There is no mouse: using a virtual mouse to generate training data for video-based pose estimation

Guido T. Meijer, Jaime Arlandis and Anne E. Urai

Over the last decade, deep artificial neural networks have revolutionized fields such as speech recognition, object detection, and drug discovery[1]. Such 'deep learning' algorithms learn which features of the data are relevant to perform a certain task. This makes hand-built 'feature extractors' unnecessary and obsolete since they are a more time consuming, inefficient, and less versatile way of dealing with this problem. Many recent, impactful applications of deep learning are based on *supervised* classification, where the only input needed by the model is a labeled dataset. For this reason, obtaining large labeled datasets has become the new bottleneck in training deep learning algorithms. As a consequence, the use of pre-trained models - also called transfer learning - has become widespread. In this approach, a model is trained using high quality datasets and can then be re-used with fewer labeled examples for other applications in a similar domain. For instance, the *Inception* algorithm, trained by Google to classify images from 1000 different categories[2], is now widely used to score conceptual similarity between different sets of images[3].

Supervised deep learning algorithms, as well as transfer learning, have found several applications in neuroscience. Pose estimation algorithms[4–7] use deep learning to track animals' body parts from video data. This obviates the need to apply visible markers on the animals' body, and considerably eases the burden of manually scoring different behaviors (common in e.g. ethology). Instead of labeling many thousands of images, a researcher can now label as few as ~200 video frames, and the pose estimation algorithm then tracks the movement of these same body parts in the remaining video. Such algorithms have recently been applied to video data from various species, and have become a popular approach for analyzing rich behavioral data[8].

Because this approach still relies on user input to acquire the labeled examples, it is sensitive to inter-individual variability in how researchers label the images, human errors, issues with occlusion, and difficult-to-label frames or anatomical structures (e.g. specific points on the animal's spine, which may be obscured by its fur). Also, while the goal of behavioral scoring is to characterize the movement patterns of a complex object in 3D space, the datasets used to train pose estimation models are generated by manually labeling projections of these objects into a 2D space of pixel values. Even in the case of multi-camera setups, which allow 3D reconstruction through a combination of the different points of view, the labeling process is still subject to these issues.

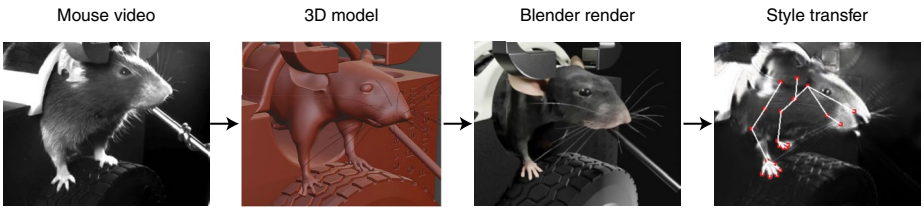

Fig. 1 | Virtual mouse workflow. (**a**) Video of a mouse using a steering wheel to report its decisions about a visual stimulus[11]. (**b**) A combined 3D model of the mouse body in the specific behavioral rig[12]. (**c**) Rendering of the 3D model in Blender[13], adding materials properties and camera angles. (**d**) Image-domain transformation is used to incorporate the video style (textures, lighting and background) into the animation. Ground-truth marker locations are overlaid in red, and can be exported to serve as labels for pose estimation algorithms. Images reprinted with permission from Bolaños et al. (2021). Springer Nature[9] and https://osf.io/h3ec5.

To overcome these difficulties in the context of mouse research, Bolaños *et al.*[9] leverage the *sim2real* approach common in robotics[10], using simulated environments to train machine learning algorithms in cases where acquiring real data (in this case, manually placed labels) is costly. They developed an anatomically realistic 3D model of the mouse body using computed tomography scans, which can be combined with a 3D model of a specific behavioral rig (Fig. 1). This model can then be animated to render synthetic video data. After animation, an image-domain transformation algorithm is used to modify the textures of the videos, making them resemble real videos from a specific behavioral task. From these synthetic videos, researchers can generate ground-truth labeled datasets. This can allow for efficient training of pose estimation models, improving tracking accuracy while significantly reducing human labelling efforts.

One advantage of this approach over hand-labeling a training set is that, after the initial effort to create the model, the training dataset can be easily augmented. For instance, researchers can add noise to the movement patterns, change the lighting, the camera position, or adjust other variables that may vary across videos in the real dataset. The virtual mouse approach can also help to reduce problems associated with user-generated labels: occlusion, inter-researcher variability, and difficult to label frames and body parts. Also, any rare postures which the animal infrequently adopts can be imitated in the virtual mouse, generating more synthetic labels to train the tracking algorithm on these uncommon movements. This opens the possibility of generating potentially infinite datasets, which in turn may allow the training of more robust and versatile behavioral tracking algorithms.

A significant limitation of the 'virtual mouse' technique is the difficulty of determining when the simulation is good enough. Finding and correcting frames in which the model fails to resemble real video data is difficult to automate, and the evaluation metrics[14] used in Bolaños *et al.*[9] may still present some flaws[15]. While rare animal postures can be simulated using the virtual mouse, the fact that these postures occur still has to be determined by a human observer. The user also has to determine when the model sufficiently covers the space of possible lightning and camera position variability.

Another major downside of the 'virtual mouse' approach is that building and animating a virtual scene using the mouse model takes a long time, and requires the experimenter to learn how the Blender software[13] works. The authors acknowledge that the initial time investment to set up such a virtual scene is large: 15-20 hours. Indeed, when testing the software to generate a virtual scene, our experience was that it took a significant amount of time to create the virtual environment and animate the mouse model. This was the case even with step-by-step tutorials and intuitive software controls. The authors argue that hand-labeling a training set for pose estimation also requires substantial user effort, making the 'virtual mouse' approach worthwhile. However, in our experience, hand-labeling 200 frames (using the DLC[4] pipeline) takes ~2 hours. Therefore, the 'virtual mouse' approach may be unsuitable for labs in which each researcher uses a different behavioral setup, each of which requires a custom virtual scene. The time it takes to create the virtual scene goes up with the complexity of the behavioral setup, as each component has to be recreated in Blender. Moreover, this extra time investment does not translate into a direct improvement of pose estimation performance on 2D videos; the performance gains are most prominent in 3D setups[9]. These issues may dissuade labs that do highly specialized, custom behavioral experiments from using this technique.

The true strength of the 'virtual mouse' technology comes to light in the context of large collaborations that use standardized experimental setups, such as the International Brain Laboratory[11] and the Allen Institute[16,17]. In such cases, a 3D model of the experimental setup is usually created during the development of standardized behavioral hardware, greatly reducing the additional time required to create a virtual scene. The virtual scene only has to be created once, and its output can be applied to all the labs that use the standardized behavioral apparatus in question. Even though care is taken to standardize each rig and video appearance, the variability between videos within a distributed collaboration will be larger than within a single lab: for instance, the exact camera angle and lighting conditions will vary slightly from rig to rig. Using the virtual scene, training data which incorporates this variability can be automatically generated, improving the robustness of video tracking. When the organization decides to modify the setup, for example by adding another camera, creating a new training set for pose estimation will come at almost no extra time investment. Furthermore, instead of designing behavioral setups from scratch, individual labs might increasingly adopt existing standardized rigs and adjust them to their specific needs. When a virtual scene already exists, these labs could modify the existing virtual scene to match their custom setup, and automatically generate a new pose estimation training set.

The 'virtual mouse' approach fits in a broad, growing ecosystem of open-source software, standardized experimental protocols and large-scale, curated datasets that can be used as benchmarks for neuroscience research. This trend of collecting more and richer data from each individual mouse fits well with principles of ethical laboratory animal research: video-based pose estimation can help reduce the number of animals studied by extracting richer information about simultaneous movement and task variables[8]. The new method proposed by Bolaños et al.[9] may increase the efficiency and reliability of marker labeling on videos from these large-scale open datasets. In this way, it can allow researchers to extract more information, and ultimately gain more scientific insight, from mouse data in neuroscience. ❐

Guido T. Meijer[1,3], Jaime Arlandis[1,3] and Anne E. Urai[2] ✉

[1]Champalimaud Center for the Unknown, Lisbon, Portugal. [2]Cognitive Psychology Unit, Institute of Psychology and Leiden Institute for Brain and Cognition, Leiden University, Leiden, The Netherlands. [3]These authors contributed equally to this work: Guido T. Meijer, Jaime Arlandis.
✉*e-mail: a.e.urai@fws.leidenuniv.nl*

Published online: 11 June 2021
https://doi.org/10.1038/s41684-021-00794-z

References

1. LeCun, Y., Bengio, Y. & Hinton, G. *Nature* **521**, 436–444 (2015).
2. Zaccone, G. & Karim, M. R. *Deep Learning with TensorFlow: Explore neural networks and build intelligent systems with Python, 2nd Edition*. (Packt Publishing Ltd, 2018).
3. Salimans, T. et al. *arXiv* **1606.03498** (2016).
4. Mathis, A. et al. *Nat. Neurosci.* **21**, 1281 (2018).
5. Pereira, T. D. et al. *Nat. Methods* **s16**, 117–125 (2019).
6. Graving, J. M. et al. *eLife* **8**, e47994 (2019).
7. Wu, A. et al. *bioRxiv* **259705** (2020). https://doi.org/10.1101/2020.08.20.259705
8. Musall, S., Urai, A. E., Sussillo, D. & Churchland, A. K. *Curr. Opin. Neurobiol.* **58**, 229–238 (2019).
9. Bolaños, L. A. et al. *Nat. Methods* **18**, 378–381 (2021).
10. Tobin, J. *et al.* in *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)* 23–30 (2017). https://doi.org/10.1109/IROS.2017.8202133
11. The International Brain Laboratory *et al. eLife* **63711**, (2021).
12. The International Brain Laboratory. 3D View Behavioral Training Rig. *Figshare* (2020). https://doi.org/10.6084/m9.figshare.13042574
13. Community, B. O. Blender - a 3D modelling and rendering package. (2018). http://www.blender.org
14. Bińkowski, M., Sutherland, D. J., Arbel, M. & Gretton, A. *arXiv,* **1801.01401** (2021).
15. Barratt, S. & Sharma, R. *arXiv,* **1801.01973** (2018).
16. de Vries, S. E. J. et al. *Nat. Neurosci.* **23**, 138–151 (2020).
17. Siegle, J. H. et al. *Nature* **592**, 86–92 (2021).

Acknowledgements

We thank Liam Paninski and Matthew Whiteway for comments on the manuscript. AEU is supported by the German National Academy of Sciences Leopoldina and the International Brain Research Organization.