International Journal of Information Technology Vol. 30 No. 2 2024

3D Pose and Shape Reconstruction of Freely Moving Organisms

Jiachen Zhao[a]

[a]Joint NTU-WeBank Research Centre on Fintech, Nanyang Technological University, 50 Nanyang Ave, Singapore 639798

zhao_jiachen@163.com

Abstract

Biological behavior directly reflects brain activity. Measuring and analyzing such behavior is one of the most fundamental tasks in neuroscience and Medicine. To avoid the time-consuming and inaccurate methods of manual observation, this paper presents a deep learning-based approach for 3D pose and shape reconstruction of freely moving organisms. Using mice as a case study, the method takes multi-view image data as input and utilizes an end-to-end neural network to estimate 2D and 3D poses. A self-supervised loss function also reduces reliance on large training data. Finally, the approach reconstructs the shape of the mouse using 3D pose data and mesh models. Visual experimental results confirm the effectiveness of this approach in capturing both skeletal movements and surface dynamics.

Keywords: 3D pose estimation, 3D shape reconstruction.

I. Introduction

Accurate measurement of animal movements and postures is essential for understanding fundamental biological processes, including motor control, sensory processing, and neurological responses to stimuli. Advanced computer vision technologies have been developed to track and

quantify animal movements automatically [1, 2, 3], offering a level of precision that was previously unattainable with manual tracking, thus enabling more detailed and large-scale behavioral studies. However, existing works have two main limitations. The first is the reliance on large amounts of training data, whereas animal motion capture data is limited compared to human data. One is the heavy reliance on large amounts of training data, while animal motion capture datasets are often much smaller compared to human data. Another issue is that 3D pose representations based on key points fail to capture the animal's surface details.

To address these challenges, this paper proposes an efficient multi-view 3D pose estimation and shape reconstruction method for animal data. 3D pose estimation aims to detect the spatial coordinates of key body joints or landmarks from images. We utilize an end-to-end trainable neural network to predict the 3D pose. This model makes the triangulation process differentiable and implements self-supervised training based on multi-view geometric consistency, reducing the reliance on labeled training data. Moreover, we utilize an optimization-based method to fit a 3D mesh model to the pose, generating a detailed and accurate representation of an object's surface.

II. Related works

A. 3D pose estimation for organisms

The 3D pose estimation methods can generally be categorized into single-view lifting, multi-view triangulation, and multi-view voxel-based regression. Single-view lifting methods directly estimate the 3D pose from a single 2D image by inferring depth information through a neural network [4, 5]. Multi-view triangulation utilizes the 2D joint positions from multiple camera views to estimate 3D joint coordinates, recent works introduce the soft-argmax and weight-based optimization to triangulation to make it end-to-end trainable [6, 7]. Voxel-based methods [8, 9] predict 3D voxel

grids and regress the pose from the volumetric representation. This paper focuses on the learnable triangulation approach.

B. 3D shape reconstruction for organisms

The 3D mesh reconstruction methods can be broadly categorized into two types: point cloud-based methods and parametric model-based methods. Point cloud-based methods use spatial point sets obtained from scanning techniques to reconstruct the surface by extracting features and performing surface reconstruction. Parametric model-based methods rely on predefined geometric shapes or templates, optimizing the parameters to fit the specific characteristics of the object. The SMPL (Skinned Multi-Person Linear) model [10] is a popular parametric model for representing human body shapes and poses. Some animal-specific models have also been proposed, adapting parametric approaches like SMPL to account for species-specific anatomy and movements [11].

III. Method

A. Overview

Fig. 1 shows the pipeline of 3D pose estimation and shape reconstruction. Our model takes synchronized multi-view videos of freely moving organisms as input (using mice as an example). First, an HRNet backbone is utilized to extract feature maps from the images. These feature maps are then processed by a heatmap decoder to predict the 2D coordinates of the joints in each view. Simultaneously, the feature maps are passed through a weight decoder to learn the weights of each view for the triangulation process. A learnable triangulation module estimates the 3D pose based on the output of decoders. All the above process is end-to-end trainable. After obtaining

the real 3D pose of the mouse, a predefined mouse mesh model is fitted to the pose, enabling the reconstruction of the mouse's 3D mesh.

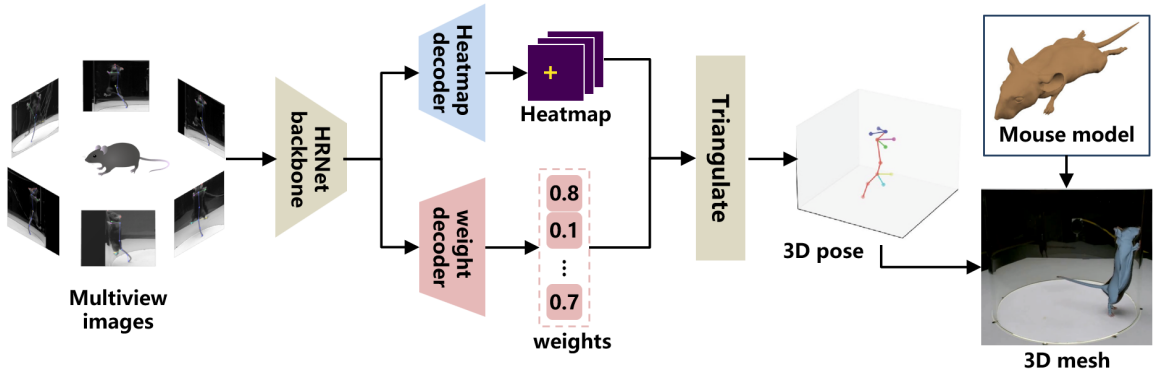


Fig 1: Pipeline of 3D pose estimation and shape reconstruction.

B. Heatmap-based 2D keypoint detection

The multi-view 3D pose estimation pipeline begins with 2D joint detection for each view. We utilize the top-down heatmap-based method. First, Faster R-CNN is applied to detect individuals within the image, followed by 2D keypoint estimation within the detected bounding box. Then we use a heatmap-based approach with HRNet as the backbone. HRNet extracts feature maps from the image, which are then passed through a heatmap decoder to generate heatmaps for each joint. Finally, the keypoint heatmaps are converted into 2D coordinates via a soft-argmax function. The soft-argmax function, by selecting the peak of each heatmap corresponding to the predicted 2D joint location in a differentiable manner, makes end-to-end training of the entire network possible.

C. 3D keypoint estimation via self-supervised triangulation

We modify the traditional triangulation to learnable self-supervised triangulation and use it to estimate the 3D keypoint coordinates. Triangulation uses the 2D coordinates from different camera views and the known camera parameters to project the 2D points back into 3D space. Each camera

provides a ray along which the 3D point lies. By finding the intersection of these rays, we can determine the 3D position of the joint. However, due to errors in real-world data, the rays from each view will not perfectly coincide at a single point. Therefore, triangulation is typically formulated as an optimization problem as shown in Eq. (1), where the goal is to minimize the reprojection error, i.e. the difference between the observed 2D points $\hat{x}_{c,j}$ and the 2D projections of the estimated 3D point X_j across all views.

$$X_j^* = \arg\min_{X_j} \sum_{c=1}^{C} ||\hat{x}_{c,j} - P_c \cdot X_j||^2 \tag{1}$$

where $\hat{x}_{c,j}$ is the detected jth 2D joint in cth view via the 2D detector, P_c is the projection matrix of cth camera, X_j is the estimated 3D joint. The optimization problem in Eq. (1) can be solved via SVD on the overdetermined system of equations on homogeneous $A_j \cdot X_j = 0$, where A_j is the triangulation matrix of jth joint.

The traditional triangulation utilizes the RANSAC to handle outliers and noise in the 2D keypoint detections, but RANSAC is not differentiable. Therefore, we introduce a weight to softly select the high-quality 2D detections, and the optimization problem could be formulated as Eq. (2), where $w_j \in \mathbb{R}^J$ is a learnable weight vector for J joints. We utilize a weight decoder following the HRNet backbone to output the weight vector w_j.

$$w_j \odot A_j \cdot X_j = 0 \tag{2}$$

We formulated the Heatmap-based 2D keypoint detection and the self-supervised triangulation in an end-to-end neural network. The total model is trained with 2D supervision loss, 3D supervision loss, and 3D self-supervision loss. The 2D supervision loss is MSE between the 2D

ground truth heatmap and the estimated heatmap by the heatmap decoder. The 3D supervision loss [6] is MSE between the 3D ground truth and 3D estimated coordinates by learnable triangulation. The 3D self-supervision loss is defined as the sum of distances between the 3D estimate and view rays, which can realign the predicted heatmap locations to the accurate point using multi-view consistency[7].

D. Optimization-based 3D shape reconstruction

We follows the pipeline of optimzation-based 3D shape reconstruction using a mesh model [11]. Similar to SMPL model, the pose of the mouse mesh model is controlled by joint angles, The mesh vertices are connected to joints through a set of weights, which determine the influence each joint has on the deformation of the mesh. To fit the mesh model to the captured pose of mice, we can minimize the distance between the model's joint locations and the captured 3D joint positions using PyTorch's automatic differentiation and backpropagation.

IV. Experiments

A. 3D pose estimation results

Fig. 2 shows the pose estimation results for a single mouse. We visually compare our method with existing approaches on the Dannce dataset[3], which includes six cameras surrounding a freely moving mouse. We captured the positions of 12 joints across the mouse's body. The first column displays the baseline, using naive triangulation; the second column shows the results of learnable triangulation with a 3D supervised loss function; and the third column demonstrates the performance of our proposed method. The example in Fig. 2 illustrates that, even when

2D detection errors occur in some views due to limited training data, our method, using a self-supervised loss function, still outputs accurate 3D poses.

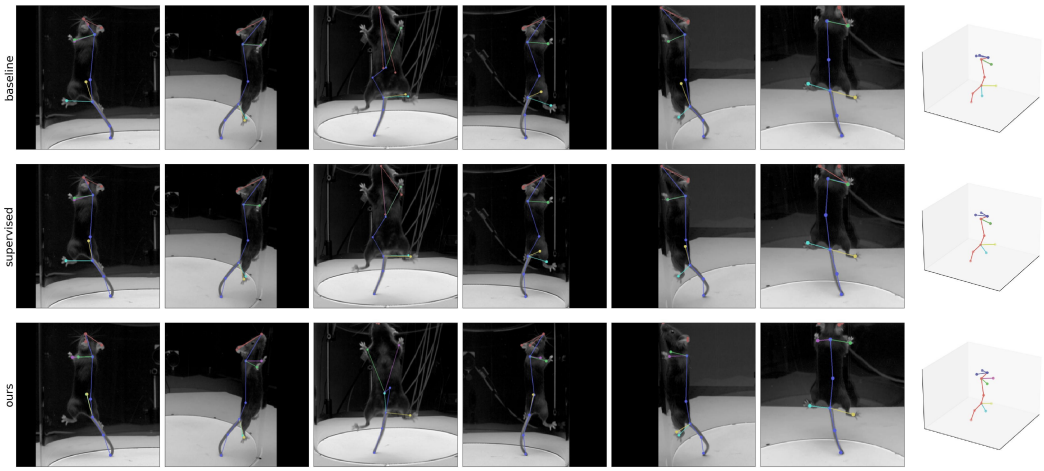
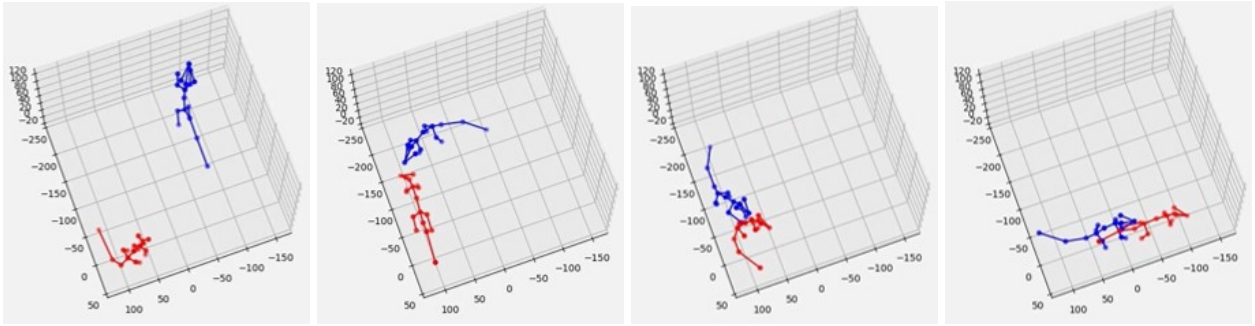


Fig 2: 3D pose capture results of a single mouse.

Fig. 3 shows the motion capture results for two interacting mice, with four pose examples extracted from a chasing interaction sequence. The blue skeleton represents the male mouse, while the red skeleton represents the female mouse. The results demonstrate that our method accurately estimates the 3D pose for interactive behaviors, including close-range sniffing and chasing actions, as seen in Examples 3 and 4.

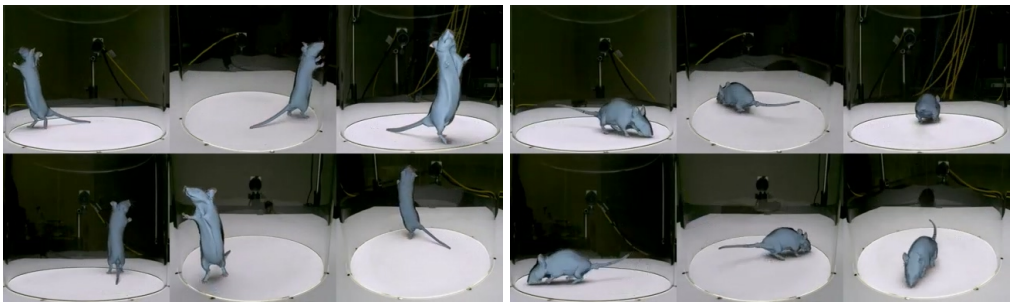


(a) Example 1 (b) Example 2 (c) Example 3 (d) Example 4

Fig 3: 3D pose results of two interacting mice.

B. 3D shape reconstruction results

Fig. 4 shows two examples of 3D mesh reconstruction on the Dannce dataset[3], with each example displaying the shape reconstruction results from the six camera views in the Dannce dataset. The two examples showcase two distinct mouse actions: standing and crawling. The results demonstrate that our method successfully reconstructs the mouse's full 3D shape, including the head, limbs, tail, and body mesh, during different poses.



(a) Example 1 (b) Example 2

Fig 4: 3D shape reconstruction results.

V. Conclusion

This paper proposes a method for 3D pose estimation and shape reconstruction of mouse, which is This pipeline is also applicable to other species, allowing for flexible adaptation to different organisms. The approach takes multi-view images as input and employs an end-to-end neural network for 2D keypoint detection and 3D triangulation. It also incorporates a self-supervised training function for data-efficient learning. Finally, an optimization-based method is used for 3D shape reconstruction, enabling precise and quantitative description of the 3D pose and shape of freely moving mice. Our future research plans include integrating skeletal and muscular models to achieve more comprehensive biological body modeling. This will enhance the accuracy of neuromechanical simulations, providing deeper insights into movement control and improving

applications in drug discovery, behavioral analysis, and bio-inspired robotics.

References

[1] Nancy Padilla-Coreano, Kanha Batra, Makenzie Patarino, Zexin Chen, Rachel R Rock, Ruihan Zhang, Sébastien B Hausmann, Javier C Weddington, Reesha Patel, Yu E Zhang, et al. Cortical ensembles orchestrate social competition through hypothalamic outputs. *Nature*, 603(7902):667–671, 2022.

[2] Alexander Mathis, Pranav Mamidanna, Kevin M. Cury, Taiga Abe, Venkatesh N. Murthy, Mackenzie Weygandt Mathis, and Matthias Bethge. DeepLabCut: markerless pose estimation of user-defined body parts with deep learning. *Nature Neuroscience*, 21(9):1281–1289, 2018.

[3] Timothy W. Dunn, Jesse D. Marshall, Kyle S. Severson, Diego E. Aldarondo, David G.C. Hildebrand, Selmaan N. Chettih, William L. Wang, Amanda J. Gellis, David E. Carlson, Dmitriy Aronov, Winrich A. Freiwald, Fan Wang, and Bence P. Ölveczky. Geometric deep learning enables 3D kinematic profiling across species and environments. *Nature Methods*, 18(5):564–573, 2021.

[4] Denis Tome, Chris Russell, and Lourdes Agapito. Lifting from the deep: Convolutional 3d pose estimation from a single image. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[5] Yangyuxuan Kang, Yuyang Liu, Anbang Yao, Shandong Wang, and Enhua Wu. 3d human pose lifting with grid convolution. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023.

[6] Karim Iskakov, Egor Burkov, Victor Lempitsky, and Yury Malkov. Learnable triangulation

of human pose. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 7717–7726, 2019.

[7] Jiachen Zhao, Tao Yu, Liang An, Yipeng Huang, Fang Deng, and Qionghai Dai. Triangulation residual loss for data-efficient 3d pose estimation. *Advances in Neural Information Processing Systems*, 36, 2024.

[8] Hanyue Tu, Chunyu Wang, and Wenjun Zeng. Voxelpose: Towards multi-camera 3d human pose estimation in wild environment. In *European Conference on Computer Vision (ECCV)*, 2020.

[9] Hang Ye, Wentao Zhu, Chunyu Wang, Rujie Wu, and Yizhou Wang. Faster voxelpose: Real-time 3d human pose estimation by orthographic projection. In *European Conference on Computer Vision (ECCV)*, 2022.

[10] Georgios Pavlakos, Vasileios Choutas, Nima Ghorbani, Timo Bolkart, Ahmed A. A. Osman, Dimitrios Tzionas, and Michael J. Black. Expressive body capture: 3d hands, face, and body from a single image. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[11] Liang An, Jilong Ren, Tao Yu, Tang Hai, Yichang Jia, and Yebin Liu. Three-dimensional surface motion capture of multiple freely moving pigs using mammal. *Nature Communications*, 14(1):7727, 2023.