simple and natural tasks such as licking for water, or complex and natural behaviors such as escaping from a looming stimulus that would rarely produce repeated trajectories. For simplicity, here we discuss how pose estimation can potentially enhance these studies along those three branches, namely complex movements (Motor control & kinematics), natural behaviors (Natural behaviors & ethologically relevant features), and during simple motor-output tasks (Revisiting classic tasks).

Motor control & kinematics

Often in neuroscience-minded motor control studies end-effector proxies (such as manipulanda or joysticks) are used to measure the motor behavior of subjects or animals. There are relatively few marker-tracking based movement neuroscience studies, in which many degrees of freedom were measured alongside neural activity, with notable exceptions like [83,84]. The ease with which kinematics of limbs and digits can now be quantified [48•,51•,56,77] should greatly simplify such studies in the future. We expect many more highly detailed kinematic studies will emerge that utilize DNN-based analyses, especially for freely moving animals, for small and aquatic animals that cannot be marked, and for motor control studies that can leverage large-scale recordings and behavioral monitoring.

Natural behaviors & ethologically relevant features

There is a trend in motor neuroscience towards natural behaviors; that is, less constrained tasks, everyday-skills, and even ‘in the wild’ studies [85]. For instance, we used DeepLabCut for 3D pose estimation in hunting cheetah’s captured via multiple Go-Pro cameras (Figure 2j; [53•]). Another ‘in the wild example’ is given by a recent study by Chambers *et al.* [86], who revisited the classic question of how people synchronize their walking, but with a modern twist by using videos from YouTube and analysis with OpenPose [19•]. Consistent with studies performed in the laboratory, they found a tendency for pairs of people to either walk in or exactly out of phase [86].

How else can pose estimation help? Specialized body parts often play a key role in ethologically relevant behaviors. For instance, ants use their antenna to follow odor trails [87], while moles use their snouts for sampling bilateral nasal cues to localize odorants [88]. To accurately measure such behaviors, highly accurate feature-detectors of often tiny, highly dexterous bodyparts are needed. This is a situation where deep learning algorithms can excel. Pose estimation algorithms can not only be used to detect the complete ‘pose’, but due to their flexibility they are extremely useful to track ethologically relevant body parts in challenging situations; incidentally DeepLabCut was created, in part, to accurately track the snouts of mice following odor trails that were printed onto a treadmill [48•]. There are of course many other specialized body parts that are hard to track: like whiskers, bee-stingers, jellyfish tentacles, or octopus arms, and we believe that studying these beautiful systems in more natural and ethologically relevant environments has now gotten easier.

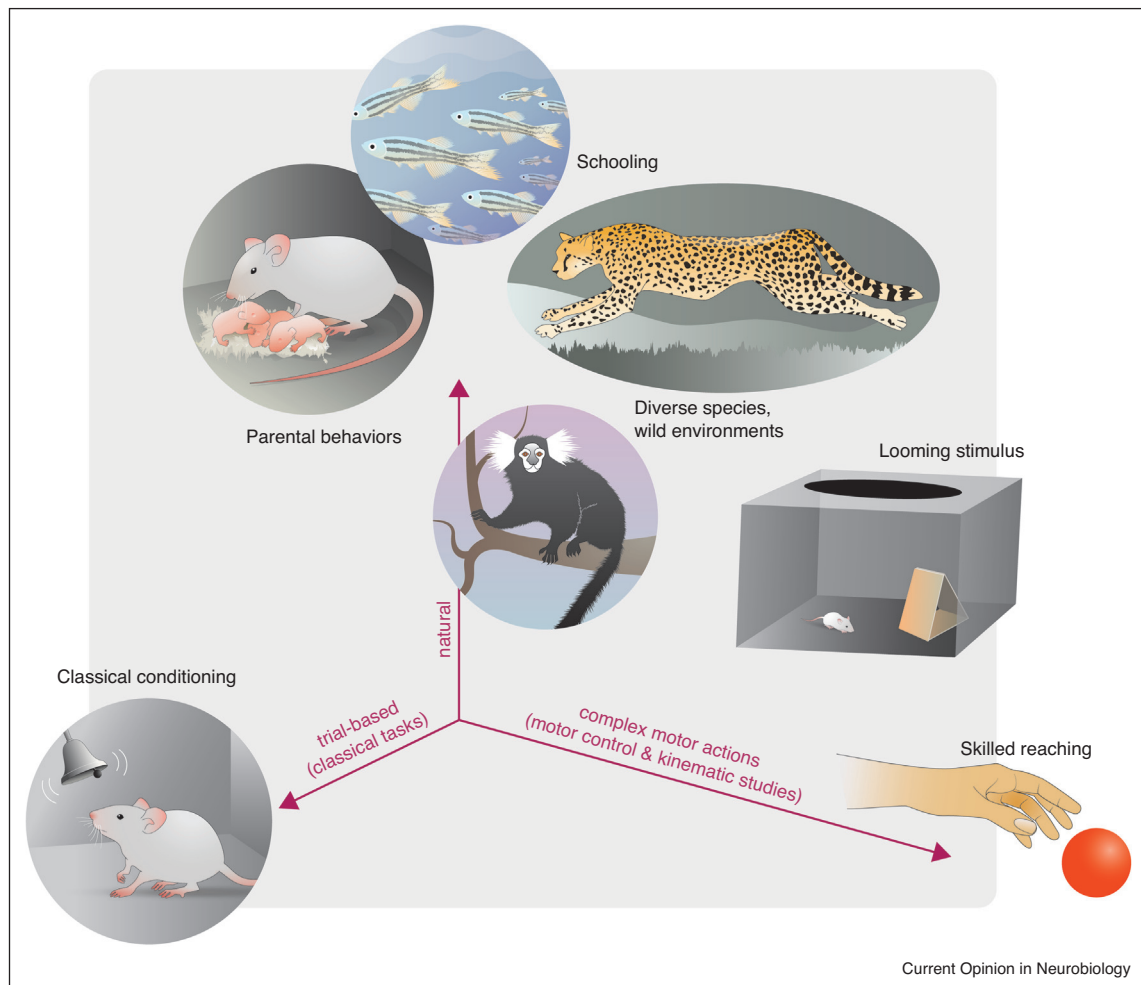
Revisiting classic tasks

Measuring behavior is already impacting ‘classic’ decision-making paradigms. For example, several groups could show broad movement encoding across the brain during decision-making tasks by carefully quantifying behavior [89,90]. Moreover, large scale efforts to use these ‘simple’ yet robust trial-based behaviors across labs and brain areas are leveraging deep learning, and comparing their utility compared to classical behavior-monitoring approaches. For example, the IBL has surmised that DeepLabCut could replace traditional methods used for eye, paw and lick detection [66]. We believe that detailed behavioral analysis will impact many paradigms, which were historically not considered ‘motor’ studies, as now it is much easier to measure movement.

Remaining challenges in pose estimation

Advances in deep learning have changed how easily posture can be measured and has impacted many studies. However, pose estimation remains a hard computer vision problem and challenges remain [16•,27,33,91–93]. In the context of multi-animal/human pose estimation, dealing with highly

Figure 3



The behavioral space in neuroscience: new applications for deep learning-assisted analysis. This diagram depicts how pose estimation with non-invasive videography can benefit behavioral paradigms that span from 'trial-based behaviors' such as classical conditioning, to 'complex motor behaviors/tasks' as in 3D reaching assays, to 'naturalistic tasks', often without any trial structure, and that are more akin to real-world 'tasks' that animals undertake. For example, a looming stimulus is ethologically relevant and complex, and pup-retrieval can be natural yet repeatable in a laboratory setting. With new tools that allow for fast and accurate analysis of movement, these types of experiments become more feasible (with much less human labor than previously required).

crowded scenes, in which different individuals cover each other, remains highly challenging [27,33,91]. In general, dealing with occlusions remains a challenge. In some experiments occlusions are hard to avoid. Thus, networks that can constraint body part detection based on anatomical relationships can be advantageous, but are computationally more complex and slower [29[•]]. As we highlighted in the robustness section, it is hard to train networks to generalize to out-of-domain scenarios [92,50,94]. Even though very large data sets have been amassed to build robust DNNs [19[•],30–32], they can still fail in sufficiently different scenarios [92,50,93]. Making robust networks will be highly useful for creating shareable behavior- or animal-specific networks that can generalize across laboratories. There will also be much work towards even faster, lighter models.

Outlook & conclusions

The field of 2D, 3D, and dense pose estimation will continue to evolve. For example, with respect to handling occlusions and robustness to out-of-domain data. Perhaps larger and more balanced datasets will be created to better span the behavioral space, more temporal information will be utilized when training networks or analyzing data, and new algorithmic solutions will be found.

Will we always need training data? A hot topic in object recognition is training from very few examples (one-shot or zero-shot learning) [95]. Can this be achieved in pose estimation? Perhaps as new architectures and training regimes come to fruition this could be possible. Alternatively, specialized networks could now be built that

leverage large datasets of specific animals. It is hard to envision a universal ‘animal pose detector’ network (for object recognition this is possible) as animals have highly diverse body plans and experimentalists often have extremely different needs. Currently many individual labs create their own specialized networks, but we plan to create shareable networks for specific animals (much like the specific networks, that is, hand network in OpenPose [96], or the human-network in DeepLabCut [29**]). For example, many open field experiments could benefit from robust and easy-to-use DNNs for video analysis across similar body points of the mouse. Indeed, efforts are underway to create networks where one can simply analyze their data without training, and we hope the community will join these efforts. Nothing improves DNNs more than more training data. These efforts, together with making code open source, will contribute to the reproducibility of science and make these tools broadly accessible.

In summary, we aimed to review the progress in computer vision for human pose estimation, how it influenced animal pose estimation, and how neuroscience laboratories can leverage these tools for better quantification of behavior. Taken together, the tremendous advance of computer vision has provided tools that are practical for the use in the laboratory, and they will only get better. They can be as accurate as human-labeling (or marker-based tracking), and are fast enough for closed-loop experiments, which is key for understanding the link between neural systems and behavior. We expect that in-light of shared, easy-to-use tools and additional deep learning advances, there will be thrilling and unforeseen advances in real-world neuroscience.

Conflict of interest statement

Nothing declared.

Acknowledgements

We thank Eldar Insafutdinov, Alex Gomez-Marin, the M Mathis Lab, and the anonymous reviewers for comments on the manuscript, and Julia Kuhl for illustrations. Funding was provided by the Rowland Institute at Harvard University (MWM) and NIHU19MH114823 (AM).

References and recommended reading

Papers of particular interest, published within the period of review, have been highlighted as:

- of special interest
- of outstanding interest

1. Uexküll JB: *Streifzüge durch die Umwelten von Tieren und Menschen Ein Bilderbuch unsichtbarer Welten*. Berlin, Heidelberg: Springer; 1956.
2. Tinbergen N: **On aims and methods of Ethology**. *Zeit Tierpsychol* 1963, **20**(4):410-433.
3. Bernstein NA: *The Co-ordination and Regulation of Movements, vol 1*. New York, Oxford: Pergamon Press; 1967.
4. Tinbergen N: *Nobel Lecture. The Nobel Prize in Physiology or Medicine 1973*. 1973.
5. Schaefer AT, Claridge-Chang A: **The surveillance state of behavioral automation**. *Curr Opin Neurobiol* 2012.
6. Anderson DJ, Perona P: **Toward a science of computational ethology**. *Neuron* 2014, **84**(1):18-31.
7. Dell AI, Bender JA, Branson K, Couzin ID, de Polavieja GG, Noldus LPJJ, Pérez-Escudero A, Perona P, Straw AD, Wikelski M et al.: **Automated image-based tracking and its application in ecology**. *Trends Ecol Evol* 2014, **29**(7):417-428.
8. Egnor R, Branson K: **Computational analysis of behavior**. *Annu Rev Neurosci* 2016, **39**:217-236.
9. Camomilla V, Bergamini E, Fantozzi S, Vannozzi G: **Trends supporting the in-field use of wearable inertial sensors for sport performance evaluation: a systematic review**. *Sensors* 2018, **18**(3):873.
10. Gomez-Marin A, Paton JJ, Kampff AR, Costa RM, Mainen ZF: **Big behavioral data: psychology, ethology and the foundations of neuroscience**. *Nat Neurosci* 2014, **17**(11):1455.
11. Poppe R: **Vision-based human motion analysis: an overview**. *Comput Vis Image Understand* 2007, **108**(1):4-18 <http://dx.doi.org/10.1016/j.cviu.2006.10.016> ISSN 1077-3142. Special Issue on Vision for Human-Computer Interaction.
12. Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, Van Der Laak JA, Van Ginneken B, Sánchez CI: **A survey on deep learning in medical image analysis**. *Med Image Anal* 2017, **42**:60-88.
- Comprehensive review of all deep learning algorithms used in medical image analysis as of 2017, as well as a discussion of most successful approaches, together with [16**] a fantastic introduction for newcomers to the field.
13. Hogg D: **Model-based vision: a program to see a walking person**. *Image Vis Comput* 1983, **1**(1):5-20 [http://dx.doi.org/10.1016/0262-8856\(83\)90003-3](http://dx.doi.org/10.1016/0262-8856(83)90003-3) ISSN 0262-8856.
14. Wren CR, Azarbayejani A, Darrell T, Pentland PA: **Pfinder: real-time tracking of the human body**. *IEEE Trans Pattern Anal Mach Intell* 1997, **19**(7):780-785.
15. Cremers D, Rousson M, Deriche R: **A review of statistical approaches to level set segmentation: integrating color, texture, motion and shape**. *Int J Comput Vis* 2007, **72**(2):195-215.
16. Serre Thomas: **Deep learning: the good, the bad, and the ugly**. *Annu Rev Vis Sci* 2019, **5**:399-426 <http://dx.doi.org/10.1146/annurev-vision-091718-014951>
- Excellent review of deep learning progress including a detailed description of recent successes as well as limitations of computer vision algorithms.
17. Alom MZ, Taha TM, Yakopcic C, Westberg S, Sidike P, Nasrin MS, Van Esen BC, Awwal AAS, Asari VK: *The History Began from Alexnet: A Comprehensive Survey on Deep Learning Approaches*. 2018arXiv:1803.01164.
18. Schmidhuber J: **Deep learning in neural networks: an overview**. *Neural Netw*. 2015, **61**:85-117 <http://dx.doi.org/10.1016/j.neunet.2014.09.003> ISSN 0893-6080.
19. Cao Z, Simon T, Wei S-E, Sheikh Y: **Realtime multi-person 2d pose estimation using part affinity fields**. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 2017
- OpenPose was the first real-time multi-person system to jointly detect human body parts by using part affinity fields, a great way to link body part proposals across individuals. The toolbox is well maintained and now boasts body, hand, facial, and foot keypoints (in total 135 keypoints) as well as 3D support.
20. Mehta D, Rhodin H, Casas D, Sotnychenko O, Xu W, Theobalt C: **Monocular 3d Human Pose Estimation Using Transfer Learning and Improved CNN Supervision**. 2016arXiv:1611.09813.
21. Alp Güler R, Neverova N, Kokkinos I: **Densepose: dense human pose estimation in the wild**. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 2018:7297-7306
- Using a large dataset of humans (50K), they build dense correspondences between RGB images and human bodies. They apply this to human ‘in the

wild', and build tools for efficiently dealing with occlusions. It is highly accurate and runs up to 25 frames per second.

22. Zuffi S, Kanazawa A, Berger-Wolf T, Black M: **Three-d safari: learning to estimate zebra pose, shape, and texture from images "in the wild"**. *ICCV. IEEE Computer Society, 08 2019*
Zuffi *et al.* push dense pose estimations by using a new SMALST model for capturing zebras pose, soft-tissue shape, and even texture 'in the wild'. This is a difficult challenge as zebras are designed to blend into the background in the safari. This paper makes significant improvements on accuracy and realism, and builds on a line of elegant work from these authors [44,46]
23. Toshev A, Szegedy C: **DeepPose: human pose estimation via deep neural networks**. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2014*:1653-1660.
24. Jain A, Tompson J, LeCun Y, Bregler C: **Modeep: A Deep Learning Framework Using Motion Features for Human Pose Estimation**. 2014arXiv:1409.7963.
25. Tompson JJ, Jain A, LeCun Y, Bregler C: **Joint training of a convolutional network and a graphical model for human pose estimation**. In *Advances in Neural Information Processing Systems 27*. Edited by Ghahramani Z, Welling M, Cortes C, Lawrence ND, Weinberger KQ. Curran Associates, Inc.; 2014:1799-1807.
26. Tompson J, Goroshin R, Jain A, LeCun Y, Bregler C: **Efficient object localization using convolutional networks**. *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2015*.
27. Andriluka M, Iqbal U, Insafutdinov E, Pishchulin L, Milan A, Gall J, Schiele B: **Posetrack: a benchmark for human pose estimation and tracking**. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2018*:5167-5176.
28. Newell A, Yang K, Deng J: **Stacked hourglass networks for human pose estimation**. *European Conference on Computer Vision; Springer: 2016*:483-499.
29. Insafutdinov E, Pishchulin L, Andres B, Andriluka M, Schiele B: **DeeperCut: a deeper, stronger, and faster multi-person pose estimation model**. *European Conference on Computer Vision 2016*:34-50. Springer
DeeperCut is a highly accurate algorithm for multi-human pose estimation due to improved deep learning based body part detectors, and image-conditioned pairwise terms to predict the location of body parts based on the location of other body parts. These terms are then used to find accurate poses of individuals via graph cutting. In ArtTrack [30] the work was extended to fast multi-human pose estimation in videos.
30. Insafutdinov E, Andriluka M, Pishchulin L, Tang S, Levinkov E, Andres B, Schiele B: **Arttrack: articulated multi-person tracking in the wild**. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2017*.
31. Kreiss S, Bertoni L, Alahi A: **Pifpaf: composite fields for human pose estimation**. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2019*:11977-11986.
32. Sun K, Xiao B, L D, Wang J: **Deep High-Resolution Representation Learning for Human Pose Estimation**. 2019arXiv:1902.09212.
33. Dang Q, Yin J, Wang B, Zheng W: **Deep learning based 2d human pose estimation: a survey**. *Tsinghua Sci Technol 2019, 24(6)*:663-676 <http://dx.doi.org/10.26599/TST.2018.9010100> ISSN 1007-0214.
34. Sarafianos N, Boteanu B, Ionescu B, Kakadiaris IA: **3d human pose estimation: a review of the literature and analysis of covariates**. *Comput Vis Image Understand 2016, 152*:1-20.
35. Martinez J, Hossain R, Romero J, Little JJ: **A simple yet effective baseline for 3d human pose estimation**. *Proceedings of the IEEE International Conference on Computer Vision 2017*:2640-2649.
36. Tomè D, Russell C, Agapito L: **Lifting from the Deep: Convolutional 3d Pose Estimation from a Single Image**. 2017arXiv:1701.00295.
37. Chen C-H, Ramanan D: **3d human pose estimation = 2d pose estimation+ matching**. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2017*:7035-7043.
38. Shotton J, Girshick RB, Fitzgibbon AW, Sharp T, Cook M, Finocchio M, Moore R, Kohli P, Criminisi A, Kipman A, Blake A: **Efficient human pose estimation from single depth images**. *IEEE Trans Pattern Anal Mach Intell 2012, 35*:2821-2840.
39. Obdržálek Š, Kurillo G, Ofli F, Bajcsy R, Seto E, Jimison H, Pavel M: **Accuracy and robustness of kinect pose estimation in the context of coaching of elderly population.. 2012 Annual International Conference of the IEEE Engineering in Medicine and Biology Society 2012**:1188-1193 <http://dx.doi.org/10.1109/EMBC.2012.6346149>.
40. Wilschko AB, Johnson MJ, Iurilli G, Peterson RE, Katon JM, Pashkovski SL, Abaira VE, Adams RP, Datta SR: **Mapping sub-second structure in mouse behavior**. *Neuron 2015, 88(6)*:1121-1135.
41. Hong W, Kennedy A, Burgos-Artizzu XP, Zelikowsky M, Navonne SG, Perona P, Anderson DJ: **Automated measurement of mouse social behaviors using depth sensing, video tracking, and machine learning**. *Proc Natl Acad Sci U S A 2015, 112(38)*:E5351-E5360.
42. Pavlakos G, Choutas V, Ghorbani N, Bolkart T, Osman AAA, Tzionas D, Black MJ: **Expressive body capture: 3d hands, face, and body from a single image**. *Proceedings IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2019*.
43. Kanazawa A, Black MJ, Jacobs DW, Malik J: **End-to-end recovery of human shape and pose**. *Computer Vision and Pattern Recognition (CVPR) 2018*.
44. Zuffi S, Kanazawa A, Jacobs D, Black M: **3d menagerie: modeling the 3d shape and pose of animals**. 2017 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2016*:11.
45. Biggs B, Roddick T, Fitzgibbon A, Cipolla R: **Creatures great and small: recovering the shape and motion of animals from video**. *Asian Conference on Computer Vision 2018*:3-19. Springer
The authors combine multiple methods in order to efficiently fit 3D shape to multiple quadrupeds from camels to bears. They also provide a novel dataset of joint annotations and silhouette segmentation for eleven animal videos.
46. Zuffi S, Kanazawa A, Black MJ: **Lions and tigers and bears: capturing non-rigid. 3d, articulated shape from images**. *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR) 2018*.
47. Nguyen G, Dlugolinsky S, Bobák M, Tran V, García ÁL, Heredia I, Malik P, Hluchý L: **Machine learning and deep learning frameworks and libraries for large-scale data mining: a survey**. *Artif Intell Rev Jun 2019, 52(1)*:77-124 <http://dx.doi.org/10.1007/s10462-018-09679-z> ISSN 1573-7462.
48. Mathis A, Mamidanna P, Cury KM, Abe T, Murthy VN, Mathis MW, Bethge M: **DeepLabCut: markerless pose estimation of user-defined body parts with deep learning**. *Nat Neurosci 2018, 21*:1281-1289
DeepLab-Cut was the first deep learning toolbox for animal pose estimation. The key advance was to benchmark a subset of the feature detectors in DeeperCut [29]. This paper showed nearly human-level performance with only 50-200 images. It benchmarked flies moving in a 3D chamber, hand articulations and open-field behavior in mice, and provided open-source tools for creating new datasets and data loaders to train deep neural networks, and post-hoc analysis. Subsequent work has improved accuracy, speed, and introduced more network variants into the Python package [50, 52, 53**].
49. Kornblith S, Shlens J, Le QV: **Do better imagenet models transfer better?** *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2019*:2661-2671.
50. Mathis A, Yüsekçönül M, Rogers B, Bethge M, Mathis MW: **Pretraining Boosts Out-of-Domain Robustness for Pose Estimation**. 2019arXiv:1909.11229.
51. Arac A, Zhao P, Dobkin BH, Carmichael T, Golshani P: **Deepbehavior: a deep learning toolbox for automated analysis of animal and human behavior imaging data**. *Front Syst Neurosci 2019, 13*:20
Arac *et al.* use three different DNN-packages (OpenPose [19*], YOLO and Tensorbox) for analyzing 3D analysis of pellet reaching, three-chamber test, social behavior in mice, and 3D human kinematics analysis

for clinical motor function assessment with OpenPose [19*]. They also provide MATLAB scripts for additional analysis (after pose estimation).

52. Mathis A, Warren RA: **On the inference speed and video-compression robustness of deeplabcut.** *bioRxiv* 2018 <http://dx.doi.org/10.1101/457242>.
53. Nath T, Mathis A, Chen AC, Patel A, Bethge M, Mathis MW: **Using deeplabcut for 3d markerless pose estimation across species and behaviors.** *Nat Protocols* 2019, **14**:2152-2176
 A Nature Protocols user-guide to DeepLabCut2.0, with 3D pose estimation of hunting cheetahs and improved network performance. The toolbox is provided as a Python package with graphical user interfaces for labeling, active-learning-based network refinement, together with Jupyter Notebooks that can be run on cloud resources such as Google Colaboratory (for free).
54. Kaplan O, Yamamoto G, Taketomi T, Plopski A, Kato H: **Video-based visualization of knee movement in cycling for quantitative and qualitative monitoring.** *2019 12th Asia Pacific Workshop on Mixed and Augmented Reality (APMAR)* 2019:1-5.
55. Pereira TD, Aldarondo DE, Willmore L, Kislin M, Wang SS-H, Murthy M, Shaevitz JW: **Fast animal pose estimation using deep neural networks.** *Nat Methods* 2019, **16**(1):117
 LEAP (LEAP estimates animal pose), a deep learning method for predicting the positions of animal body parts. This framework consists of a graphical interface for labeling of body parts and training the network. Training and inference times are fast due to the lightweight architecture. The authors also analyzed insect gait based on unsupervised behavioral methods [61], which they directly applied to the posture, rather than to compressed image features.
56. Bova A, Kernodle K, Mulligan K, Leventhal D: **Automated rat single-pellet reaching with 3-dimensional reconstruction of paw and digit trajectories.** *J Visual Exp* 2019.
57. Sriram B, Li L, Cruz-Martin A, Ghosh A: **A sparse probabilistic code underlies the limits of behavioral discrimination.** *Cereb Cortex* 2019.
58. Moore DD, Walker JD, MacLean JN, Hatsopoulos NG: **Comparing pose estimation performance using deeplabcut and marker-based tracking.** *Soc Neurosci* 2019.
59. Badrinarayanan V, Kendall A, Cipolla R: **Segnet: a deep convolutional encoder-decoder architecture for image segmentation.** *IEEE Trans Pattern Anal Mach Intell* 2017, **39**(12):2481-2495.
60. Graving JM, Chae D, Naik H, Li L, Koger B, Costelloe BR, Couzin ID: **DeepPoseKit, a software toolkit for fast and robust animal pose estimation using deep learning.** *eLife* 2019, **8**:e47994 <http://dx.doi.org/10.7554/eLife.47994> ISSN 2050-084X.
61. Berman GJ, Choi DM, Bialek W, Shaevitz JW: **Mapping the stereotyped behaviour of freely moving fruit flies.** *J R Soc Interface* 2014, **11**(99) <http://dx.doi.org/10.1098/rsif.2014.0672> ISSN 1742-5689.
62. Forsy B, Xiao D, Gupta P, Boyd JD, Murphy TH: **Real-time markerless video tracking of body parts in mice using deep neural networks.** *bioRxiv* 2018:482349.
63. Štih V, Petrucco L, Kist AM, Portugues R: **Stytra: an open-source, integrated system for stimulation, tracking and closed-loop behavioral experiments.** *PLoS Comput Biol* 2019, **15**(4):e1006699.
64. Wiegand T, Sullivan GJ, Bjontegaard G, Luthra A: **Overview of the H.264/AVC video coding standard.** *IEEE Trans Circuits Systems Video Technol* 2003, **13**(7):560-576.
65. Shorten C, Khoshgftaar TM: **A survey on image data augmentation for deep learning.** *J Big Data* Jul 2019, **6**(1):60 <http://dx.doi.org/10.1186/s40537-019-0197-0> ISSN 2196-1115.
66. Meijer G, Schartner M, Chapuis G, Svoboda K: **Video hardware and software for the international brain laboratory.** *Int Brain Lab* 2019.
67. Karashchuk P: *lambdaloop/Anipose: v0.5.0.* 2019.
68. Günel S, Rhodin H, Morales D, Campagnolo JH, Ramdya P, Fua P: **Deepfly3d, a deep learning-based approach for 3d limb and appendage tracking in tethered, adult Drosophila.** *eLife* 2019, **8**:e48571 <http://dx.doi.org/10.7554/eLife.48571> ISSN 2050-084X.
69. Zhang Y, Park HS: **Multiview Supervision by Registration.** *2019arXiv:1811.11251v2*.
70. Sridhar VH, Roche DG, Gingins S: **Tracktor: image-based automated tracking of animal movement and behaviour.** *Methods Ecol Evol* 2019, **10**(6):815-820 <http://dx.doi.org/10.1111/2041-210X.13166>.
71. Romero-Ferrero F, Bergomi MG, Hinz RC, Heras FJH, de Polavieja GG: **idtracker. ai: tracking all individuals in small or large collectives of unmarked animals.** *Nat Methods* 2019, **16**(2):179.
72. Jiang Z, Liu Z, Chen L, Tong L, Zhang X, Lan X, Crookes D, Yang M-H, Zhou H: **Detection and Tracking of Multiple Mice Using Part Proposal Networks.** *2019arXiv:1906.02831*.
73. Cronin NJ, Rantalainen T, Ahtiainen JP, Hynynen E, Waller B: **Markerless 2d kinematic analysis of underwater running: a deep learning approach.** *J Biomech* 2019, **87**:75-82.
74. Zhang W, Yartsev MM: **Correlated neural activity across the brains of socially interacting bats.** *Cell* 2019:178.
75. Owen SF, Liu MH, Kreitzer AC: **Thermal constraints on in vivo optogenetic manipulations.** *Nat Neurosci* 2019:22.
76. Bidaye SS, Laturney M, Chang AK, Liu Y, Bockemuhl T, Buschges A, Scott K: **Two brain pathways initiate distinct forward walking programs in drosophila.** *bioRxiv* 2019:798439.
77. Azevedo AW, Gurung P, Venkatasubramanian L, Mann R, Tuthill JC: **A size principle for leg motor control in drosophila.** *bioRxiv* 2019:730218.
78. Tzouanas CN, Kim S, Badhiwala KN, Avants BW, Robinson J: **Thermal stimulation temperature is encoded as a firing rate in a hydra nerve ring.** *bioRxiv* 2019:787648.
79. De Bari B, Dixon JA, Kay BA, Kondepudi D: **Oscillatory dynamics of an electrically driven dissipative structure.** *PLOS ONE* 2019, **14**(5):e0217305.
80. Zaghi-Lara R, Gea MÁ, Camí J, Martínez LM, Gomez-Marin A: **Playing Magic Tricks to Deep Neural Networks Untangles Human Deception.** *2019arXiv:1908.07446*.
81. Clifton G, Holway D, Gravish Nick: **Rough substrates constrain walking speed in ants through modulation of stride frequency and not stride length.** *bioRxiv* 2019:731380.
82. Yue Y, Xu P, Liu Z, Chen Z, Su J, Chen A, Ash RT, Barut E, Simha R, Smirnakis S, Zeng C, Lu H: **Mecp2 deletion impaired layer 2/3-dominant dynamic reorganization of cortical circuit during motor skill learning.** *bioRxiv* 2019 <http://dx.doi.org/10.1101/786822>.
83. Vargas-Irwin CE, Shakhnarovich G, Yadollahpour P, Mislow JMK, Black MJ, Donoghue JP: **Decoding complete reach and grasp actions from local primary motor cortex populations.** *J Neurosci* 2010, **30**(29):9659-9669.
84. Schaffelhofer S, Agudelo-Toro A, Scherberger H: **Decoding a wide range of hand configurations from macaque motor, premotor, and parietal cortices.** *J Neurosci* 2015, **35**(3):1068-1081.
85. Mathis A, Pack AR, Maeda RS, McDougale SD: **Highlights from the 29th annual meeting of the society for the neural control of movement.** *J Neurophysiol* 2019, **122**:1777-1783 <http://dx.doi.org/10.1152/jn.00484.2019>.
86. Chambers C, Kong G, Wei K, Kording K: **Pose estimates from online videos show that side-by-side walkers synchronize movement under naturalistic conditions.** *PLOS ONE* 2019, **14**(6):e0217861.
87. Draft RW, McGill MR, Kapoor V, Murthy VN: **Carpenter ants use diverse antennae sampling strategies to track odor trails.** *J Exp Biol* 2018, **221**(22):jeb185124.
88. Catania KC: **Stereo and serial sniffing guide navigation to an odour source in a mammal.** *Nat Commun* 2013, **4**:1441.

89. Stringer C, Pachitariu M, Steinmetz N, Reddy CB, Carandini M, Harris KD: **Spontaneous behaviors drive multidimensional, brainwide activity.** *Science* 2019, **364**:6437 <http://dx.doi.org/10.1126/science.aav7893> ISSN 0036-8075..
90. Kaufman MT, Juavinett AL, Gluf S, Churchland AK: **Single-trial neural dynamics are dominated by richly varied movements.** *Nat Neurosci* 2019:22 <http://dx.doi.org/10.1038/s41593-019-0502-4>.
91. Golda T, Kalb T, Schumann A, Beyerer J: *Human Pose Estimation for Real-World Crowded Scenarios*. 2019arXiv:1907.06922.
92. Rhodin H, Spörri J, Katircioglu I, Constantin V, Meyer F, Müller E, Salzmann M, Fua P: **Learning monocular 3d human pose estimation from multi-view images.** *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 2018:8437-8446.
93. Seethapathi N, Wang S, Saluja R, Blohm G, Kording KP: *Movement Science Needs Different Pose Tracking Algorithms*. 2019arXiv:1907.10226.
94. Michaelis C, Mitzkus B, Geirhos R, Rusak E, Bringmann O, Ecker AS, Bethge M, Brendel Wieland: *Benchmarking Robustness in Object Detection: Autonomous Driving When Winter is Coming*. 2019arXiv:1907.07484.
95. Xian Y, Lampert CH, Schiele B, Akata Z: **Zero-shot learning — a comprehensive evaluation of the good, the bad and the ugly.** *IEEE Trans Pattern Anal Mach Intell* 2019, **41**(9):2251-2265 <http://dx.doi.org/10.1109/TPAMI.2018.2857768> ISSN 0162-8828.
96. Simon T, Joo H, Matthews I, Sheikh Y: **Hand keypoint detection in single images using multiview bootstrapping.** *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 2017.