eLife

Reviewed Preprint
v2 • March 24, 2025
Revised by authors

Reviewed Preprint
v1 • May 14, 2024

Neuroscience

Anti-drift pose tracker (ADPT): A transformer-based network for robust animal pose estimation cross-species

Guoling Tang, Yaning Han, Xing Sun, Ruonan Zhang, Minghu Han, Quanying Liu, Pengfei Wei

Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China • University of Chinese Academy of Sciences, Beijing, China • Guangxi University of Science and Technology, Liuzhou, China • Shenzhen University of Advanced Technology, Shenzhen, China • Department of Biomedical Engineering, Southern University of Science and Technology, Shenzhen, China

eLife Assessment

This **useful** study introduces a deep learning-based algorithm that tracks animal postures with reduced drift by incorporating transformers for more robust keypoint detection. The efficacy of this new algorithm for single-animal pose estimation was demonstrated through comparisons with two popular algorithms. The strength of evidence is **solid** but would benefit from consideration of issues in multi-animal tracking. This work will be of interest to those interested in animal behavior tracking.

https://doi.org/10.7554/eLife.95709.2.sa2

Abstract

Deep learning-based methods have advanced animal pose estimation, enhancing accuracy and efficiency in quantifying animal behavior. However, these methods frequently experience tracking drift, where noise-induced jumps in body point estimates compromise reliability. Here, we present the Anti-Drift Pose Tracker (ADPT), a transformer-based tool that mitigates tracking drift in behavioral analysis. Extensive experiments across cross-species datasets—including proprietary mouse and monkey recordings and public Drosophila and macaque datasets—demonstrate that ADPT significantly reduces drift and surpasses existing models like DeepLabCut and SLEAP in accuracy. Moreover, ADPT achieved 93.16% identification accuracy for 10 unmarked mice and 90.36% accuracy for freely interacting unmarked mice, which can be further refined to 99.72%, enhancing both anti-drift performance and pose estimation accuracy in social interactions. With its end-to-end design, ADPT is computationally efficient and suitable for real-time analysis, offering a robust solution for reproducible animal behavior studies. The ADPT code is available at https://github.com/tangguoling/ADPT.

Introduction

Animal behavior is a complex and dynamic phenomenon that is shaped by a wide range of factors, including environment, genetics, diseases, cognitive states, and social interactions Robinson et al. (2008). Understanding the underlying mechanisms and neural correlates of animal behaviors requires accurate and detailed pose tracking as they move freely Pereira et al. (2020); Krakauer et al. (2017). Recently, deep learning-based tools such as DeepLabCut, SLEAP, and DeepPoseKit have offered the feasibility of automatically quantifying complex freely-moving animal behaviors from videos recorded by contactless cameras Mathis et al. (2018); Pereira et al. (2022); Graving et al. (2019). Nevertheless, these deep learning methods are susceptible to uncertainty and noise interference, leading to tracking drift due to errors in the detection of one or more keypoints, in the estimated keypoint dynamics Weinreb et al. (2024); Hsu and Yttri (2021); Lonini et al. (2022). Such drift in keypoints estimates can broadly affect subsequent animal behavior statistics and downstream tasks, such as behavior classification, individual identification, and social behavior clustering Sheppard et al. (2022); Huang et al. (2021). It severely jeopardizes the reliability and repeatability of ethological studies. Thus, there is an urgent need for an anti-drift pose tracking tool for animal behavior analysis.

Tracking drift of pose estimation, occurring at the upstream behavioral analysis, generally hinders all downstream behavior-related studies. For example, animal gait analysis relies on accurate tracking of limbs and paws Sheppard et al. (2022), and behavioral classification relies on the dynamics of body keypoints Huang et al. (2021); Han et al. (2022). So far, deep learning pose estimation has not achieved the reliability of classical kinematic analysis, which often involves post-processing in real-world applications Niknejad et al. (2023); Aljovic et al. (2022); Weinreb et al. (2024). One major reason is tracking drift. The drifted keypoints may be unsystematically distributed within each predefined behavior class, misleading the decision boundaries of the behavior class, thereby reducing the performance of supervised behavior classification Gabriel et al. (2022) or unsupervised behavior representation Huang et al. (2021). Even the state-of-the-art (SOTA) deep learning methods such as DeepLabCut and SLEAP have no effective strategies to avoid the tracking drift Weinreb et al. (2024); Mathis et al. (2018); Pereira et al. (2022); Graving et al. (2019); Lauer et al. (2022). Inherited from the tracking drifts, the inaccuracy of pose estimation, gait analysis, and behavioral classification may result in wrong behavioral discoveries, such as those investigating behavioral correlates of genes, neural circuits, and neuropsychiatric diseases Sheppard et al. (2022); Huang et al. (2021); Liu et al. (2021); Han et al. (2022). These concerns, along with issues related to the safety of deep learning tools, have slowed the widespread application of deep learning-based methods in behavioral analysis and limited the development of ethology.

There are three strategies to eliminate tracking drift in current SOTA methods of animal pose estimation. The first strategy is human refinement or human in the loop Mathis et al. (2018); Pereira et al. (2022). DeepLabCut (DLC) and SLEAP both embed a user interface to allow humans to exclude and rectify outliers frame by frame Mathis et al. (2018); Pereira et al. (2022). Although it would be the golden criterion to reduce the tracking drift, this strategy restricts the efficiency of the biological experiment when the human faces millions of drifted frames. The second strategy is signal processing filters such as median filter and low pass filter Stenum et al. (2021); Pereira et al. (2019); Luxem et al. (2022); Weinreb et al. (2024); Han et al. (2024); Li and Lee (2021). They can efficiently remove most of the drifted points without human intervention, but they will also remove the subtle behaviors with high-frequency features such as self-grooming in autism mouse models Huang et al. (2021) or tremor in animal models of Parkinson's disease Baker et al. (2022). The third strategy is fitting the drifted frames using linear dynamic models such as Keypoint-Moseq Weinreb et al. (2024) and adaptive Kalman filter Huang et al. (2022). They can reduce the drift and maintain the high-frequency behaviors at the

same time. Nevertheless, the performance of these models would drop sharply when processing continuous and long-duration drifted frames. These three strategies are only expedient to reduce tracking drift after pose estimation, whose performances are also restricted by the tracking accuracy of raw frames. Therefore, the elimination of tracking drift should be tackled from the beginning of the deep learning pose estimation step.

The structure design of the artificial neural network (ANN) is the first step to correct tracking drift. DeepLabCut, SLEAP, and DeepPoseKit all take the convolutional neural network (CNN) as the main component of pose estimation ANNs, which is the core problem causing tracking drift Mathis et al. (2018); Pereira et al. (2022); Graving et al. (2019); Lauer et al. (2022). The limited working memory of the CNN makes it easy to be influenced by the content-independent parameters to predict the wrong locations of keypoints and finally cause tracking drift Yang et al. (2021). To avoid this drawback, the Transformer becomes a better option to construct pose estimation ANNs because it is more efficient to capture global dependent features of images Yang et al. (2021); Stoffl et al. (2021); Xu et al. (2022b). Although Transformer-based ANNs have achieved new SOTA in lots of human pose estimation datasets, it is rarely applied in animal pose estimation. Different from human poses, animal poses have more indistinct body structures because they are covered by furs Vidal et al. (2021). In addition, the well-annotated animal pose datasets are not abundant enough to cover various experiment settings. Experimenters always need to make customized datasets for their specific applications Han et al. (2024). Therefore, the application of the Transformer to reduce tracking drift in the animal pose estimation task still needs an elaborate design of ANN structures.

To import the Transformer to overcome the tracking drift of animals, we designed an anti-drift pose tracker (ADPT) following the characteristics of animal behavior data. CNN and Transformer are cascaded with skip connections to capture subtle animal appearance features from only hundreds of labeled frames He et al. (2016); LeCun et al. (2015); Vaswani et al. (2017). This structure design makes ADPT show significantly fewer tracking drifts than DeepLabCut and SLEAP Mathis et al. (2018); Pereira et al. (2022). The effect of anti-drift of ADPT is universally validated in the public datasets and our customized datasets including Drosophilas, mice, and macaques, which demonstrates that ADPT is robust in broad application scenarios cross-species Pereira et al. (2019); Bala et al. (2020); Han et al. (2024). ADPT also achieves robust pose estimation and identity recognition of free-interactive mice combined with a mix-up dataset generation strategy. The results of markerless identity recognition show that the feature extraction of ADPT is reliable enough to cover both multi-animal pose estimation and identity recognition tasks, which are more difficult than the single-animal pose estimation task Agezo and Berman (2022); Lauer et al. (2022); Han et al. (2024). It reduces the computational time cost and increases the throughput of behavioral data processing because ADPT does not need a multi-stage neural network such as SIPEC Marks et al. (2022) or Social Behavior Atlas Han et al. (2024). Together, ADPT would be an accessible tool to reduce the pose tracking shift across species from the upstream of behavior analysis. ADPT has the potential to improve the reliability of computational ethology-based biological studies.

Results

Anti-drift pose tracker

Existing deep learning-based methods often produce some unreliable pose estimation, such as interference caused by similar objects, keypoint drift, and failures of body part detection (**Figure 1A**). To clarify these errors, we use “track” or “tracking” to refer to the tracking of all body points or poses of an individual, and “detect” or “detection” when referring to specific keypoints. These estimation errors largely compromise the robustness of pose estimation in freely behaving animals, which can affect the statistical results of behavioral analyses and sometimes even lead to

erroneous scientific findings. In this study, we present a reliable animal behavioral analysis tool, called the anti-drift pose tracker (ADPT). ADPT can effectively eliminate estimated drifts (**Figure 1B**). ADPT is a heatmap-based pose estimation network that inferences input images to confidence heatmap, location refinement, and low-resolution semantic segmentation (LRSS) (**Figure 1C**). In the network architecture of ADPT (**Figure 1D**), we utilize the convolutional structure to extract local information on the one hand, and the transformer attention mechanism to learn the long-term global dependencies on the other hand. Compared with purely attention-based network structures (such as ViT Yang et al. (2021); Stoffl et al. (2021); Xu et al. (2022b), our CNN-transformer structure can significantly reduce the number of model parameters and therefore requires fewer training data samples. It is particularly suitable for data-limited applications such as animal behavior analysis.

Customized behavioral videos for testing ADPT

The identification of drifting keypoints relies heavily on videos generated during inference or visualized coordinates. Yet there is no publicly available video dataset specifically designed for anti-drift evaluation. To fill this gap, we collected behavioral data from mice and monkeys (see supplement videos 1). For single animal pose estimation, we recorded videos from free-moving mice and monkeys with 4 cameras and then hand-labeled randomly extracted frames. For mice, we labeled these frames with 16 keypoints, including nose, eyes, ears, front limbs, front claws, back, hind limbs, hind claws, root tail, mid tail and tip tail. For monkeys, we labeled these frames with 17 keypoints, including nose, eyes, ears, shoulders, elbows, hands, hips, knees, and ankle. Given the popularity of mouse behavioral study, mice served as our primary subjects for evaluation, with videos obtained from 4 different perspectives involving 4 distinct individuals. Each mouse video spanned 15 minutes. The training dataset comprised 440 randomly extracted images from these videos and other collected videos(training:validation=95%:5%). Monkey videos, on the other hand, encompassed 8 different viewpoints, featuring multiple individuals, from which a 30 minutes video was used for performance evaluation. The training set consisted of 3488 randomly sampled images (training:validation=95%:5%). For social or multi-animal pose estimation, we recorded 10-minute videos of freely socializing mice in a homecage from three different perspectives. We manually labeled 1200 images for training and validation (training = 95%:5%) and also labeled the back locations of two mice during the first minute to evaluate tracking accuracy. Using our dataset, we trained ADPT, DeepLabCut, and SLEAP models, separately, to detect body keypoints from behavioral videos. The behavioral data is available at *https://github.com/tangguoling/ADPT/data*.

ADPT demonstrates the remarkable anti-drift performance

Firstly, we visualized the time course of seventeen estimated key body parts from a one-minute segment of mouse videos (**Figure 2A**), demonstrating the anti-drift effects of ADPT. In contrast, the other two deep learning-based methods suffer from drift and misses of body parts. Then, we zoomed into the frames of failures in **Figure 2B**. The quantitative results of 240-minute videos from two mice were shown in **Figure 2C&D**. Interestingly, DeepLabCut has almost the same probability of generating drift and misses, while SLEAP was more prone to misses. As presented in **Figure 2D**, the tip tail was the most challenging part of the body for both drifts and missed due to the long distance from the tip tail to the rest of the body. For CNN-based methods such as DeepLabCut and SLEAP, learning such long-range tail-body relationships is particularly difficult, while the attention mechanism of ADPT allows it to learn long-range dependencies. Due to frequent occlusion in the video, the left and right claws could be easily missed or drifted. Our model evaluations show that ADPT has significantly lower root mean squared errors compared to SLEAP and achieves comparable or improved accuracy compared to DeepLabCut (**Figure 2E**), suggesting that ADPT can reliably detect the hind claws, offering a potential tool for gait analysis and tail-related behavior paradigms.

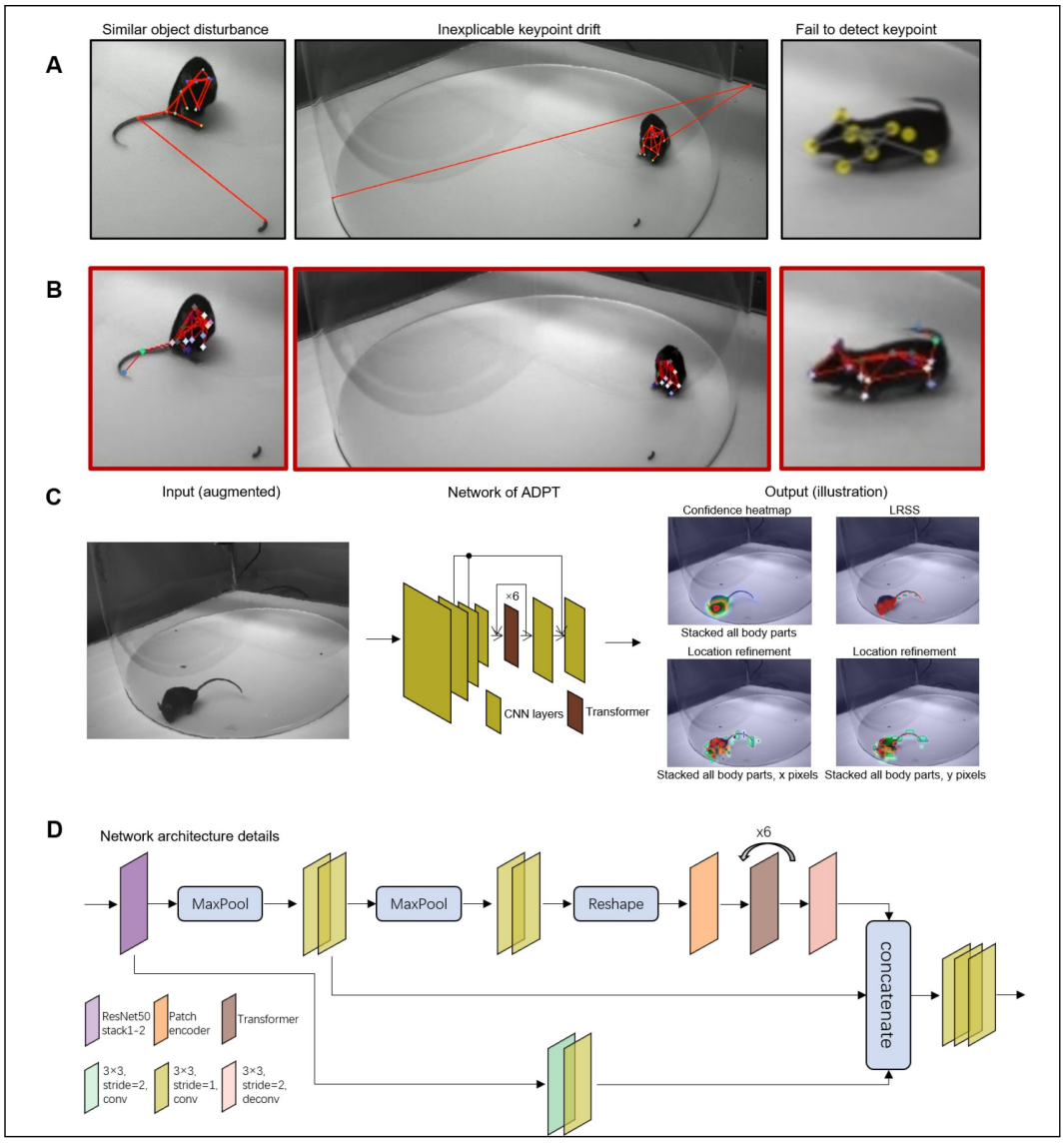

Figure 1.

Anti-drift pose tracker (ADPT). **A** Three examples of drifts in deep learning-based animal behavioral analysis. Similar object disturbance means that the object similar to a specific body part misleads the deep learning-based methods. Inexplicable keypoint drift is caused by the high confidence score predicted on the wrong place by the network. Failure to detect the keypoint is probably caused by the the predicted low confidence score. **B** The anti-drift effects of ADPT. **C** The general workflow of ADPT. The network is trained to predict confidence heatmap, LRSS, and location refinement. **D** The network architecture of ADPT.

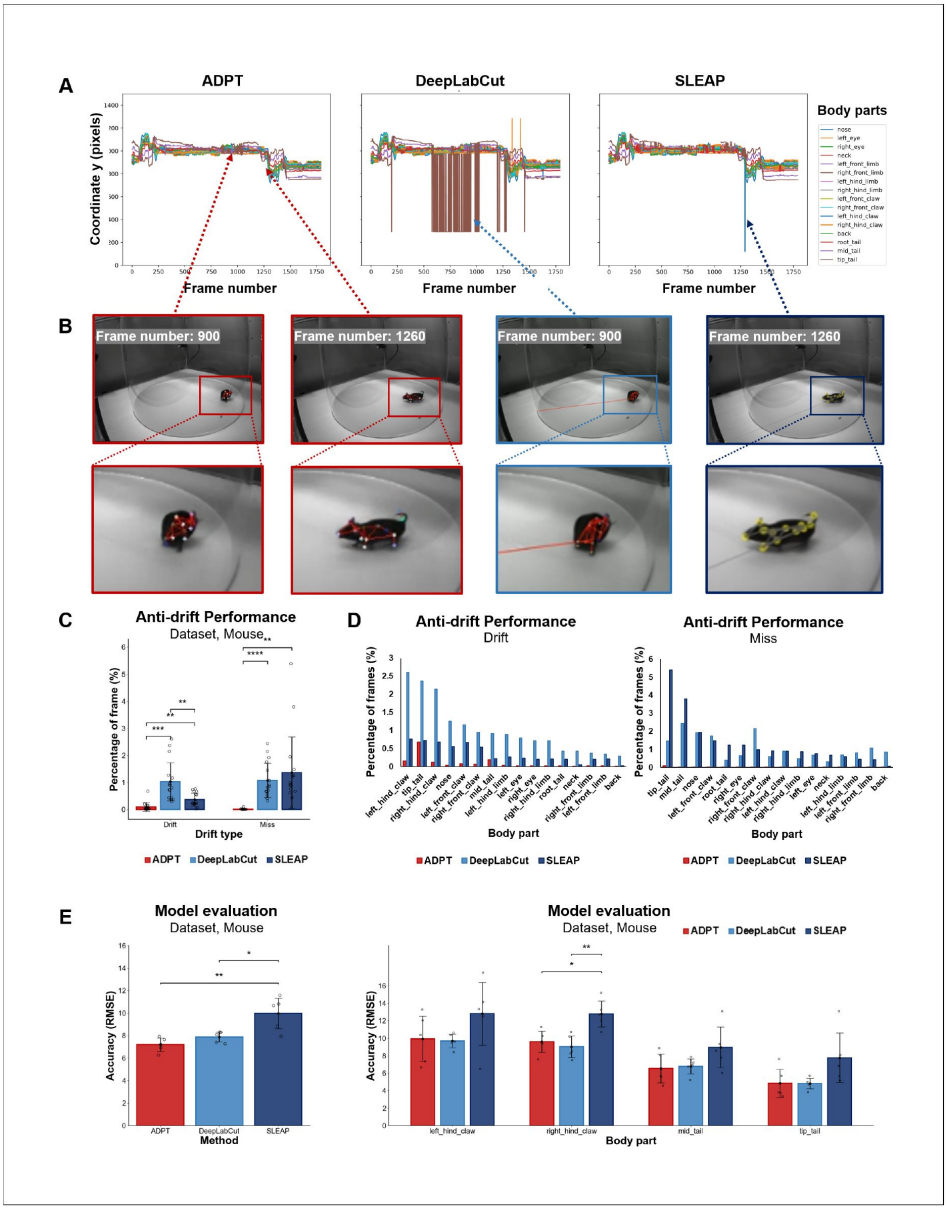

Figure 2.

Analysis of ADPT's anti-drift performance in a mouse dataset collected by our lab. **A** The time course of the *y*-axis position of sixteen body parts extracted from a one-minute video using ADPT, DeepLabCut and SLEAP tools. It showed that ADPT successfully detected all 16 body parts of a mouse, whereas DeepLabCut and SLEAP encountered inexplicable tracking drifts. **B** Two anti-drift examples from ADPT, where the tail was drifted by DeepLabCut and the hind claw failed to detect by SLEAP. **C** Overall percentage of tracking drift and failing to detect (miss) frames from three methods. ADPT demonstrated a significantly lower drift percentage than other methods. **D** The percentage of frames with tracking drift (left) and failing to detect (right). Drifts were mainly from the top four body parts, including the tip tail, the left and the right hind claws, and the middle tail. **E** The averaged RMSE across all body parts (left) and RMSE of the top four body parts with drifts (right). ADPT achieved the smallest RMSE than other two tools when thresholded at 0.2. *: P<0.05, **: P<0.01, ***: P<0.001, ****: P<0.0001. RMSE: root mean square error.

Anti-drift performance remains consistent irrespective of the video background and individual animals

Any measuring tool that exhibits biased measurement errors towards specific subjects introduces inaccuracies in its assessments. For example, if the model accurately estimates the posture of mouse A but experiences greater posture drift in estimating mouse B, this discrepancy leads to measurement errors, impacting subsequent behavioral analyses. Hence, to evaluate the independence of posture estimation's anti-drift effect concerning individual animals or background factors, we conducted one-way ANOVA on the tracking results. We trained ADPT, DeepLabCut, and SLEAP five times each and applied these models to infer behavioral videos across different individuals and video backgrounds. Firstly, we compared ADPT's anti-drift performance across different individuals and backgrounds. The results showed that ADPT exhibited significantly lower drift percentages than the other two methods across different individuals and video backgrounds (**Figure 3 A&C**).Then, the inference results were grouped based on individual animals and video backgrounds, respectively, for five individual one-way ANOVA analyses. The results of these five ANOVA analyses are presented in **Figure 3 B&D**. Our analyses revealed that drift occurrences were more significantly affected by backgrounds in DeepLabCut, while individual variations had a more significant impact on SLEAP. However, ADPT showed slight susceptibility to background influence. Consequently, we assert that in comparison to DeepLabCut and SLEAP, ADPT only demonstrates a lesser susceptibility to the influence of individual animals and background factors. This resilience significantly mitigates biases in tracking results. ADPT's ability to generate fewer biases due to individual or background factors during inference holds promise for achieving better consistency in downstream behavioral analyses.This analysis also underscores the importance, when using ADPT, of minimizing background variations, ideally maintaining consistent backgrounds.

Cross-species anti-drift capability of ADPT is reliable

While ADPT has demonstrated exceptional anti-drift abilities in mice, numerous other animal models are employed in behavioral studies. To validate the robustness of ADPT in tracking different species, particularly those posing significant tracking challenges, we selected cynomolgus monkey as a specie known for its complexities in tracking. We utilzed the models to track a video in which both humans and monkey appeared simultaneously, presenting similar objects in the scene. Visualizing the keypoint tracking results from 1-minute time course featuring both entities allowed us to showcase the anti-drift efficacy of ADPT (**Figure 4A**). In contrast, the other two methods exhibited tracking failures when humans were present, as illustrated in the zoomed-in frames of failure in **Figure 4B**. When humans were present, both DeepLabCut and SLEAP exhibited instances of tracking drift, whereas ADPT remained unaffected by the presence of similar objects. Similarly, we evaluated the performance of 'drift' and 'miss' for various body parts in this scenario. We observed that ADPT consistently outperformed the other two methods overall (**Figure 4C&D**). However, given the more complex experimental setup and animal movements, ADPT exhibited slight instances of drift and 'fail to detect' effects.

Consequently, our findings suggest that our approach demonstrates remarkable anti-drift performance, cross-individual and cross-view capabilities. Notably, our anti-drift performance was more pronounced in consistent experimental scenarios. Our experiments with monkeys substantiated our method's profound cross-species anti-drift capability, emphasizing its significance in behavioral studies involving diverse animal models.

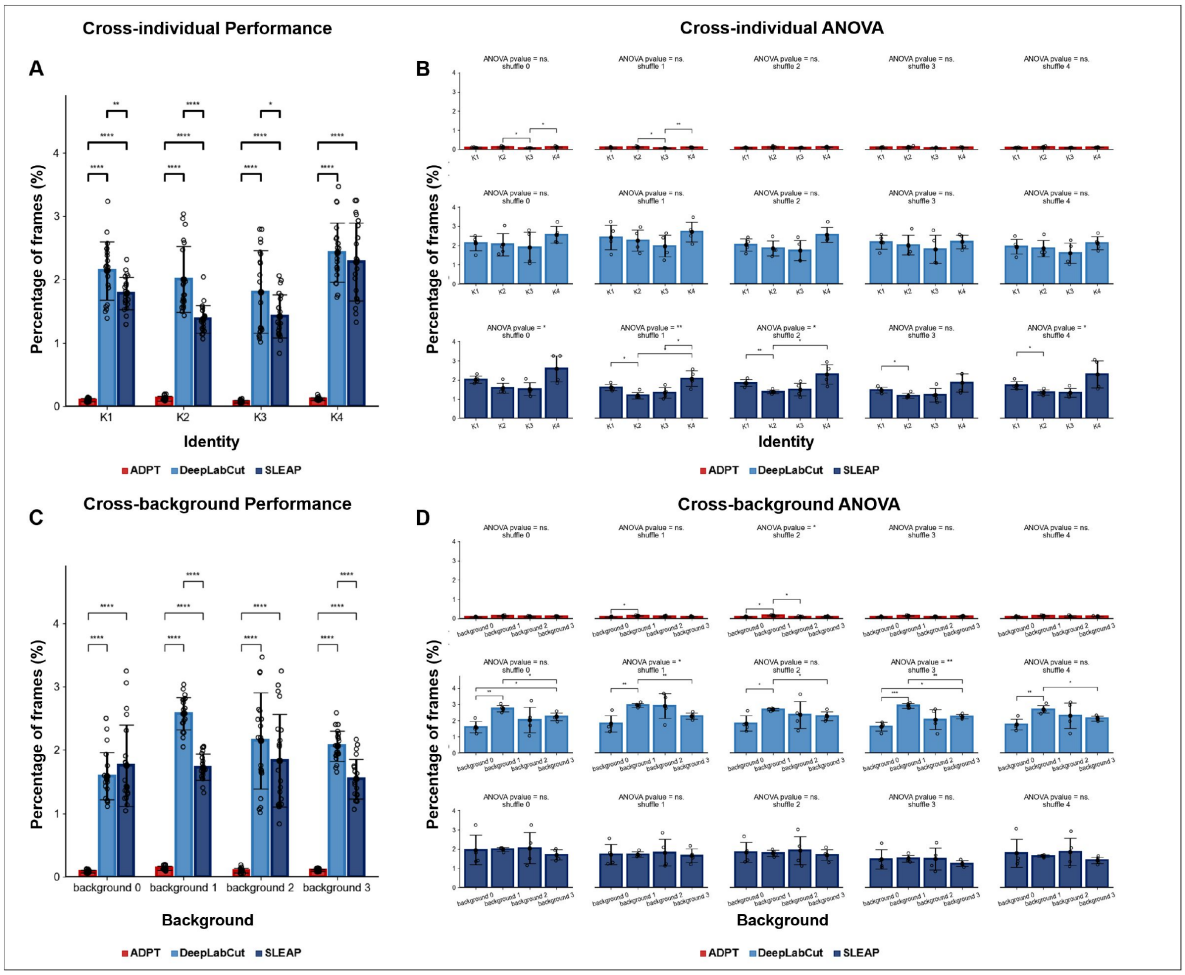

Figure 3.

Anti-drift performance cross background and individual, where the percentage of frames includes two types of drift phenomena: drift and miss. **A** The overall cross-individual anti-drift performance of ADPT and the other methods. The drift percentage of ADPT is significant lower than other methods. **B** After training the model 5 times on the dataset shuffle, the cross-individual drift percentage for each shuffle was analyse using one-way ANOVA. The ANOVA results revealed that there are differences in the inference results of the SLEAP model among individual, and there were no differences for ADPT or DeepLabCut. **C** The overall cross-background anti-drift performance of ADPT and the other methods. The drift percentage of ADPT is significant lower than other methods. **D** The cross-background drift percentage for each shuffle was analyse using one-way ANOVA. The ANOVA results revealed that there are slight differences in the inference results of the DeepLabCut model among individual, and there were no differences for ADPT or SLEAP. ns.: no significant, *: P<0.05, **: P<0.01, ***: P<0.001, ****: P<0.0001.

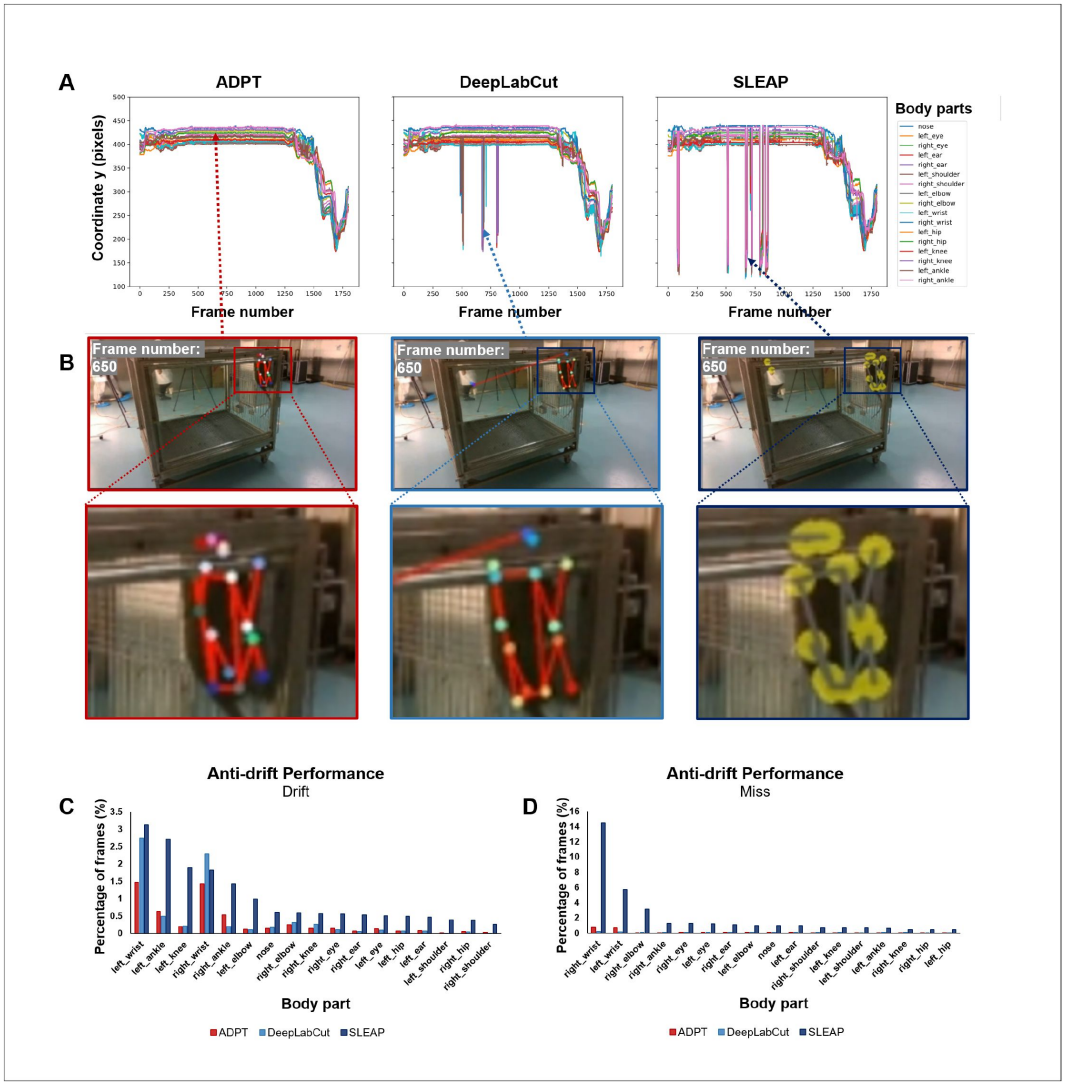

Figure 4.

Analysis of ADPT's anti-drift performance on monkey data, showing the cross species anti-drift ability. **A** The time course of the *y*-axis position of sixteen body parts extracted from a one-minute video using ADPT, DeepLabCut and SLEAP tools. It showed that ADPT successfully detected all 17 body parts of a monkey, while the other two methods encountered tracking drift because of the appearance of humans. **B** DeepLabCut and SLEAP both mistakenly located the monkey's eyes on humans when they appeared, while ADPT can achieve robust tracking. **C, D** The percentage of frames with tracking drift and failing to detect (miss). The occurrence of drift was mainly concentrated in the limbs, because the appearance of humans.

Public datasets confirm the outperformance of ADPT in precision and practicality

In adddition to evaluating ADPT's performance on behavioral study videos, we recognized the significance of image datasets as benchmarks for assessing pose estimation effectiveness. Thus, to comprehensively evaluate the generalizability of ADTP performance to animals in skeletal complexity and body size, and the background complexity of videos, we used two public datasets, a single fly dataset (**Figure 5A**, Pereira et al. (2019), and a macaque OMS_Datase Bala et al. (2020). The single fly dataset contains 1500 annotated frames of 32-node skeleton fly. To ensure a fair comparison, we followed the same dataset split strategy and data augmentation strategy described in Pereira et al. (2022). The evaluation metric used was mean Average Precision (mAP), which measures the accuracy of keypoint localization for all body parts, following the protocol established in Pereira et al. (2022). On the other hand, The OMS_Dataset Bala et al. (2020) is a large database of annotated macaque images (**Figure 5F**). To evaluate the performance of our methods, we randomly selected 5000 images out of 195,228 images from this dataset and resized them to 368 × 368 resolution. We split the dataset into 40% training data and 60% validation data. We employed the same strategy used in the default configuration of DeepLabCut toolbox to augment the data. The average distance (root square mean errors, RMSE) between the ground truth and predicted keypoints and the mAP were used as evaluation metrics. **Figure 5A** and **Figure 5F** presented several examples annotated by ADPT on these two datasets, respectively. Furthermore, to verify the practicality of ADPT, we also evaluated the amount of required training data and the inference speed of the model. Finally, we evaluated the scalability of ADPT on the StanfordExtra dataset Biggs et al. (2020). Our results demonstrated the capability of ADPT on non-laboratory dogs (Supplementary Fiigure 1 and Supplementary Video 2). These evaluations underscore ADPT's versatility, showcasing its robustness and accuracy in diverse animal contexts, thereby affirming its potential as a highly adaptable tool for comprehensive hehavioral studies.

ADPT offers higher tracking accuracy than existing SOTA methods

The tracking performance of ADPT was compared with the existing SOTA methods, such as DeepLabCut and SLEAP (**Figure 5B, G** and **H**). On the single-fly dataset, ADPT exceled with an average mAP of 92.83%, surpassing both DeepLabCut and SLEAP (**Figure 5B**). On the OMS Dataset, ADPT exhibited significant advantages in terms of mAP, RMSE (threshold = 0.2), RMSE (threshold = 0.6), achieving values of 30.9%, 8.32, and 6.25, which were significantly superior to SLEAP, and slightly outperforming DeepLabCut when the threshold set as 0.6 (**Figure 5G**, supplement Table 1). Moreover, we further examined the detection of macaque hip and tail on OMS_Dataset (**Figure 5H**). We found that ADPT's tracking performance of tail is better than DeepLabCut and SLEAP, while the hip tracking is equivalent to DeepLabCut and better than SLEAP. This further demonstrates the superiority of ADPT in tail-related behavior paradigms. By conducting evaluations on these diverse datasets, we aimed to assess the robustness and generalizability of our methods across more different animal species, pose complexities, and environmental conditions. The results obtained from these evaluations provide solid proof of the performance and potential of our methods for single-animal pose estimation.

ADPT only needs a small amount of annotated data

Since annotating behavioral data is tedious, a deep learning-based method that does not require large amounts of annotated data is crucial. Here, we studied how the accuracy of ADPT changes with the amount of annotated data. Notably, ADPT achieved acceptable performance using only 350 annotated images (**Figure 5D**), indicating that ADPT is data efficient.

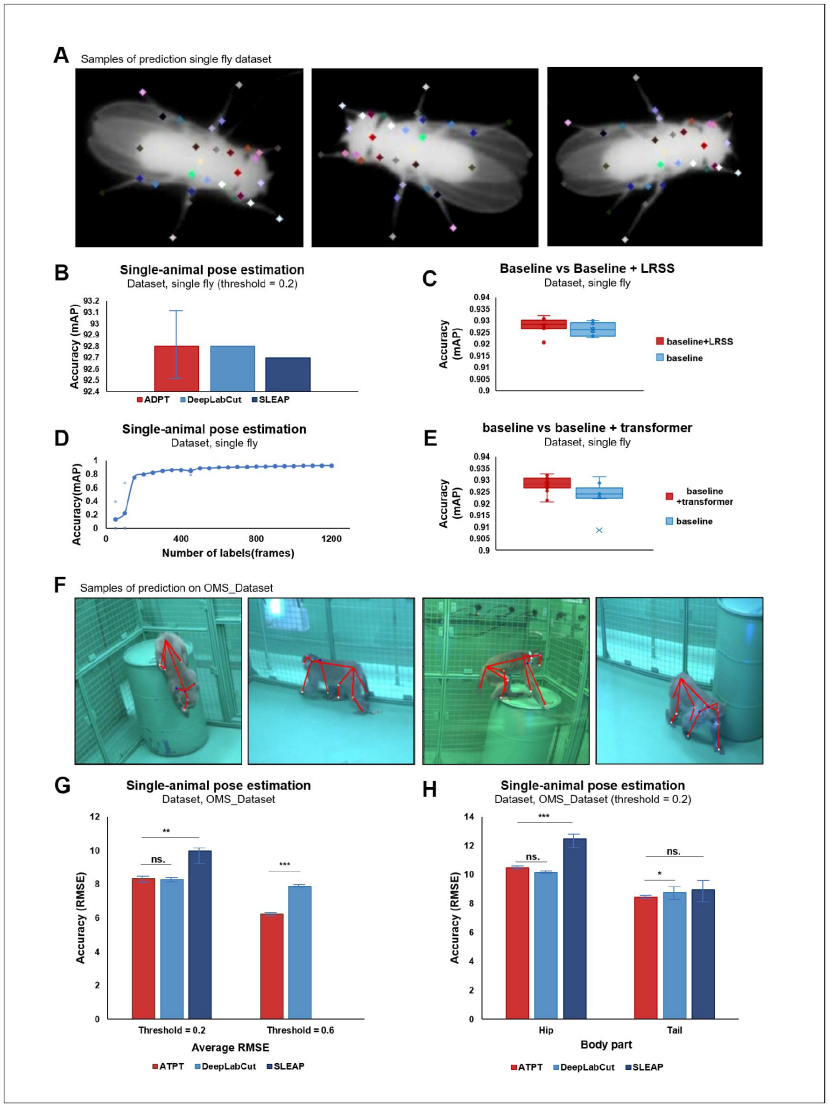

Figure 5.

Results of public datasets evaluation. **A** Samples of prediction on single fly dataset. **B** Mean average precision (mAP) on fly dataset, where ADPT achieved average 92.8% accuracy (the best model achieved 93.27%). **C** LRSS improved the average accuracy by 0.3% on single fly dataset. **D** Relationship between annotated image and accuracy of ADPT on fly dataset where ADPT achieved acceptable performance with only 350 annotated images in a simple laboratory environment. Points indicate the validation accuracy of model training on specific number of labels dataset. **E** Transformer improved the average accuracy by 0.4% on single fly dataset. **F** Samples of prediction on OMS_Dataset. **G** Root mean square error (RMSE) on OMS_Dataset, where ADPT achieved smaller RMSE than SLEAP when threshold = 0.2, and smaller than DeepLabCut when threshold = 0.6. P value, **: 0.001862, ns.: 0.243472, ***: 8.700e-06. **H** RMSE comparison on hip and tail of OMS_Dataset.P value, ***: 0.000561, Hip ns. :0.023766, Tail ns. :0.336642, *: 0.035782.

ADPT's fast inference enables real-time applications

Here we evaluate the inference speed of ADPT. We compared it with DeepLabCut and SLEAP on mouse videos at 1288 × 964 resolution. Our method exhibited an impressive prediction speed of 90±4 frames per second (fps), faster than DeepLabCut (44±2 fps) and equivalent to SLEAP (106±4 fps). These results highlighted the efficient inference capabilities of ADPT, which is crucial for real-time applications and the analysis of large-scale behavioral data.

LRSS and transformer help improve tracking accuracy

To examine the contribution of the low-resolution semantic segmentation (LRSS) and the transformer architecture to ADPT, we conducted two ablation studies using the fly dataset. We compared multiple variants to uncover the impacts of the LRSS module and the transformer module on pose estimation performance. Firstly, we explored the influence of LRSS by comparing the performance of the complete ADPT with the one removed LRSS. As shown in **Figure 5C**, LRSS module can improve the average accuracy by 0.2%. Moreover, to assess the role of transformer architecture, we conducted a comparative analysis between the complete ADPT with the transformer and a variant of the model where the transformer was removed. As shown in **Figure 5E**, the transformer improved the average accuracy by 0.4%, suggesting the benefits of the transformer architecture in pose estimation.

ADPT can accurately track the non-laboratory dog

To test the generalizability of our approach beyond laboratory-behavior animals, we applied ADPT to the keypoint detection task for the non-laboratory dog. The dataset is from Biggs et al. (2020). We randomly divided the dataset into 85% and 15% training and validation data. ADPT was instantiated with the same network architecture for laboratory animal pose estimation, showcasing the versatility of ADPT. We followed Biggs et al. (2020) and used Percentage of Correct Keypoints (PCK) metric to evaluate the accuracy of keypoint detection. The results showed that ADPT achieved an average 86.54% PCK score (legs: 85.54%, tail: 79.89%, ears: 88.61%, face: 95%). Examples of identified keypoints of dogs were shown in Supplement **Figure 1** and Supplement Video 2. These results supported the flexibility of ADPT in different animal species and potentially more challenging real-world scenarios.

ADPT can be adapted for end-to-end pose estimation and identification of freely social animals

We adapted ADPT to end-to-end tracking of the social interacting mice with similar appearances. To this end, we added additional heads after feature concatenation and utilized LRSS to confirm the identities of the mice. We generated a multi-animal dataset for social tracking by mixing up two labeled frames from single mouse videos (**Figure 6**). The evaluation of our social tracking capability was performed by visualizing the predicted video data (see supplement Videos 3 and 4).

Prior to social tracking, we evaluated identity-tracking accuracy using a dataset consisting of 10 mouse videos of different individuals. The overall workflow of these extended applications is depicted in **Figure 7**. Initially, we utilized a variant of ADPT (empowering LRSS with identity information) for simultaneous animal pose estimation and identity-synchronized tracking. For each frame, identity recognition was based on the LRSS output by ADPT (**Figure 7A**). Although the appearance of the mice is very similar, our experimental results showcased a remarkable 93.16% accuracy in identity recognition (**Figure 7B**). This approach demonstrates LRSS's capability to record individual identities like semantic segmentation masks. The outcomes, showcased in supplement Videos 3, manifested synchronized tracking of identity and pose estimation.

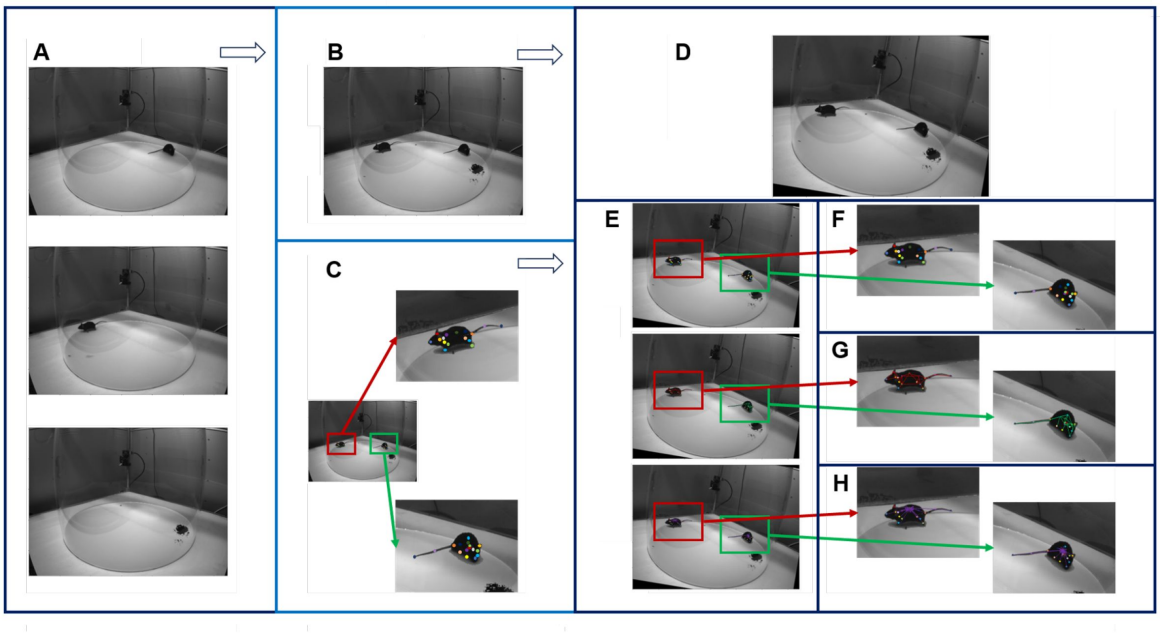

Figure 6.

Illustration for mix-up social animal dataset generation. **A** Frames originating from different videos and corresbonding background. **B** Mix-up image. **C** Represents schematic diagrams illustrating the keypoint generated from single animal pose estimation of ADPT. **D** Represents an augmented mix-up image. **E** Represents schematic diagrams of augmented annotation. **F** Represents augmented keypoints. **G** Represents augmented LRSS. **H** Represents schematic diagrams of augmented Body Affinity Fields(BAF), inspired by Part Affinity Fileds(Cao et al. (2021)).

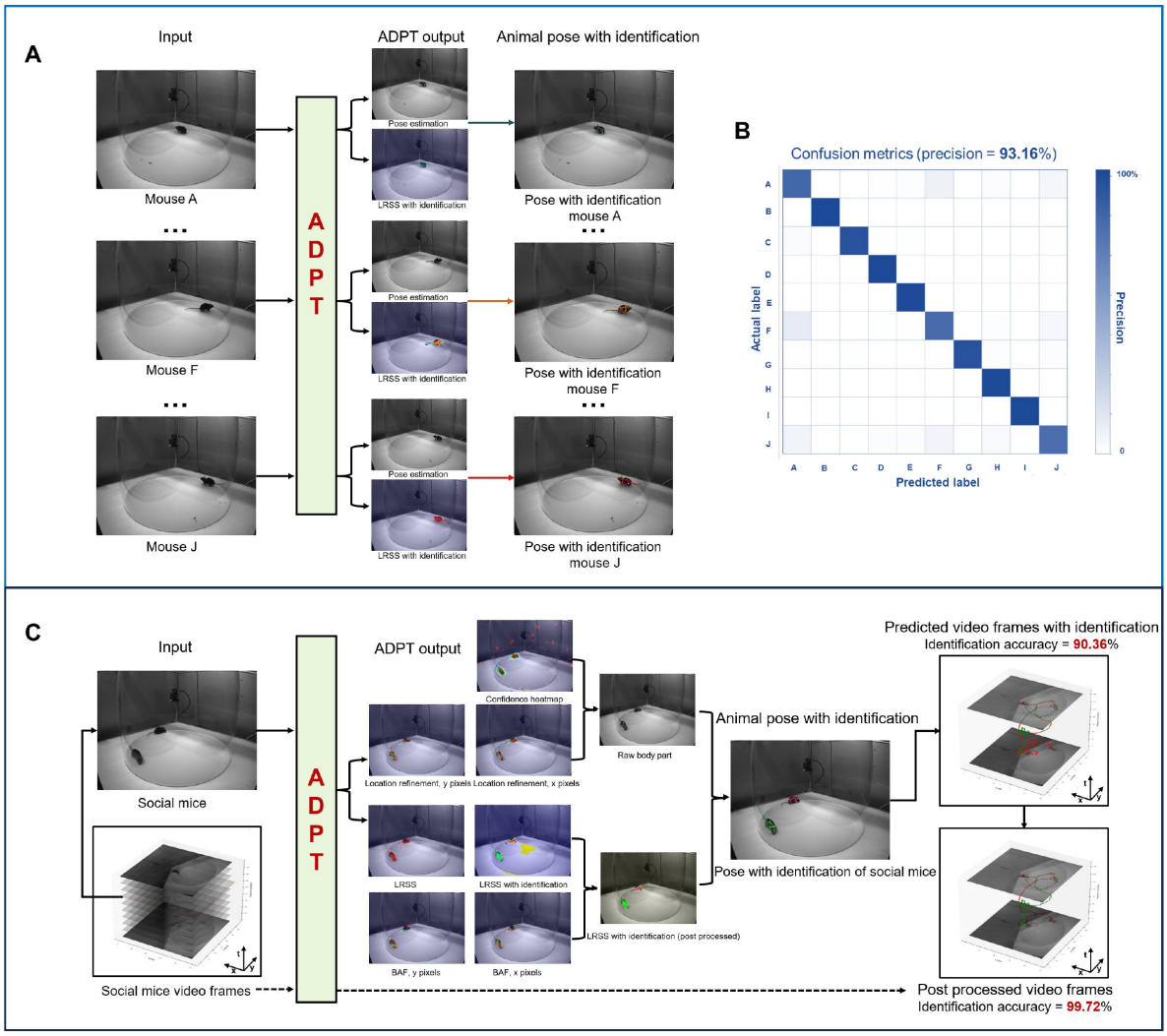

Figure 7.

Applications of ADPT for multi-animal pose tracking.

A Left: The pipeline for the multi-animal identity-pose tracking task. **B** Confusion matrix of the 10-mice classification (accuracy=93.16%).**C** Social mice tracking pipeline with identification accuracy of 99.72%.

Subsequently, we tested the tracking performance with free-social animals. Inspired by Part Affinity Fileds Cao et al. (2021), we created Body Affinity Fileds (BAF) to help distinguish different individuals. BAF and LRSS were used together to identify individuals. In the first scenario, We trained ADPT on the Mix-up social animal dataset and employed it to predict 1-minute free-social video of mice with similar appearance. Without additional temporal post-processing, ADPT achieved a 90.36% accuracy in identity recognition, as referenced in supplement Video 4A. Following temporal identity correction, ADPT remarkably achieved a 99.72% accuracy in identity recognition (**Figure 7C**), as shown in supplement Video 4B. The pose estimation accuracy was acceptable, but we recognized that there are detection error of tail or tracking error when animals are very closed sometimes which may be due to the lack of real-world training data.

In the second scenario, we trained ADPT on manual labeled homecage social mice dataset, a set of real-world training data (**Figure 8A**.) and used it to predict a 1-minute free-social video. We evaluated anti-drift performance and found that ADPT outperformed Deeplabcut and SLEAP, achieving 15% improvement of pose estimation accuracy (**Figure 8D, E**) and almost 5 to 10 times improvement of tracking accuracy (**Figure 8C**). **Figure 8B** illustrates ADPT's prediction of the x-coordinates of different mice, demonstrating less keypoint drift. In **Figure 8C**, we compare the anti-drift performance of the raw predictions from the three methods, highlighting ADPT's superior tracking performance compared to DeepLabCut and SLEAP. Furthermore, we assessed pose estimation accuracy between ADPT and DeepLabCut / SLEAP, showing that ADPT has better pose estimation accuracy than both SLEAP and DLC (**Figure 8D, E**). Lastly, an ablation study confirmed that BAF improves pose estimation accuracy (**Figure 8F**). The tracking result of this scenario was shown in supplement Video 5.

In addition to mice, we evaluated the pose estimation accuracy of ADPT on the marmoset dataset, a publicly available resource Mathis et al. (2018). We adhered to the default marmoset configuration of DeepLabCut, randomly dividing the dataset into training and validation sets while employing the same data augmentation strategy. Under the evaluation metrics used by DeepLabCut, ADPT achieved an average accuracy of 6.14 ± 0.19 pixels, whereas DeepLabCut reached 6.63 ± 0.09 pixels. Additionally, when assessed using SLEAP evaluation metrics, ADPT achieved an average accuracy of 7.02 ± 0.11 pixels, compared to SLEAP's 11.39 ± 0.31 pixels.

Together, these different applications demonstrate the versatility of ADPT, ranging from single animal pose estimation to complex situations involving social interactions. ADPT's versatility and adaptability paves the way for comprehensive behavioral studies.

Discussion

Here, we have presented ADPT, a Transformer-based pose tracker, to address the pose drift problem in animal pose estimation. The core of ADPT is the elaborate combination of the convolutional network LeCun et al. (2015) and transformer layers Vaswani et al. (2017), with the goal of capturing both local details and global context. This architecture helps ADPT achieve a more reliable feature extraction on animal objects, resulting in higher accuracy in tracking the poses frame by frame with less drifts or misses, compared to DeepLabCut and SLEAP Lauer et al. (2022); Pereira et al. (2022). In addition, we presented the procedure for the data generation of Mix-up social animals, which is convenient and effective for exponentially synthesizing new data to improve the performance of ADPT. We showed that ADPT can be used for multi-animal pose estimation and identification. These two tasks were considered much more difficult than single-animal pose estimation Lauer et al. (2022). The end-to-end network structure of ADPT only needs to calculate one model loss so it is more computationally efficient than the multi-stage methods such as SIPEC and Social Behavior Atlas Marks et al. (2022); Han et al. (2024). These

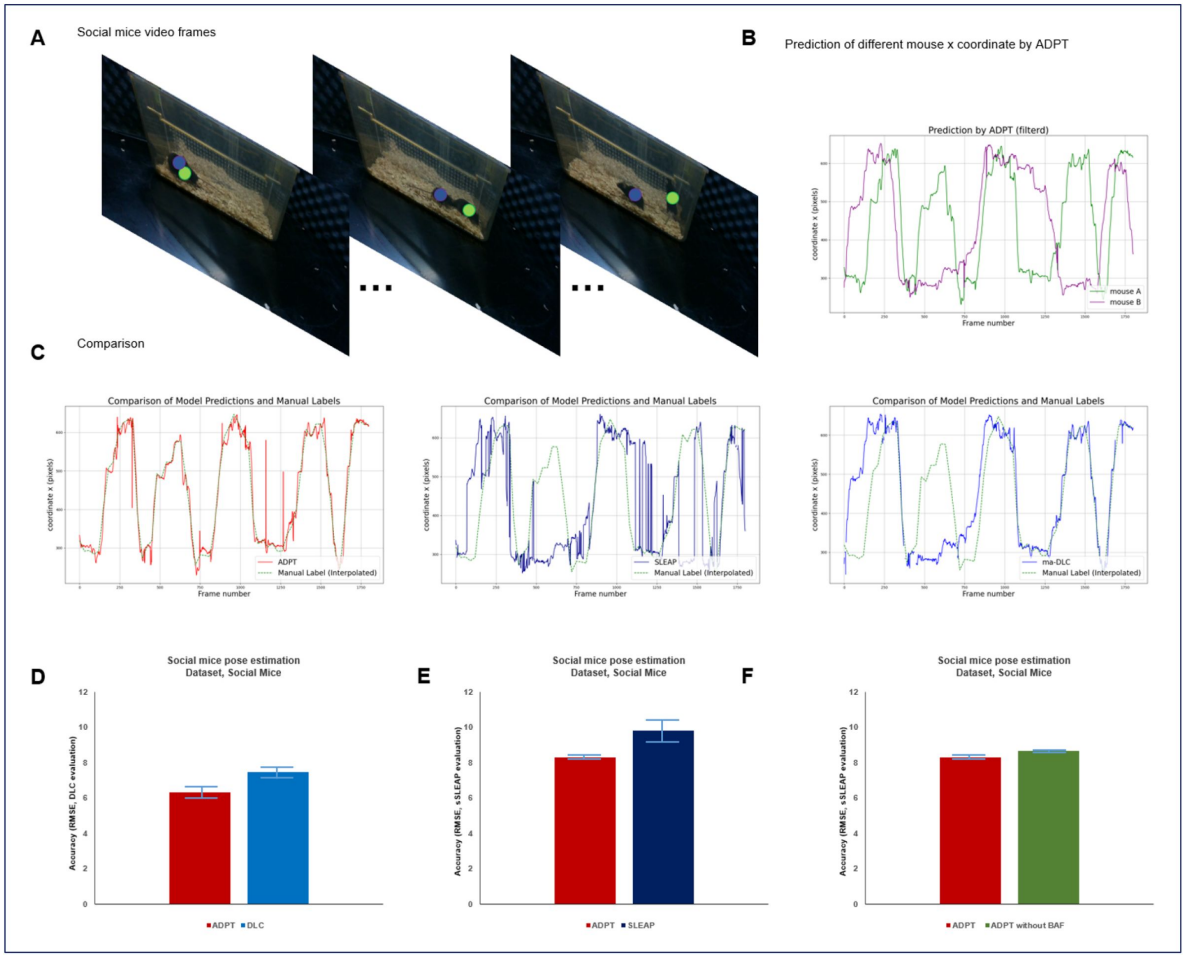

Figure 8.

Evaluation of ADPT for homecage social mice scenario.

A Illustration of homecage social mice dataset. **B** Filtered predicted back locations of different mice by ADPT. **C** Comparison of different methods and manual labels. We trained each model three times, and this figure presents the results from one of those training sessions. We calculated the average RMSE between predictions and manual labels, demonstrating that ADPT achieved an average RMSE of 15.8 ± 0.59 pixels, while DeepLabCut (DLC) and SLEAP recorded RMSEs of 113.19 ± 42.75 pixels and 94.76 ± 1.95 pixels, respectively. **D** Pose estimation accuracy comparison between ADPT and DLC based on the DLC evaluation metric. ADPT achieved an accuracy of 6.35 ± 0.14 pixels across all body parts of the mice, while DLC reached 7.49 ± 0.2 pixels. **E** Pose estimation accuracy comparison between ADPT and SLEAP using the SLEAP evaluation metric. ADPT achieved 8.33 ± 0.19 pixels across all body parts of the mice, compared to SLEAP's 9.82 ± 0.57 pixels. **F** Body Affinity Fields (BAF) improved pose estimation accuracy by 0.4 pixels under the SLEAP evaluation metric.

advances show that ADPT is an accurate, universal, and efficient method, suggesting broader application scenarios in neuroscience, genetics, and drug discovery. Now, the toolbox of ADPT has also been released at *https://github.com/tangguoling/ADPT-TOOLBOX*

As the higher resolution of microscopy promotes the discovery of biological microstructures, the higher precision of animal pose estimation helps to detect subtle behavior structures and patterns, advancing ethology research. Behavior structures have been proven to be the signatures, fingerprints, and biomarkers to indicate disease developments Bohic et al. (2023); Gschwind et al. (2023), genetic mutations Liu et al. (2021); Huang et al. (2021); Han et al. (2024), and drug effects Wiltschko et al. (2020); ?. Although these studies refine the behavior to module level Wiltschko et al. (2015), this spatiotemporal scale of behavior structures is not sufficient to support finer animal studies such as decoding millisecond neural recordings with un-drifted poses Schneider et al. (2023). Therefore, improving the accuracy and reliability of animal pose estimation is of high need for behavioral studies. ADPT provides such a tool for animal pose estimation.

ADPT enables a wide range of downstream applications, for instance, aligning behavioral manifold from keypoint dynamics with the neural manifold from large-scale neural recordings Urai et al. (2022). Recent advances in neural decoding of speech Li et al. (2023); Metzger et al. (2023) and vision Schneider et al. (2023); Takagi and Nishimoto (2023) have achieved incredible performance, but the accurate neural decoding of poses is still an existing problem. ADPT can quantify the poses of animals to reach a high resolution like the microphone for speech acquisition or visual pixels, which is an improvement from the aspect of behavior data acquisition. The second application is the gait analysis for 3D movements. Non-human primates are not restricted to moving on the ground, and the 3D gait would reflect their abnormal state after modeling treatment Liang et al. (2023); Thota and Alberts (2013). ADPT decreases the pose drift caused by body occlusion of single-view frames, which would reduce the error of 3D gait reconstruction. It also reduces the number of cameras for view angle compensation except for the profound understanding of 3D gait-related disorders Bala et al. (2020). The third application is behavior-based drug screening Wiltschko et al. (2020). Although MoSeq has built up the relationship between behavior syllables and psychoactive drugs Wiltschko et al. (2015, 2020), the resolution of behavior only exists at the syllable level. It is predictable for ADPT to improve the behavior resolution of MoSeq even Keypoint-MoSeq to a finer level to be not limited to the screen of psychoactive drugs Wiltschko et al. (2015); Weinreb et al. (2024). In summary, solving the anti-drift problem from the very beginning of ADPT determines that it has widespread applications.

One potential improvement of ADPT is the design of positional encoding. With the increase in image size, the positional encoding would occupy more memory of the graphics processing unit. The process of high-resolution videos has to resize the frame to avoid being out of memory, in which the pixel-level information could be missed. Conditional positional encoding would be a possible solution to improve ADPT to face high-resolution frames Chu et al. (2021). Another improvement of ADPT is using a more powerful backbone neural network. To facilitate the comparison between ADPT with other methods, the ResNet50 is used in all of the validation He et al. (2016). Recent advances in the backbone such as RegNet Xu et al. (2022a) could be the better choice to replace ResNet and improve the performance of ADPT.

Methods and materials

In this section, we first present ADPT method, then introduce the datasets used in each experiment, and finally describe the details of multi-animal experiments.

The details of ADPT

Here we present the key components and details of ADPT. We also provide the code for ADPT at *https://github.com/tangguoling/ADPT/code*

The network architecture

Applying transformer in freely behaving animal pose estimation can help us alleviate keypoint tracking drift. Thus, we created a heatmap-based pose estimation model, called ADPT. The overall structure of the method and network is illustrated in **Figure 1C** and **D**. Initially, ADPT will resize image to a scale (a hyper parameter, the same as global_scale in DLC). Then, the network employs the stack1-2 of the ResNet50 model to extract shallow-level features from the input images. At this stage, the images are extracted into features with a size of one-fourth of their original dimensions. Subsequently, network separately process these features in three branches, compute features at scale of one-fourth, one-eight and one-sixteenth, and generate one-eight scale features using convolution layer or deconvolution layer. Of particular significance is the utilization of the one-sixteenth scale feature, which is input into a transformer module for computation. This large-scale feature's involvement in the multi-head attention mechanism substantially enhances the model's ability to capture global relationships within the data. Finally, model concatenates these features by skip connections and compute them using convolution layers to generating output, including keypoint position confidence heatmaps, location refinement maps, low-resolution semantic segmentation map, and body affinity fields map.

Low resolution semantic segmentation

In addition to generating the animal's skeletal keypoints, we also create a low-resolution semantic segmentation map (LRSS) of the animal. This segmentation map captures the coarse-level information about the different body parts or regions of the animal. By connecting the skeletal keypoints, the model can infer the boundaries, shapes and identities of these regions. According to keypoints set *kps* of all individuals in frame, the pixel *p* value at segmentation map is defined as,

$$\mathcal{M}(p) = \begin{cases} identity, & \text{if } p \text{ on } limb_{ij}, \text{for } i, j \text{ in } kps \\ 0, & \text{otherwise} \end{cases} \quad (1)$$

The low-resolution map plays a crucial role in training our model. It allows the model to learn the correlation between the skeletal keypoints and the semantic information of the animal's body. By incorporating the segmentation map into the training process, the model can better understand the spatial relationships between different keypoints and improve the accuracy and robustness of pose estimation.

Network training details

In our animal pose estimation tasks, we employed specific training configurations to optimize the performance of our models. The following training details were utilized. We trained the models for a total of 190 epochs. Additionally, we included 10 warm-up epochs at the beginning of the training process. The batch size used during training was set to 8. We utilized the *AdamW* optimizer, and the weight decay rate was set to 1e-4. We employed a warmup cosine decay schedule for the learning rate. Initially, the learning rate was warmed up from 1e-5 to 1e-3 over the warm-up epochs. Subsequently, the learning rate gradually decayed to 1e-5 using a cosine decay function. For optimizing the keypoint confidence heatmaps and location refinement maps, we utilized root square error (RMSE) as the loss function. RMSE measures the average squared difference between the predicted and ground truth key points, providing a measure of the accuracy of the model's predictions. Additionally, for training the low-resolution semantic

segmentation map, we used sparse categorical cross-entropy loss, which is suitable for multi-class segmentation tasks. We early stop the training procedure when it reaches a plateau for 30 epochs according to validation loss. These training details were carefully chosen to ensure effective training and optimization of our models for single animal pose estimation. For data augmentation, we followed DeepLabCut augmentation strategy Mathis et al. (2018) in training ADPT, and followed Pereira et al. (2022) specifically for single fly dataset. The image inputs of ADPT were resized to a size that can be trained on the computer which was defined as 'global_scale' in configuration file. For mouse images, it was reduced to half of the original size. For monkey images, it was reduced to 0.8 of the original size. For macaque and fruit flies, there were no resizing, while for dogs, it was resized to 224 × 224 resolution. For homecage social mice images and marmoset images, there were no resizing.

The specific values and configurations may vary depending on the dataset, network architecture, and specific requirements of the task.

Network implementation

We implementated ADPT in the Python programming language(python 3.9). We used tensorflow 2.9.1 for all deep learning models. We used imgaug for image and annotation augmentation. We used OpenCV for video reading/writing and matplotlib for image reading. The hardware condition includes RTX4090 GPU, Intel 12900K CPU, Samsung 980 Pro hard disk, and 128 GB DDR5 memory. For comparison, we used DeepLabCut 2.2.1 with default configuration during training, in which 'global_scale' parameter was adjusted to match with ADPT resizing configuration. Similarly, SLEAP 1.2.9 was used with the baseline_medium_rf.single configuration, adjusting the 'input scaling' to align with ADPT's resizing configuration.

Datasets

To comprehensively evaluate the robust performance of ADPT, we selected datasets consider factors such as skeletal complexity, body size, and background complexity. However, there exists no publicly available video dataset specifically designed for anti-drift evaluation. Therefore, we also collected behavioral video data involving mice and monkeys. We also provide code to transfer DeepLabCut format labeled dataset to our ADPT format dataset, which may allow users to make deeper study toward the past behavioral data. Code is available at *https://github.com/tangguoling/ADPT/data/dlc2adpt.py*.

Mouse dataset

The mouse dataset is a customized single animal dataset collected by ourselves. We recorded a C57BL/6 mouse freely behaving in an open field from four different view. The dataset contained 440 labeled image in 1288 × 964 resolution across 4 different backgrounds and 11 individuals, 16 single mouse videos with the same resolution across 4 different individuals and 4 backgrounds. Each video spans 15 minutes.

Monkey dataset

The monkey dataset is a customized single animal dataset collected by ourselves. We recorded a cynomolgus monkey freely behaving in behavioral cage. The dataset contained 3488 labeled image in 640 × 360 resolution across 8 different backgrounds and multiple individuals, and one specific 30 minutes video in which a monkey and people appeared simultaneously.

Single fly dataset

The single fly dataset is a benchmark dataset used in animal pose estimation Pereira et al. (2022). It contained 1500 manual labeled frames which was split into 1,200 training, 150 validation and 150 test frames. The fly in the dataset was annotated with 32-node skeleton.

OpenMonkeyStudio Dataset

The OpenMonkeyStudio dataset is a macaque pose estimation dataset, containing 195,228 labeled frames with 13-node skeletons Bala et al. (2020). we randomly selected 5000 images and resized them to 368 × 368 resolution to evaluate the performance of our methods. We randomly divided this selected dataset into a 40% - 60% training and validation split.

StanfordExtradataset

StanfordExtradataset is a large-scale dog dataset with 2D keypoint and silhouette annotations, containing 12,000 images of dogs with 24-node skeletons Biggs et al. (2020). We randomly split the dataset into 85% training and 15% validation.

Mouse videos of different individuals

Videos in 1288 × 964 resolution across 4 different backgrounds and 10 individuals. Each video spans 15 minutes during which the first 12 minutes was used for training identity lrss while the rest was used for validation.

Free-social mice video

A 1 minute video in 1288 × 964 resolution of free-social mice.

Homecage social mice dataset

The homecage social mice dataset is a customized animal dataset collected by ourselves. We recorded two markerless C57BL/6 mice freely behaving in a homecage from three different view. The dataset contained 1200 labeled images in 960 × 540 resolution across 3 different backgrounds and 2 paired individuals. Each video spans 10 minutes during which the first 1 minute video was use for anti-drift performance evaluation. We manual annotated the position location in the 1 minute video every 30 frames.

Marmoset

Marmoset is a dataset with released by multi-animal DeepLabCut for marmosets pose estimation, containing 5,316 images of marmoset with 15-node skeletons Lauer et al. (2022). We resized the images to 368 × 368 resolution to evaluate the performance of our methods.

Mix-up social animal dataset generation

To address the challenge of acquiring labeled datasets for multi-animal pose estimation, we introduce a novel data augmentation strategy. This strategy involves mixing up a background picture and 2 labeled frames from single animal videos predicted by single animal model, generating synthetic data with multiple animals. The process is illustrated in **Figure 6**, and the algorithm is detailed in **Algorithm 1**. Initially, we employ the ADPT model to predict keypoint position for two images originating from different videos, resulting in two frame annotation sets of keypoints. Using these frames and the corresponding background image (**Figure 6 A**), we create a mix-up image, as shown in **Figure 6B**. We utilize two frame annotations to generate Mix-up annotation heatmaps. These heatmaps associate each keypoint with its corresponding location on the mix-up image, as shown in **Figure 6 C**. For the augmented image as shown in **Figure 6 D**, we generated augmented annotations as shown in **Figure 6 E**, and **Figure 6 F** represents augmented keypoints. Importantly, we leverage LRSS to distinguish between animals' identities, as indicated in the **Figure 6 G**. Finally, we leverage body affinity fields (BAF) to match the body parts and identity, as indicated in the **Figure 6 H** in which we set back as the center point.

Algorithm 1: Generation of Mix-up Social Animal Data
Algorithm 1: Generation of Mix-up Social Animal Data

Data: video1, video2, backgrounds
Result: Mix-up frame, Mix-up annotation
Tool: ADPT;
Select randomly;
$frame1 \in video1$;
$frame2 \in video2$;
$background \in backgrounds$;
Label frame;
$frame1\_annotation = ADPT(frame1)$;
$frame2\_annotation = ADPT(frame2)$;
$Mix\text{-}up\ annotation = \{frame1\_annotation, frame2\_annotation\}$;
Mix up image;
$mouse1 = frame1[where((frame1 - background) >= delta)]$;
$mouse2 = frame2[where((frame2 - background) >= delta \& mouse1 == 0)]$;
$Mix\text{-}up\ frame = mouse1 + mouse2 + background[where((mouse1 + mouse2) == 0)]$;

Body affinity fields

Inspired by PAF, we create Body Affinity Fields for associating body part to instance identity. Considering all individuals in frame, the pixel p value at BAF map is defined as,

$$\mathcal{M}(p) = \begin{cases} (p_x - center_x, p_y - center_y), & \text{if } p \text{ on } body_i, \text{for } i \text{ in } instances \\ (0,0), & \text{otherwise} \end{cases} \quad (2)$$

where p_x and p_y represents pixel p's location (x and y coordination), and $center_{x,y}$ represents the center location.

Combining BAF and LRSS, we can infer pixels identities. We only used this map in social animal tracking.

Experiments for ten mice identity tracking

In this experiment, we used videos featuring different identified mice, allocating 80% of the data for model training and the remaining 20% for accuracy validation. We configured the output channels of the model's LRSS to be 11 (1 background channel + 10 identity channels). Finally, we determined the identity of mice in the image by analyzing the proportion of each category within the LRSS image. For data augmentation, random rotation (±30°), random pixel translation (x: [-100,100], y:[-30,15]) and random scale (0.9,1.1) were used in training ADPT.

Following metrics was used for identity determination:

$$identity = argmax(\sum p_{identity}) \quad (3)$$

where $p_{identity}$ represents pixel p value at LRSS.

Experiments for social mice tracking

In this experiment, we randomly selected two mice. We created a Mix-up Social Keypoint Dataset using individual videos of these mice and randomly captured background. We computed the BAF centered on the back of the mice. For the social interaction task, the LRSS channels of the model were set to 3 (1 background channel and 2 identity channels), while 2 channels were introduced for the newly incorporated BAF (representing a two-dimensional vector). Random pixel translation (x:[-100,100], y:[-30,15]) was the only augmentation method used in training ADPT.

We trained the model on this mix-up dataset and used it to predict real social interaction videos of mice spanning 1 minute. In practical application, we employed a bidirectional approach both bottom-up and top-down to ascertain mouse identities. Specifically, we utilized the BAF image to confirm the center position pointed by each pixel. Then, based on the identity information from LRSS corresponding to the center positions, we determined the identity information of each pixel (body pixels) to generate an identity map. Finally, by matching the location heatmap with the identity map, we calculated the posture information of the interacting animals.

Both manual verification and following metrics was used for evaluating identity exchange rate:

$$changerate@\alpha = \frac{1}{F}\sum_{i=2}^{F}\delta(\sqrt{(y_i - y_{i-1})^2} \leq \alpha) \tag{4}$$

where y_i represents center location of each individual, and α represent drift distance threshold which was set as 75 pixels.

In our ten mice identity tracking and social mice tracking task, we trained the model for a total of 300 epochs with 10 warm-up epochs. We early stop the training procedure when it reaches a plateau for 30 epochs according to training loss. The batch size used during training was set to 8. Each epoch has 250 iterations for the first task and 50 iterations for the social task. To optimize the BAF maps, we utilized RMSE as the loss function.

Evaluation metrics

To evaluate keypoint tracking drift, we use following metrics: for each keypoint,

$$drift@\alpha = \frac{1}{F}\sum_{i=2}^{F}\delta(\sqrt{(y_i - y_{i-1})^2} \leq \alpha) \tag{5}$$

where F represents the total number of frames, y_i represent a predicted keypoint position, α represent the drift distance threshold which was set as 50 pixels on mice, and 30 pixels on monkey, and δ is an indicator function that equals 1 when $\sqrt{(y_i - y_{i-1})^2} \leq \alpha$, and 0 otherwise.

$$failtodetect = \frac{1}{F}\sum_{i=1}^{F}\delta(y_{i,confidencescore} \leq 0.2) \tag{6}$$

where $y_{i,confidencescore}$ represents the confidence score of the predicted heatmap of i-th frames.

We used the following metrics for single animal pose estimation: PCK@0.15, RMSE, mAP

$$PCK@0.15 = \frac{1}{N}\sum_{i=1}^{N}\delta(d_i \leq 0.15 \cdot L_i) \tag{7}$$

where N represents the total number of keypoints, d_i is the Euclidean distance for the i-th keypoint, L_i is the normalized limb scale associated with the i-th keypoint.

$$OKS = \frac{\sum_{i=1}^{N} \exp\left(-\frac{d_i^2}{2\alpha(2s)^2}\right)\delta_{v_i>0}}{\sum_{i=1}^{N}\delta_{v_i>0}} \tag{8}$$

where α is the bounding box area occupied by the GT instance, v_i is a visibility flag for the i-th keypoint, and s is the uncertainty factor(set to 0.025 for all measurements, the same as SLEAP)

$$AP@\alpha = \frac{1}{N}\sum_{i=1}^{N}\delta(OKS_i > \alpha) \tag{9}$$

where α represent the accuracy threshold.

$$mAP = \frac{1}{10}(AP@0.5 + AP@0.55 + AP@0.6 + ... + AP@0.95) \tag{10}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - y_{\text{true},i})^2} \tag{11}$$

where y_i represent a predicted keypoint position and $y_{\text{true},i}$ is its' ground truth.

Additional information

Author contributions

Guoling Tang, Conceptualization, Data acquisition, Code, Data analysis, Validation, Investigation, Visualization, Methodology, Writing - original draft, Writing - review and editing; Yaning Han, Conceptualization, Data acquisition, Validation, Investigation, Visualization, Methodology, Writing - review and editing; Xing Sun, Data acquisition, Methodology, Validation, Visualization; Ruonan Zhang, Methodology, Validation; Minghu Han, Supervision, Funding acquisition; Quanying Liu, Conceptualization, Supervision, Funding acquisition, Validation, Investigation, Visualization, Writing - review and editing; Pengfei Wei, Conceptualization, Supervision, Funding acquisition, Data acquisition, Validation, Investigation, Methodology, Writing - review and editing.

Funding

Funder	Grant reference number	Author
National Natural Science Foundation of China	32222036	Pengfei Wei
Research Fund for International Senior Scientists	T2250710685	Pengfei Wei
Shenzhen Science and Technology Innovation Committee	2022410129	Quanying Liu

Ethics

All experimental procedures of mice in this study were approved by Animal Care and Use Committees at the Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences. And all experimental procedures of monkey adhered to the Guidelines for the Care and Use of Laboratory Animals established by Jinan University.

Acknowledgements

We acknowledge the effort from Wenhao Liu who recorded the mouse behavioral data and Professor Sen Yan's laboratory who recorded the monkey behavioral data. This work was supported in part by National Natural Science Foundation of China (32222036 to PF. W), the Research Fund for International Senior Scientists (T2250710685 to PF. W), and Shenzhen Science and Technology Innovation Committee (2022410129 to Q. L). We thank ChatGPT for the English language editing of this paper.

Additional files

Supplementary Figure 1

Supplementary Table 1

Supplementary Video 1

Supplementary Video 2

Supplementary Video 3

Supplementary Video 4

Supplementary Video 5

Appendix 1

Deep learning pose estimation

Pose estimation is a well-established computer vision task that has achieved significant advancements in human pose estimation. Traditional CNN-based algorithms for human pose estimationNewell et al. (2016); Cao et al. (2021); Toshev and Szegedy (2014); Chen et al. (2018); Wei et al. (2016); Insafutdinov et al. (2016); Sun et al. (2019) have been widely applied and have shown promising results. With the recent rise of transformer-based models, researchers have explored the use of transformers for human pose estimationYang et al. (2021); Li et al. (2021); Xu et al. (2022b); Mao et al. (2021), leading to improved accuracy and performance. At the same time, some of these worksNewell et al. (2016); Wei et al. (2016); Insafutdinov et al. (2016); Xu et al. (2022b)has also been extended to the field of animal pose estimation. Notably, keypoint detection methods typically employ two main approaches: heatmap-based and regression-based methods. Heatmap-based methods generate keypoint heatmaps, calculate the index of the maximum confidence score within these heatmaps, and obtain keypoint coordinates. Heatmap-based methods have the advantage of providing confidence scores, allowing researchers to gauge the reliability of each keypoint's estimate. However, they can be computationally intensive due to the generation of multiple heatmaps. Conversely, regression-based methods directly output keypoint coordinates from the model. Regression-based methods are often computationally efficient and can provide accurate results. However, they may lack the ability to express the confidence or uncertainty associated with each keypoint prediction, which heatmap-based methods can provide. The choice between these methods depends on the specific requirements of the pose estimation task.

In the domain of behavioral studies, specific estimation methods have been developed and widely used. Notable examples include DeepLabCut Mathis et al. (2018), SLEAP Pereira et al. (2022), and DeepPoseKit Graving et al. (2019). These methods have found extensive application in experimental animal pose estimation, where the estimated poses are used for quantifying and analyzing animal behavior. They are heatmap-based methods. DeepLabCut is a popular toolbox utilized for animal pose estimation, employing CNNs such as ResNets He et al. (2016) or MobileNets Sandler et al. (2018) that initial pretrained on ImageNet Russakovsky et al. (2015) to accurately estimate animal poses. It has been widely adopted in various experimental settings, enabling researchers to track and analyze animal behavior with high precision. Similarly, SLEAP is another widely used tool for multi-animal pose estimation, leveraging U-NET Ronneberger et al. (2015) liked CNN architectures to estimate poses and facilitate behavior analysis in animals. Additionally, DeepPoseKit is another notable software toolkit using Stacked DenseNet for behavioral animal pose estimation. The results of pose estimation serve as a critical component in quantifying and analyzing animal behavior. By accurately estimating animal poses, researchers can extract valuable insights into the kinematics Monsees et al. (2022), dynamics Luxem et al. (2022), and patterns of animal movements Huang et al. (2021). This information further contributes to a better understanding of animal behavior, cognition, and underlying neural mechanisms.

According to literature report Pereira et al. (2022), SLEAP and DeepLabCut have similar accuracy on a benchmark single-fly datasets Pereira et al. (2019), with mean average precision scores(mAP) of 92.7% and 92.8%, respectively. Their accuracies are significantly higher than that of DeepPoseKit(86.4%). Additionally, SLEAP demonstrates the highest inference speed among the three tools. Therefore, currently, SLEAP and DeepLabCut are considered to have the best performance in freely behaving animal pose estimation. However, these methods are still limited by their robustness, which refers to the presence of uncertainty or noise interference in the estimated positions of keypoints due to the inherent limitations of the algorithms or noise in the image. For instance, the limited receptive fields of convolutional kernels may hinder their ability to capture the global dependencies within an image. This constraint can be particularly relevant in tasks that require modeling complex spatial relationships or long-range interactions. ADPT primarily aims to compare and improve upon these two methods.

In summary, various pose estimation methods, including DeepLabCut, SLEAP, and Deep-PoseKit, have been developed and extensively employed in the field of experimental animal pose estimation. These methods leverage CNN-based models to estimate animal poses, enabling researchers to conduct detailed behavior quantification and analysis.

References

Agezo S, Berman GJ (2022) **Tracking together: estimating social poses** *Nature Methods* **19**:410–411

Aljovic A, Zhao S, Chahin M, de la Rosa C, Steenbergen VV, Kerschensteiner M, Bareyre FM (2022) **A deep learning-based toolbox for Automated Limb Motion Analysis (ALMA) in murine models of neurological disorders** *Communications Biology* **5** https://doi.org/10.1038/s42003-022-03077-6

Baker S, Tekriwal A, Felsen G, Christensen E, Hirt L, Ojemann SG, Kramer DR, Kern DS, Thompson JA (2022) **Automatic extraction of upper-limb kinematic activity using deep learning-based markerless tracking during deep brain stimulation implantation for Parkinson's disease: a proof of concept study** *Plos one* **17**:e0275490

Bala PC, Eisenreich BR, Yoo SBM, Hayden BY, Park HS, Zimmermann J. (2020) **Automated markerless pose estimation in freely moving macaques with OpenMonkeyStudio** *Nature Communications* **11** https://doi.org/10.1038/s41467-020-18441-5

Biggs B, Boyne O, Charles J, Fitzgibbon A, Cipolla R. (2020) **Who left the dogs out? 3d animal reconstruction with expectation maximization in the loop** In: *Computer Vision–ECCV 2020: 16th European Conference* Springer pp. 195–211

Bohic M, Pattison LA, Jhumka ZA, Rossi H, Thackray JK, Ricci M, Mossazghi N, Foster W, Ogundare S, Twomey CR, et al. (2023) **Mapping the neuroethological signatures of pain, analgesia, and recovery in mice** *Neuron* **111**:2811–2830

Cao Z, Hidalgo G, Simon T, Wei SE, Sheikh Y. (2021) **OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields** *IEEE Transactions on Pattern Analysis and Machine Intelligence* **43** https://doi.org/10.1109/TPAMI.2019.2929257

Chen Y, Wang Z, Peng Y, Zhang Z, Yu G, Sun J. (2018) **Cascaded pyramid network for multi-person pose estimation** In: *Proceedings of the IEEE conference on computer vision and pattern recognition* pp. 7103–7112

Chu X, Tian Z, Zhang B, Wang X, Wei X, Xia H, Shen C. (2021) **Conditional positional encodings for vision transformers** *arXiv*

Gabriel CJ, Zeidler Z, Jin B, Guo C, Goodpaster CM, Kashay AQ, Wu A, Delaney M, Cheung J, Difazio LE, Sharpe MJ, Aharoni D, Wilke SA, Denardo LA (2022) **Behavior DEPOT is a simple, flexible tool for automated behavioral detection based on marker less pose tracking** *eLife* **11** https://doi.org/10.7554/eLife.74314

Graving JM, Chae D, Naik H, Li L, Koger B, Costelloe BR, Couzin ID (2019) **Deepposekit, a software toolkit for fast and robust animal pose estimation using deep learning** *eLife* **8** https://doi.org/10.7554/eLife.47994

Gschwind T, Zeine A, Raikov I, Markowitz JE, Gillis WF, Felong S, Isom LL, Datta SR, Soltesz I. (2023) **Hidden behavioral fingerprints in epilepsy** *Neuron* **111**:1440–1452

Han Y, Chen K, Wang Y, Liu W, Wang Z, Wang X, Han C, Liao J, Huang K, Cai S, et al. (2024) **Multi-animal 3D social pose estimation, identification and behaviour embedding with a few-shot learning framework** *Nature Machine Intelligence* **6**:48–61

Han Y, Huang K, Chen K, Pan H, Ju F, Long Y, Gao G, Wu R, Wang A, Wang L, et al. (2022) **MouseVenue3D: A markerless three-dimension behavioral tracking system for matching two-photon brain imaging in free-moving mice** *Neuroscience Bulletin* :1–15

He K, Zhang X, Ren S, Sun J. (2016) **Deep residual learning for image recognition** In: *Proceedings of the IEEE conference on computer vision and pattern recognition* pp. 770–778

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

Huang K, Han Y, Chen K, Pan H, Zhao G, Yi W, Li X, Liu S, Wei P, Wang L. (2021) **A hierarchical 3D-motion learning framework for animal spontaneous behavior mapping** *Nature Communications* **12** https://doi.org/10.1038/s41467-021-22970-y

Huang K, Yang Q, Han Y, Zhang Y, Wang Z, Wang L, Wei P. (2022) **An Easily Compatible Eye-tracking System for Freely-moving Small Animals** *Neuroscience Bulletin* **38** https://doi.org/10.1007/s12264-022-00834-9

Insafutdinov E, Pishchulin L, Andres B, Andriluka M, Schiele B. (2016) **Deepercut: A deeper, stronger, and faster multi-person pose estimation model** In: *Computer Vision–ECCV 2016: 14th European Conference* Springer pp. 34–50

Krakauer JW, Ghazanfar AA, Gomez-Marin A, MacIver MA, Poeppel D. (2017) **Neuroscience Needs Behavior: Correcting a Reductionist Bias** *Neuron* **93** https://doi.org/10.1016/j.neuron.2016.12.041

Lauer J, Zhou M, Ye S, Menegas W, Schneider S, Nath T, Rahman MM, Di Santo V, Soberanes D, Feng G, et al. (2022) **Multi-animal pose estimation, identification and tracking with DeepLabCut** *Nature Methods* **19**:496–504

LeCun Y, Bengio Y, Hinton G. (2015) **Deep learning** *nature* **521**:436–444

Li C, Lee GH (2021) **From synthetic to real: Unsupervised domain adaptation for animal pose estimation** In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* pp. 1482–1491

Li K, Wang S, Zhang X, Xu Y, Xu W, Tu Z. (2021) **Pose recognition with cascade transformers** In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* pp. 1944–1953

Li Y, Anumanchipalli GK, Mohamed A, Chen P, Carney LH, Lu J, Wu J, Chang EF (2023) **Dissecting neural computations in the human auditory pathway using deep neural networks for speech** *Nature Neuroscience* :1–13

Liang F, Yu S, Pang S, Wang X, Jie J, Gao F, Song Z, Li B, Liao WH, Yin M. (2023) **Non-human primate models and systems for gait and neurophysiological analysis** *Frontiers in Neuroscience* **17**:1141567

Liu N, Han Y, Ding H, Huang K, Wei P, Wang L. (2021) **Objective and comprehensive re-evaluation of anxiety-like behaviors in mice using the Behavior Atlas** *Biochemical and Biophysical Research Communications* **559**:1–7

Lonini L, Moon Y, Embry K, Cotton RJ, McKenzie K, Jenz S, Jayaraman A. (2022) **Video-Based Pose Estimation for Gait Analysis in Stroke Survivors during Clinical Assessments: A Proof-of-Concept Study** *Digital Biomarkers* **6** https://doi.org/10.1159/000520732

Luxem K, Mocellin P, Fuhrmann F, Kürsch J, Miller SR, Palop JJ, Remy S, Bauer P. (2022) **Identifying behavioral structure from deep variational embeddings of animal motion** *Communications Biology* **5** https://doi.org/10.1038/s42003-022-04080-7

Mao W, Ge Y, Shen C, Tian Z, Wang X, Wang Z. (2021) **Tfpose: Direct human pose estimation with transformers** *arXiv*

Marks M, Jin Q, Sturman O, von Ziegler L, Kollmorgen S, von der Behrens W, Mante V, Bohacek J, Yanik MF (2022) **Deep-learning-based identification, tracking, pose estimation and behaviour classification of interacting primates and mice in complex environments** *Nature Machine Intelligence* **4** https://doi.org/10.1038/s42256-022-00477-5

Mathis A, Mamidanna P, Cury KM, Abe T, Murthy VN, Mathis MW, Bethge M. (2018) **DeepLabCut: markerless pose estimation of user-defined body parts with deep learning** *Nature Neuroscience* **21** https://doi.org/10.1038/s41593-018-0209-y

Metzger SL, Littlejohn KT, Silva AB, Moses DA, Seaton MP, Wang R, Dougherty ME, Liu JR, Wu P, Berger MA, et al. (2023) **A high-performance neuroprosthesis for speech decoding and avatar control** *Nature* **620**:1037–1046

Monsees A, Voit KM, Wallace DJ, Sawinski J, Charyasz E, Scheffler K, Macke JH, Kerr JND (2022) **Estimation of skeletal kinematics in freely moving rodents** *Nature Methods* **19** https://doi.org/10.1038/s41592-022-01634-9

Newell A, Yang K, Deng J. (2016) **Stacked hourglass networks for human pose estimation** In: *Computer Vision–ECCV 2016: 14th European Conference* Springer pp. 483–499

Niknejad N, Caro JL, Bidese-Puhl R, Bao Y, Staiger EA (2023) **Equine kinematic gait analysis using stereo videography and deep learning: stride length and stance duration estimation** *Journal of the ASABE* **66** https://doi.org/10.13031/ja.15386

Pereira TD, Aldarondo DE, Willmore L, Kislin M, Wang SSH, Murthy M, Shaevitz JW (2019) **Fast animal pose estimation using deep neural networks** *Nature Methods* **16** https://doi.org/10.1038/s41592-018-0234-5

Pereira TD, Shaevitz JW, Murthy M. (2020) **Quantifying behavior to understand the brain** *Nature Neuroscience* **23** https://doi.org/10.1038/s41593-020-00734-z

Pereira TD, Tabris N, Matsliah A, Turner DM, Li J, Ravindranath S, Papadoyannis ES, Normand E, Deutsch DS, Wang ZY, McKenzie-Smith GC, Mitelut CC, Castro MD, D'Uva J, Kislin M, Sanes DH, Kocher SD, Wang SSH, Falkner AL, Shaevitz JW, et al. (2022) **SLEAP: A deep learning system for multi-animal pose tracking** *Nature Methods* **19** https://doi.org/10.1038/s41592-022-01426-1

Robinson GE, Fernald RD, Clayton DF (2008) **Genes and social behavior** *Science* **322** https://doi.org/10.1126/science.1159277

Ronneberger O, Fischer P, Brox T. (2015) **U-net: Convolutional networks for biomedical image segmentation** In: *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference* Springer pp. 234–241

Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, Huang Z, Karpathy A, Khosla A, Bernstein M, Berg AC, Fei-Fei L. (2015) **ImageNet Large Scale Visual Recognition Challenge** *International Journal of Computer Vision* **115** https://doi.org/10.1007/s11263-015-0816-y

Sandler M, Howard A, Zhu M, Zhmoginov A, Chen LC (2018) **Mobilenetv2: Inverted residuals and linear bottlenecks** In: *Proceedings of the IEEE conference on computer vision and pattern recognition* pp. 4510–4520

Schneider S, Lee JH, Mathis MW (2023) **Learnable latent embeddings for joint behavioural and neural analysis** *Nature* :1–9

Sheppard K, Gardin J, Sabnis GS, Peer A, Darrell M, Deats S, Geuther B, Lutz CM, Kumar V. (2022) **Stride-level analysis of mouse open field behavior using deep-learning-based pose estimation** *Cell reports* **38**

Stenum J, Rossi C, Roemmich RT (2021) **Two-dimensional video-based analysis of human gait using pose estimation** *PLoS Computational Biology* **17** https://doi.org/10.1371/journal.pcbi.1008935

Stoffl L, Vidal M, Mathis A. (2021) **End-to-end trainable multi-instance pose estimation with transformers** *arXiv*

Sun K, Xiao B, Liu D, Wang J. (2019) **Deep high-resolution representation learning for human pose estimation** In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* pp. 5693–5703

Takagi Y, Nishimoto S. (2023) **High-resolution image reconstruction with latent diffusion models from human brain activity** In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* pp. 14453–14463

Thota AK, Alberts JL (2013) **Novel use of retro-reflective paint to capture 3d kinematic gait data in non-human primates** In: *2013 29th Southern Biomedical Engineering Conference* IEEE pp. 113–114

Toshev A, Szegedy C. (2014) **Deeppose: Human pose estimation via deep neural networks** In: *Proceedings of the IEEE conference on computer vision and pattern recognition* pp. 1653–1660

Urai AE, Doiron B, Leifer AM, Churchland AK (2022) **Large-scale neural recordings call for new insights to link brain and behavior** *Nature neuroscience* **25**:11–19

Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser L, Polosukhin I. (2017) **Attention is all you need** *Advances in neural information processing systems* **30**

Vidal M, Wolf N, Rosenberg B, Harris BP, Mathis A. (2021) **Perspectives on individual animal identification from biology and computer vision** *Integrative and comparative biology* **61**:900–916

Wei SE, Ramakrishna V, Kanade T, Sheikh Y. (2016) **Convolutional pose machines** In: *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition* pp. 4724–4732

Weinreb C, Pearl JE, Lin S, Osman MAM, Zhang L, Annapragada S, Conlin E, Hoffmann R, Makowska S, Gillis WF, et al. (2024) **Keypoint-MoSeq: parsing behavior by linking point tracking to pose dynamics** *Nature Methods* **21**:1329–1339

Wiltschko AB, Johnson MJ, Iurilli G, Peterson RE, Katon JM, Pashkovski SL, Abraira VE, Adams RP, Datta SR (2015) **Mapping sub-second structure in mouse behavior** *Neuron* **88**:1121–1135

Wiltschko AB, Tsukahara T, Zeine A, Anyoha R, Gillis WF, Markowitz JE, Peterson RE, Katon J, Johnson MJ, Datta SR (2020) **Revealing the structure of pharmacobehavioral space through motion sequencing** *Nature neuroscience* **23**:1433–1443

Xu J, Pan Y, Pan X, Hoi S, Yi Z, Xu Z. (2022a) **RegNet: self-regulated network for image classification** In: *IEEE Transactions on Neural Networks and Learning Systems*

Xu Y, Zhang J, Zhang Q, Tao D. (2022b) **Vitpose: Simple vision transformer baselines for human pose estimation** *Advances in Neural Information Processing Systems* **35**:38571–38584

Yang S, Quan Z, Nie M, Yang W. (2021) **Transpose: Keypoint localization via transformer** In: *Proceedings of the IEEE/CVF International Conference on Computer Vision* pp. 11802–11812

Author information

Guoling Tang[†]

Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China, University of Chinese Academy of Sciences, Beijing, China
ORCID iD: 0009-0008-2318-2624

[†]These authors contributed equally to this work

Yaning Han[†]

Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China, University of Chinese Academy of Sciences, Beijing, China
ORCID iD: 0000-0002-1650-2262

[†]These authors contributed equally to this work

Xing Sun

Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China, University of Chinese Academy of Sciences, Beijing, China
ORCID iD: 0000-0002-7355-2100

Ruonan Zhang

Guangxi University of Science and Technology, Liuzhou, China
ORCID iD: 0009-0005-8337-217X

Minghu Han

Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China, Shenzhen University of Advanced Technology, Shenzhen, China

ORCID iD: 0000-0001-8613-4824

Quanying Liu
Department of Biomedical Engineering, Southern University of Science and Technology, Shenzhen, China
ORCID iD: 0000-0002-2501-7656

For correspondence: liuqy@sustech.edu.cn

Pengfei Wei
Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China, University of Chinese Academy of Sciences, Beijing, China
ORCID iD: 0000-0003-1845-8856

For correspondence: pf.wei@siat.ac.cn

Editors

Reviewing Editor
Gordon Berman
Emory University, Atlanta, United States of America

Senior Editor
Kate Wassum
University of California, Los Angeles, Los Angeles, United States of America

Reviewer #2 (Public review):

Summary:

The authors present a new model for animal pose estimation. The core feature they highlight is the model's stability compared to existing models in terms of keypoint drift. The authors test this model across a range of new and existing datasets. The authors also test the model with two mice in the same arena. For the single animal datasets the authors show a decrease in sudden jumps in keypoint detection and the number of undetected keypoints compared with DeepLabCut and SLEAP. Overall average accuracy, as measured by root mean squared error, generally shows generally similar but sometimes superior performance to DeepLabCut and better performance compared to SLEAP. The authors confusingly don't quantify the performance of pose estimation in the multi (two) animal case instead focusing on detecting individual identity. This multi-animal model is not compared with the model performance of the multi-animal mode of DeepLabCut or SLEAP.

Strengths:

The major strength of the paper is successfully demonstrating a model that is less likely to have incorrect large keypoint jumps compared to existing methods. As noted in the paper, this should lead to easier-to-interpret descriptions of pose and behavior to use in the context of a range of biological experimental workflows.

Weaknesses:

There are two main types of weaknesses in this paper. The first is a tendency to make unsubstantiated claims that suggest either model performance that is untested or misrepresents the presented data, or suggest excessively large gaps in current SOTA capabilities. One obvious example is in the abstract when the authors state ADPT

"significantly outperforms the existing deep-learning methods, such as DeepLabCut, SLEAP, and DeepPoseKit." All tests in the rest of the paper, however, only discuss performance with DeepLabCut and SLEAP, not DeepPoseKit. At this point, there are many animal pose estimation models so it's fine they didn't compare against DeepPoseKit, but they shouldn't act like they did. Similar odd presentation of results are statements like "Our method exhibited an impressive prediction speed of 90{plus minus}4 frames per second (fps), faster than DeepLabCut (44{plus minus}2 fps) and equivalent to SLEAP (106{plus minus}4 fps)." Why is 90{plus minus}4 fps considered "equivalent to SLEAP (106{plus minus}4 fps)" and not slower? I agree they are similar but they are not the same. The paper's point of view of what is "equivalent" changes when describing how "On the single-fly dataset, ADPT excelled with an average mAP of 92.83%, surpassing both DeepLabCut and SLEAP (Figure 5B)" When one looks at Figure 5B, however, ADPT and DeepLabCut look identical. Beyond this, oddly only ADPT has uncertainty bars (no mention of what uncertainty is being quantified) and in fact, the bars overlap with the values corresponding to SLEAP and DeepPoseKit. In terms of making claims that seem to stretch the gaps in the current state of the field, the paper makes some seemingly odd and uncited statements like "Concerns about the safety of deep learning have largely limited the application of deep learning-based tools in behavioral analysis and slowed down the development of ethology" and "So far, deep learning pose estimation has not achieved the reliability of classical kinematic gait analysis" without specifying which classical gait analysis is being referred to. Certainly, existing tools like DeepLabCut and SLEAP are already widely cited and used for research.

The other main weakness in the paper is the validation of the multi-animal pose estimation. The core point of the paper is pose estimation and anti-drift performance and yet there is no validation of either of these things relating to multi-animal video. All that is quantified is the ability to track individual identity with a relatively limited dataset of 10 mice IDs with only two in the same arena (and see note about train and validation splits below). While individual tracking is an important task, that literature is not engaged with (i.e. papers like Walter and Couzin, eLife, 2021: https://doi.org/10.7554/eLife.64000) and the results in this paper aren't novel compared to that field's state of the art. On the other hand, while multi-animal pose estimation is also an important problem the paper doesn't engage with those results either. The two methods already used for comparison in the paper, SLEAP and DeepPoseKit, already have multi-animal modes and multi-animal annotated datasets but none of that is tested or engaged with in the paper. The paper notes many existing approaches are two-step methods, but, for practitioners, the difference is not enough to warrant a lack of comparison. The authors state that "The evaluation of our social tracking capability was performed by visualizing the predicted video data (see supplement Videos 3 and 4)." While the authors report success maintaining mouse ID, when one actually watches the key points in the video of the two mice (only a single minute was used for validation) the pose estimation is relatively poor with tails rarely being detected and many pose issues when the mice get close to each other.

Finally, particularly in the methods section, there were a number of places where what was actually done wasn't clear. For example in describing the network architecture, the authors say "Subsequently, network separately process these features in three branches, compute features at scale of one-fourth, one-eight and one-sixteenth, and generate one-eight scale features using convolution layer or deconvolution layer." Does only the one-eight branch have deconvolution or do the other branches also? Similarly, for the speed test, the authors say "Here we evaluate the inference speed of ADPT. We compared it with DeepLabCut and SLEAP on mouse videos at 1288 x 964 resolution", but in the methods section they say "The image inputs of ADPT were resized to a size that can be trained on the computer. For mouse images, it was reduced to half of the original size." Were different image sizes used for training and validation? Or Did ADPT not use 1288 x 964 resolution images as input which would obviously have major implications for the speed comparison? Similarly, for the individual ID experiments, the authors say "In this experiment, we used videos featuring

different identified mice, allocating 80% of the data for model training and the remaining 20% for accuracy validation." Were frames from each video randomly assigned to the training or validation sets? Frames from the same video are very correlated (two frames could be just 1/30th of a second different from each other), and so if training and validation frames are interspersed with each other validation performance doesn't indicate much about performance on more realistic use cases (i.e. using models trained during the first part of an experiment to maintain ids throughout the rest of it.)

Editors' note: None of the original reviewers responded to our request to re-review the manuscript. The attached assessment statement is the editor's best attempt at assessing the extent to which the authors addressed the outstanding concerns from the previous round of revisions.

https://doi.org/10.7554/eLife.95709.2.sa1

Author response:

The following is the authors' response to the original reviews.

> ***eLife Assessment***
>
> *This study introduces a useful deep learning-based algorithm that tracks animal postures with reduced drift by incorporating transformers for more robust keypoint detection. The efficacy of this new algorithm for single-animal pose estimation was demonstrated through comparisons with two popular algorithms. However, the analysis is incomplete and would benefit from comparisons with other state-of-the-art methods and consideration of multi-animal tracking.*

First, we would like to express our gratitude to the eLife editors and reviewers for their thorough evaluation of our manuscript. ADPT aims to improve the accuracy of body point detection and tracking in animal behavior, facilitating more refined behavioral analyses. The insights provided by the reviewers have greatly enhanced the quality of our work, and we have addressed their comments point-by-point.

In this revision, we have included additional quantitative comparisons of multi-animal tracking capabilities between ADPT and other state-of-the-art methods. Specifically, we have added evaluations involving homecage social mice and marmosets to comprehensively showcase ADPT's advantages from various perspectives. This additional analysis will help readers better understand how ADPT effectively overcomes point drift and expands its applicability in the field.

> ***Reviewer #1:***
>
> *In this paper, the authors introduce a new deep learning-based algorithm for tracking animal poses, especially in minimizing drift effects. The algorithm's performance was validated by comparing it with two other popular algorithms, DeepLabCut and LEAP.The accessibility of this tool for biological research is not clearly addressed, despite its potential usefulness. Researchers in biology often have limited expertise in deep learning training, deployment, and prediction. A detailed, step-by-step user guide is crucial, especially for applications in biological studies.*

We appreciate the reviewers' acknowledgment of our work. While ADPT demonstrates superior performance compared to DeepLabCut and SLEAP, we recognize that the absence of a user-friendly interface may hinder its broader application, particularly for users with a background solely in biology. In this revision, we have enhanced the command-line version

of the user tutorial to provide a clear, step-by-step guide. Additionally, we have developed a simple graphical user interface (GUI) to further support users who may not have expertise in deep learning, thereby making ADPT more accessible for biological research.

> *The proposed algorithm focuses on tracking and is compared with DLC and LEAP, which are more adept at detection rather than tracking.*

In the field of animal pose estimation, the distinction between detection and tracking is often blurred. For instance, the title of the paper "SLEAP: A deep learning system for multi-animal pose tracking" refers to "tracking," while "detection" is characterized as "pose estimation" in the body text. Similarly, "Multi-animal pose estimation, identification, and tracking with DeepLabCut" uses "tracking" in the title, yet "detection" is also mentioned in the pose estimation section. We acknowledge that referencing these articles may have contributed to potential confusion.

To address this, we have clarified the distinction between "tracking" and "detection" Results section under " Anti-drift pose tracker." (see lines 118-119). In this paper, we now explicitly use “track” to refer to the tracking of all body points or poses of an individual, and “detect” for specific keypoints.

> ***Reviewer #1 recommendations:***
>
> *(1) DLC and LEAP are mainly good in detection, not tracking. The authors should compare their ADPT algorithm with idtracker.ai, ByteTrack, and other advanced tracking algorithms, including recent track-anything algorithms.*
>
> *(2) DeepPoseKit is outdated and no longer maintained; a comparison with the T-REX algorithm would be more appropriate.*

We appreciate the reviewer's suggestion for a more comprehensive comparison and acknowledge the importance of including these advanced tracking algorithms. However, we have not yet found suitable publicly available datasets for such comparative testing. We appreciate this insight and will consider incorporating T-REX into future comparisons.

> *(3) The authors primarily compared their performance using custom data. A systematic comparison with published data, such as the dataset reported in the paper "Multi-animal pose estimation, identification, and tracking with DeepLabCut," is necessary. A detailed comparison of the performances between ADPT and DLC is required.*

In the previous version of our manuscript, we included the SLEAP single-fly public dataset and the OMS_dataset from OpenMonkeyStudio for performance comparisons. We recognize that these datasets were not comprehensive. In this revision, we have added the marmoset dataset from "Multi-animal pose estimation, identification, and tracking with DeepLabCut" and a customized homecage social mice dataset to enhance our comparative analysis of multi-animal pose estimation performance. Our comprehensive comparison reveals that ADPT outperforms both DLC and SLEAP, as discussed in the Results section under "ADPT can be adapted for end-to-end pose estimation and identification of freely social animals.". (Figure 1, see lines 303-332)

> *(4) Given the focus on biological studies, an easy-to-use interface and introduction are essential.*

In this revision, we have not only developed a GUI for ADPT but also included a more detailed tutorial. This can be accessed at https://github.com/tangguoling/ADPT-TOOLBOX

> ***Reviewer #2:***
>
> *The authors present a new model for animal pose estimation. The core feature they highlight is the model's stability compared to existing models in terms of keypoint drift. The authors test this model across a range of new and existing datasets. The authors also test the model with two mice in the same arena. For the single animal datasets the authors show a decrease in sudden jumps in keypoint detection and the number of undetected keypoints compared with DeepLabCut and SLEAP. Overall average accuracy, as measured by root mean squared error, generally shows similar but sometimes superior performance to DeepLabCut and better performance compared to SLEAP. The authors confusingly don't quantify the performance of pose estimation in the multi (two) animal case instead focusing on detecting individual identity. This multi-animal model is not compared with the model performance of the multi-animal mode of DeepLabCut or SLEAP.*

We appreciate the reviewer's thoughtful assessment of our manuscript. Our study focuses on addressing the issue of keypoint drift prevalent in animal pose estimation methods like DeepLabCut and SLEAP. During the model design process, we discovered that the structure of our model also enhances performance in identifying multiple animals. Consequently, we included some results related to multi-animal identity recognition in our manuscript.

In recent developments, we are working to broaden the applicability of ADPT for multi-animal pose estimation and identity recognition. Given that our manuscript emphasizes pose estimation, we have added a comparison of anti-drift performance in multi-animal scenarios in this revision. This quantifies ADPT's capability to mitigate drift in multi-animal pose estimation.

Using our custom Homecage social mice dataset, we compared ADPT with DeepLabCut and SLEAP. The results indicate that ADPT achieves more accurate anti-drift pose estimation for two mice, with superior keypoint detection accuracy. Furthermore, we also evaluated pose estimation accuracy on the publicly available marmoset dataset, where ADPT outperformed both DeepLabCut and SLEAP. These findings are discussed in the Results section under "ADPT can be adapted for end-to-end pose estimation and identification of freely social animals."

> *The first is a tendency to make unsubstantiated claims that suggest either model performance that is untested or misrepresents the presented data, or suggest excessively large gaps in current SOTA capabilities. One obvious example is in the abstract when the authors state ADPT "significantly outperforms the existing deep-learning methods, such as DeepLabCut, SLEAP, and DeepPoseKit." All tests in the rest of the paper, however, only discuss performance with DeepLabCut and SLEAP, not DeepPoseKit. At this point, there are many animal pose estimation models so it's fine they didn't compare against DeepPoseKit, but they shouldn't act like they did.*

We appreciate the reviewer's feedback regarding unsubstantiated claims in our manuscript. Upon careful review, we acknowledge that our previous revisions inadvertently included statements that may misrepresent our model's performance. In particular, we have revised the abstract to eliminate the mention of DeepPoseKit, as our comparisons focused exclusively on DeepLabCut and SLEAP.

In addition to this correction, we have thoroughly reviewed the entire manuscript to address other instances of ambiguity and ensure that our claims are well-supported by the data presented. Thank you for bringing this to our attention; we are committed to maintaining the integrity of our claims throughout the paper.

> *In terms of making claims that seem to stretch the gaps in the current state of the field, the paper makes some seemingly odd and uncited statements like "Concerns about the safety of deep learning have largely limited the application of deep learning-based tools in behavioral analysis and slowed down the development of ethology" and "So far, deep learning pose estimation has not achieved the reliability of classical kinematic gait analysis" without specifying which classical gait analysis is being referred to. Certainly, existing tools like DeepLabCut and SLEAP are already widely cited and used for research.*

In this revision, we have carefully reviewed the entire manuscript and addressed the instances of seemingly odd and unsubstantiated claims. Specifically, we have revised the statements "largely limited" to "limited" to ensure accuracy and clarity. Additionally, we thoroughly reviewed the citation list to ensure proper attribution, incorporating references such as "A deep learning-based toolbox for Automated Limb Motion Analysis (ALMA) in murine models of neurological disorders" to better substantiate our claims and provide a clearer context.

We have also added an additional section to comprehensively discuss the applications of widely-used tools like DeepLabCut and SLEAP in behavioral research. This new section elaborates on the challenges and limitations researchers encounter when applying these methods, highlighting both their significant contributions and the areas where improvements are still needed.

> *The other main weakness in the paper is the validation of the multi-animal pose estimation. The core point of the paper is pose estimation and anti-drift performance and yet there is no validation of either of these things relating to multi-animal video. All that is quantified is the ability to track individual identity with a relatively limited dataset of 10 mice IDs with only two in the same arena (and see note about train and validation splits below). While individual tracking is an important task, that literature is not engaged with (i.e. papers like Walter and Couzin, eLife, 2021: https://doi.org/10.7554/eLife.64000) and the results in this paper aren't novel compared to that field's state of the art. On the other hand, while multi-animal pose estimation is also an important problem the paper doesn't engage with those results either. The two methods already used for comparison in the paper, SLEAP and DeepPoseKit, already have multi-animal models and multi-animal annotated datasets but none of that is tested or engaged with in the paper. The paper notes many existing approaches are two-step methods, but, for practitioners, the difference is not enough to warrant a lack of comparison.*

We appreciate the reviewer's insights regarding the validation of multi-animal pose estimation in our paper. While our primary focus has been on pose estimation and anti-drift performance, we recognize the importance of validating these aspects within the context of multi-animal videos.

In this revision, we have included a comparison of ADPT's anti-drift performance in multi-animal pose estimation, utilizing our custom Homecage social mouse dataset (Figure 1A). Our findings indicate that ADPT achieves more accurate pose estimation for two mice while significantly reducing keypoint drift, outperforming both DeepLabCut and SLEAP. (see lines 311-322). We trained each model three times, and this figure presents the results from one of those training sessions. We calculated the average RMSE between predictions and manual labels, demonstrating that ADPT achieved an average RMSE of 15.8 ± 0.59 pixels, while DeepLabCut (DLC) and SLEAP recorded RMSEs of 113.19 ± 42.75 pixels and 94.76 ± 1.95 pixels, respectively (Figure 1C). ADPT achieved an accuracy of 6.35 ± 0.14 pixels based on the DLC evaluation metric across all body parts of the mice, while DLC reached 7.49 ± 0.2 pixels (Figure 1D). ADPT achieved 8.33 ± 0.19 pixels using the SLEAP evaluation Metric across all body parts of the mice, compared to SLEAP's 9.82 ± 0.57 pixels (Figure 1E).

Furthermore, we have conducted pose estimation accuracy evaluations on the publicly available marmoset dataset from DeepLabCut, where ADPT also demonstrated superior performance compared to DeepLabCut and SLEAP. These results can be found in the "ADPT can be adapted for end-to-end pose estimation and identification of freely social animals" section of the Results. (see lines 323-329)

We acknowledge the existing literature on multi-animal tracking, such as the work by Walter and Couzin (2021). While individual tracking is crucial, our primary focus lies in the effective tracking of animal poses and minimizing drift during this process. This dual emphasis on pose tracking and anti-drift performance distinguishes our work and aligns with ongoing advancements in the field. Engaging with relevant literature, highlights the importance of contextualizing our results within the broader tracking literature, demonstrating that while our findings may overlap with existing methods, the unique focus on improving tracking stability and reducing drift presents valuable contributions to the field. Thank you for your valuable feedback, which has helped us improve the robustness of our manuscript.

> *The authors state that "The evaluation of our social tracking capability was performed by visualizing the predicted video data (see supplement Videos 3 and 4)." While the authors report success maintaining mouse ID, when one actually watches the key points in the video of the two mice (only a single minute was used for validation) the pose estimation is relatively poor with tails rarely being detected and many pose issues when the mice get close to each other.*

We acknowledge that there are indeed challenges in pose estimation, particularly when the two mice get close to each other, leading to tracking failures and infrequent detection of tails in the predicted videos. The reasons for these issues can be summarized as follows:

Lack of Training Data from Real Social Scenarios: The training data used for the social tracking assessment were primarily derived from the Mix-up Social Animal Dataset, which does not fully capture the complexities of real social interactions. In future work, we plan to incorporate a blend of real social data and the Mix-up data for model training. Specifically, we aim to annotate images where two animals are in close proximity or interacting to enhance the model's understanding of genuine social behaviors.

Challenges in Tail Tracking in Social Contexts: Tracking the tails of mice in social situations remains a significant challenge. To validate this, we have added an assessment of tracking performance in real social settings using homecage data. Our findings indicate that using annotated data from real environments significantly improves tail tracking accuracy, as demonstrated in the supplementary video.

We appreciate your feedback, which highlights critical areas for improvement in our model.

> *Finally, particularly in the methods section, there were a number of places where what was actually done wasn't clear.*

We have carefully reviewed and revised the corresponding parts to clarify the previously incomprehensible statements. Thank you for your valuable feedback, which has helped enhance the clarity of our methods.

> *For example in describing the network architecture, the authors say "Subsequently, network separately process these features in three branches, compute features at scale of one-fourth, one-eight and one-sixteenth, and generate one-eight scale features using convolution layer or deconvolution layer." Does only the one-eight branch have deconvolution or do the other branches also?*

We apologize for the confusion this has caused. Upon reviewing our manuscript, we identified an error in the diagram. In the revised version, we have clarified that the model samples feature maps at multiple resolutions and ultimately integrates them at the 1/8 resolution for feature fusion. Specifically, the 1/4 feature map from ResNet50's stack 2 is processed through max-pooling and convolution to generate a 1/8 feature map. Additionally, the 1/4 feature map from ResNet50's stack 2 is also transformed into a 1/8 feature map using a convolution operation with a stride of 2. Finally, both the input and output of the transformer are at the 1/16 resolution, which can be trained on a 2080Ti GPU. The 1/16 feature map is then upsampled to produce the final 1/8 feature map. We have updated the manuscript to reflect these changes, and we also modified the model architecture diagram for better clarity.

> *Similarly, for the speed test, the authors say "Here we evaluate the inference speed of ADPT. We compared it with DeepLabCut and SLEAP on mouse videos at 1288 x 964 resolution", but in the methods section they say "The image inputs of ADPT were resized to a size that can be trained on the computer. For mouse images, it was reduced to half of the original size." Were different image sizes used for training and validation? Or Did ADPT not use 1288 x 964 resolution images as input which would obviously have major implications for the speed comparison?*

For our inference speed evaluation, all models, including ADPT, used images with a resolution of 1288 x 964. In ADPT's processing pipeline, the first layer is a resizing layer designed to compress the images to a scale determined by the global scale parameter. For the mouse images, we set the global scale to 0.5, allowing our GPU to handle the data at that resolution during transformer training.

We recorded the time taken by ADPT to process the entire 15-minute mouse video, which included the time taken for the resizing operation, and subsequently calculated the frames per second (FPS). We have clarified this process in the manuscript, particularly in the "Network Architecture" section, where we specify: "Initially, ADPT will resize the images to a390 scale (a hyperparameter, consistent with the global scale in the DLC configuration)."

> *Similarly, for the individual ID experiments, the authors say "In this experiment, we used videos featuring different identified mice, allocating 80% of the data for model training and the remaining 20% for accuracy validation." Were frames from each video randomly assigned to the training or validation sets? Frames from the same video are very correlated (two frames could be just 1/30th of a second different from each other), and so if training and validation frames are interspersed with each other validation performance doesn't indicate much about performance on more realistic use cases (i.e. using models trained during the first part of an experiment to maintain ids throughout the rest of it.)*

In our study, we actually utilized the first 80% of frames from each video for model training and the remaining 20% for testing the model's ID tracking accuracy. We have revised the relevant description in the manuscript to clarify this process. The updated description can be found in the "Datasets" section under "Mouse Videos of Different Individuals."

https://doi.org/10.7554/eLife.95709.2.sa0