SUPPLEMENTARY A

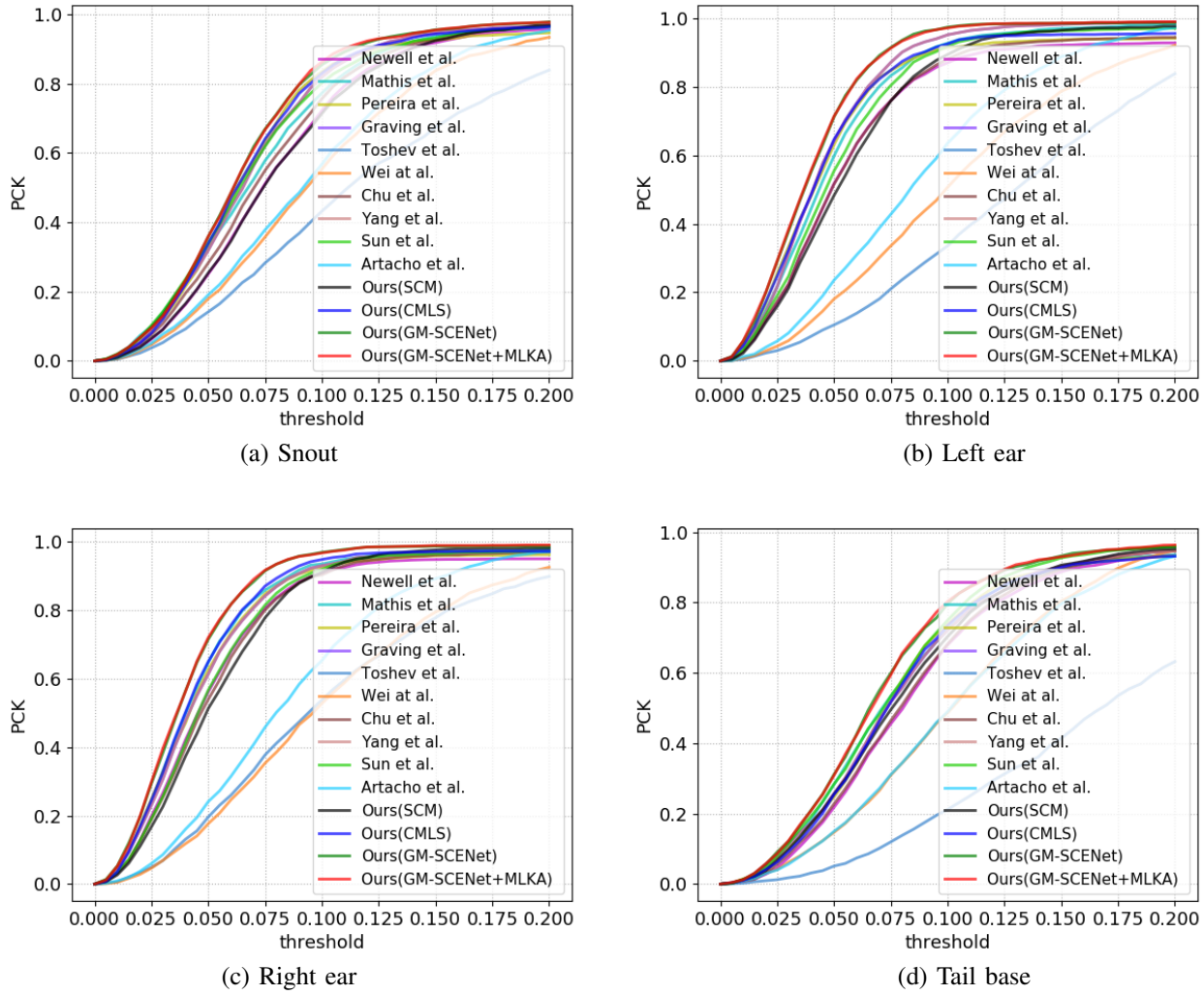
TABLE S1
ABLATION EXPERIMENTS OF THE STRUCTURED CONTEXT MIXER ON DEEPLABCUT MOUSE POSE TEST DATASET (RMSE(PIXELS) AND PCK@0.2).

Methods	RMSE	Snout	Left ear	Right ear	Tail base	Mean
Baseline	3.83	89.35	90.28	90.28	93.06	90.74
SCM(Conv(3×3)*)*						
SCM(Iteration 1)	4.70	91.20	95.83	95.37	96.30	94.68
SCM(Iteration 2)	4.11	91.08	**97.18**	**96.24**	**98.12**	**95.66**
SCM(Iteration 3)	14.87	91.08	76.53	69.95	88.26	81.46
SCM(Conv(1×1)*)*						
SCM(Iteration 1)	4.09	94.37	93.43	92.96	94.84	93.90
SCM(Iteration 2)	**3.67**	**95.77**	96.24	90.61	**98.12**	95.19
SCM(Iteration 3)	4.52	92.49	90.61	93.90	94.84	92.96
SCM(Conv(1×1)*+Conv*(3×3)*)*						
SCM(Iteration 1)	3.97	95.31	97.65	88.26	95.30	94.13
SCM(Iteration 2)	3.93	95.77	96.71	89.67	95.31	94.37
SCM(Iteration 3)	8.55	91.08	89.67	84.04	81.69	86.62

TABLE S2
ABLATION EXPERIMENTS OF THE CASCADED MULTI-LEVEL SUPERVISION MODULE ON THE DEEPLABCUT MOUSE POSE TEST SET (RMSE(PIXELS) AND PCK@0.2).

Methods	RMSE	Snout	Left ear	Right ear	Tail base	Mean
Baseline	3.83	89.35	90.28	90.28	93.06	90.74
MLS(All Cas1)	7.85	90.28	91.67	95.37	**98.61**	93.98
CMLS(Start Cas2)	**3.52**	**94.84**	**97.65**	**95.77**	97.65	**96.48**
CMLS(Middle Cas2)	5.35	90.14	90.61	88.26	98.59	91.90
CMLS(Final Cas2)	4.51	89.20	93.42	95.77	98.12	94.13
CMLS(All Cas2)	8.01	81.22	93.43	92.02	96.24	90.73

SUPPLEMENTARY B

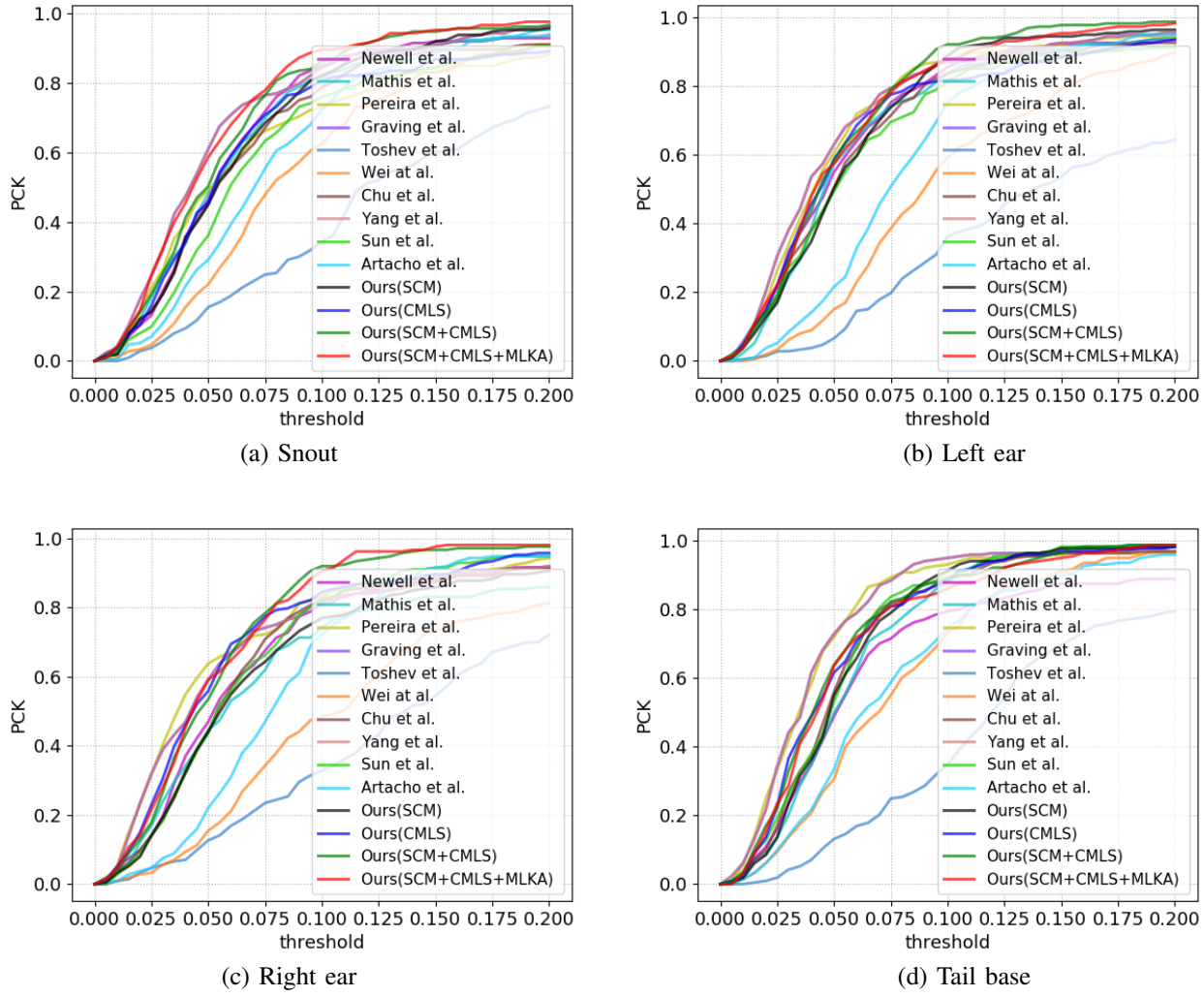


(a) Snout (b) Left ear

(c) Right ear (d) Tail base

Fig. S1. Comparisons of the PCK curves for each part on the PDMB test set.

SUPPLEMENTARY C



(a) Snout (b) Left ear

(c) Right ear (d) Tail base

Fig. S2. Comparisons of the PCK curves for each part on the DeepLabCut Mouse Pose test set.

SUPPLEMENTARY D

TABLE S3
COMPARISONS OF RMSE AND PCK@0.2 SCORE ON THE ZEBRA TEST SET.

Methods	RMSE	Snout	Head	Neck	ForelegL1	ForelegR1	HindlegL1	HindlegR1	Tail base	Tail tip	Mean
Newell et al. [26]	2.09	81.67	85.0	98.33	96.67	92.78	97.22	97.22	97.78	76.67	91.48
Mathis et al. [17]	1.71	81.67	90.0	100.0	97.22	97.22	97.78	98.33	**100.0**	76.67	93.21
Pereira et al. [23]	4.13	72.22	80.56	90.56	91.11	94.44	93.33	92.22	93.33	80.56	87.59
Graving et al. [22]	1.62	83.89	92.78	**100.0**	**97.78**	96.67	**98.89**	**98.89**	**100.0**	77.22	94.01
Toshev et al. [38]	2.35	77.22	65.0	97.78	95.0	91.11	91.67	91.11	98.33	72.78	86.67
Wei at al. [27]	3.00	66.67	71.11	95.0	55.56	65.56	57.22	65.56	95.56	60.0	70.25
Chu et al. [34]	1.87	85.56	85.0	98.33	96.67	95.0	96.67	98.33	98.89	82.22	92.96
Sun et al. [28]	2.91	81.11	90.0	97.78	95.56	93.33	93.89	95.56	95.0	74.44	90.74
Artacho et al. [37]	2.90	71.11	73.89	91.11	58.33	68.33	59.44	62.78	96.11	65.56	71.85
Ours(SCM)	1.85	83.33	93.33	96.67	97.22	97.22	97.22	98.33	98.33	83.89	93.95
Ours(CMLS)	1.94	83.33	86.11	99.44	97.22	96.67	97.22	98.33	98.33	75.56	92.47
Ours(GM-SCENet)	1.60	87.22	93.33	**100.0**	97.22	**97.78**	99.44	97.78	99.44	**86.11**	95.37
Ours(GM-SCENet+MLKA)	**1.57**	**88.33**	**93.89**	99.44	**97.78**	**97.78**	**98.89**	98.33	99.44	85.56	**95.49**

TABLE S4
COMPARISONS OF RMSE AND PCK@0.2 SCORE ON THE FLY TEST SET. WE CHOOSE 8 DIFFICULT KEYPOINTS OF THE FLY, I.E.,MIDLEGR3, HINDLEGR2, HINDLEGR3, HINDLEGR4, MIDLEGL4, HINDLEGL2, HINDLEGL3, HINDLEGL4 FOR COMPARISON.

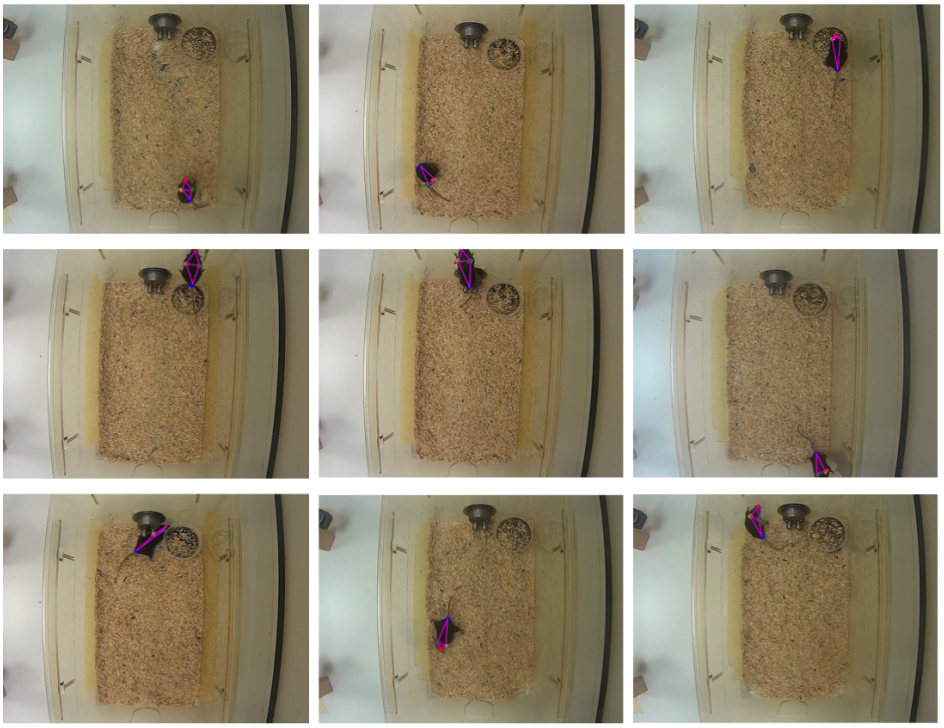
Methods	RMSE	MidlegR3	HindlegR2	HindlegR3	HindlegR4	MidlegL4	HindlegL2	HindlegL3	HindlegL4	Mean
Newell et al. [26]	2.88	96.67	90.0	85.67	87.0	95.0	93.67	89.0	87.67	90.58
Mathis et al. [17]	1.94	96.67	93.0	92.0	83.67	93.67	94.0	92.0	88.0	91.63
Pereira et al. [23]	2.99	94.67	89.67	86.0	82.33	91.67	91.0	89.0	86.33	88.83
Graving et al. [22]	**1.91**	96.67	92.67	92.0	85.67	95.33	94.0	92.33	87.0	91.96
Toshev et al. [38]	3.26	90.33	87.33	85.0	59.0	89.33	89.33	84.33	72.0	82.08
Wei at al. [27]	3.03	93.67	89.33	88.67	80.67	94.0	89.0	88.0	83.33	88.33
Chu et al. [34]	2.59	95.67	90.0	90.33	88.33	91.67	90.33	87.0	91.0	90.54
Sun et al. [28]	7.43	55.0	91.67	93.0	82.0	94.0	92.67	92.33	87.33	86.0
Artacho et al. [37]	1.99	98.0	94.0	88.67	82.33	95.67	94.67	86.67	87.67	90.96
Ours(SCM)	2.40	98.33	92.67	91.67	88.33	95.0	92.67	90.33	91.33	92.54
Ours(CMLS)	3.17	96.67	91.0	88.67	87.0	92.67	94.0	90.67	88.0	91.08
Ours(GM-SCENet)	2.56	98.33	**94.0**	**94.0**	**90.67**	**96.33**	**96.33**	**94.0**	94.33	**94.75**
Ours(GM-SCENet+MLKA)	2.51	**98.67**	**94.0**	93.33	**90.67**	95.67	**96.33**	**94.0**	**94.67**	94.67

TABLE S5
COMPARISONS OF PCK@0.2 SCORE ON THE LSP TEST SET.

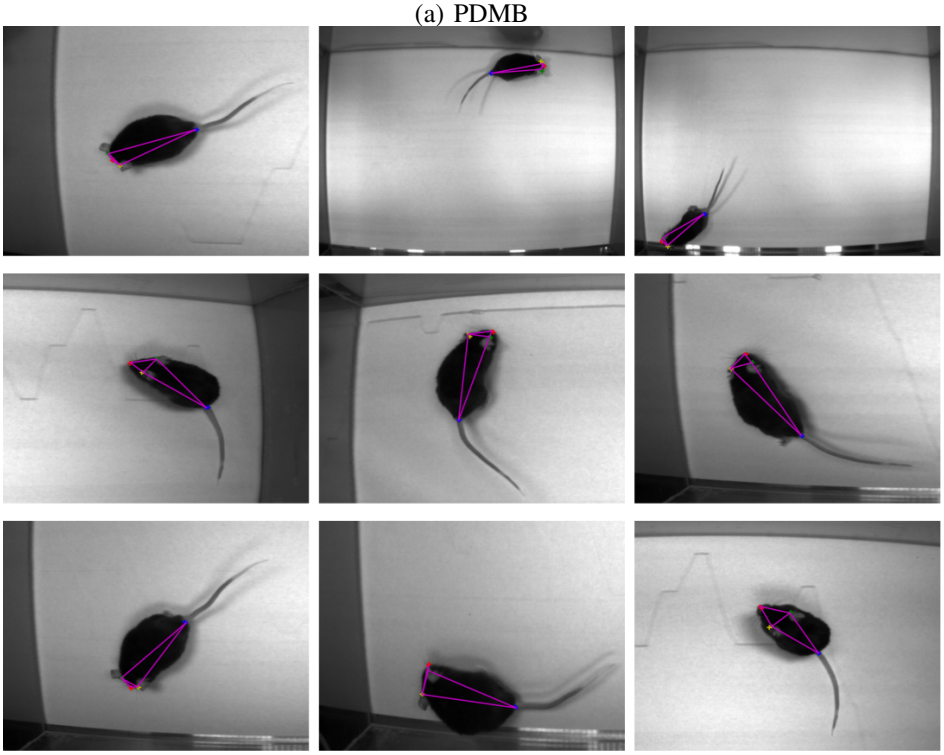
Methods	Head	Sho.	Elb.	Wri.	Hip	Knee	Ank.	Mean
Lifshitz et al.*	96.8	89.0	82.7	79.1	90.9	86.0	82.5	86.7
Wei at al. [27]	97.8	92.5	87.0	83.9	91.5	90.8	89.9	90.5
Bulat at al.*	97.2	92.1	88.1	85.2	92.2	91.4	88.7	90.7
Chu et al. [34]	98.1	93.7	89.3	86.9	93.4	94.0	92.5	92.6
Shu et al.*	97.9	93.6	89.0	85.8	92.9	91.2	90.5	91.6
Yang et al.*	98.3	94.5	92.2	88.9	94.4	95.0	93.7	93.9
Zhang et al.*	98.4	94.8	92.0	89.4	94.4	94.8	93.8	94.0
Cao et al.*	**98.6**	95.1	92.1	89.8	94.7	94.9	93.9	94.2
Zhang et al.*	97.3	92.3	86.8	84.2	91.9	92.2	90.9	90.8
Artacho et al. [37]	–	–	–	–	–	–	–	94.5
Tang et al.*	98.3	95.9	93.5	90.7	95.0	96.6	95.7	95.1
Xiao et al.*	98.3	96.3	**96.2**	**95.1**	96.0	**96.7**	**95.9**	**96.4**
Ours(GM-SCENet)	93.1	**96.6**	96.1	94.6	**96.4**	95.9	92.7	95.1

* The corresponding references are on the Supplementary H. For data preprocessing, we follow the default setting in [37] where we simply use the original image size as the rough scale, and the image center as the rough position of the target human to crop the image patches.

SUPPLEMENTARY E



(a) PDMB



(b) DeepLabCut Mouse Pose

Fig. S3. Examples of the estimated mouse poses on the PDMB and DeepLabCut test sets (best viewed in electronic form with 4× zoom in). Our proposed method deals well with deformable mouse body on the PDMB dataset. The first, second and third row in (a) show occlusions, invisible keypoints and abnormal poses, respectively. On the DeeplabCut Mouse Pose dataset, our method is robust against scale variations.

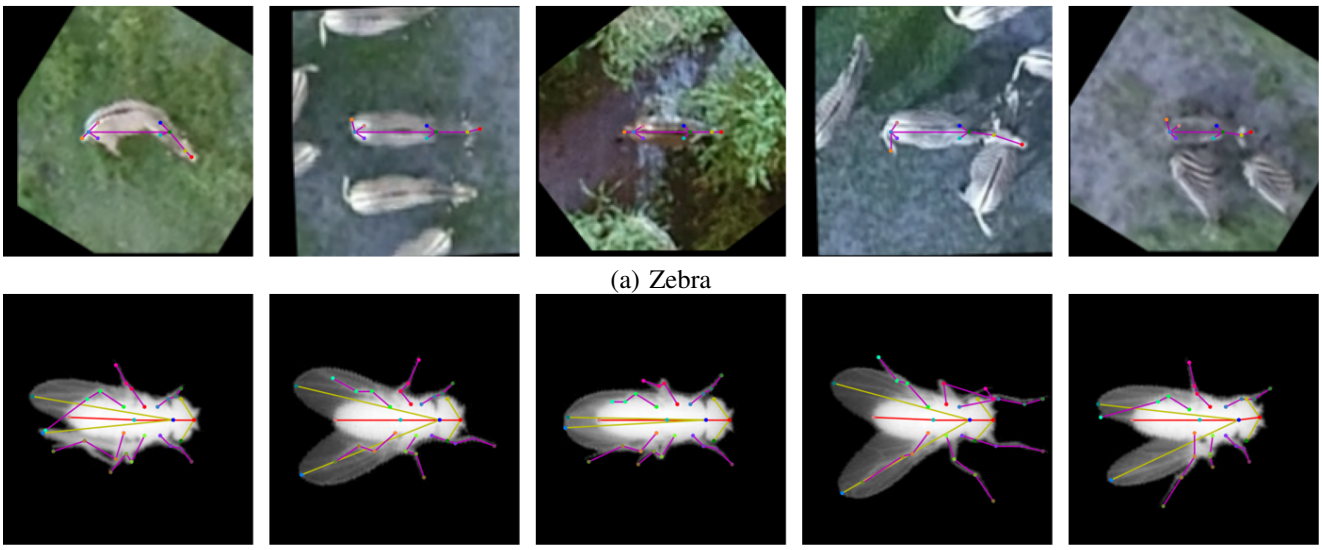
SUPPLEMENTARY F

TABLE S6
ZEBRA, FLY AND LSP DATASETS USED FOR MODEL COMPARISONS.

Name	Species	Resolution	Images	Keypoints	Individuals
Zebra [22]	Equus grevyi	160*160	900	9	Multiple
Fly [23]	Drosophila melanogaster	192*192	1500	32	Single
LSP [48]	Human	–	12000	14	Multiple

The herds of zebras are recorded in the wild. This dataset features multiple interacting individuals with highly-variable environments and lighting conditions. When multiple zebras occurs in an image, the owner of the dataset only provides the grount-truth annotation (i.e., the locations of 9 keypoints) of one of these zebras, which lies in the center of the image, as shown in Fig. S4(a). In our experiments, we randomly split the Zebra dataset into a training set of 720 images and a test set of 180 images. Fly dataset is recorded in a laboratory setting. Flies move freely in a backlit 100-mm-diameter circular arena covered by a 2-mm-tall clear polyethylene terephthalate glycol dome. This dataset is divided into training set (1200 images) and testing set (300 images). These datasets are freely-available from https://github.com/jgraving/deepposekit-data.

The Leeds Sports Pose (LSP) dataset is leveraged for single person pose estimation. Images for LSP are collected from Flickr for a wide range of individuals performing sports activities. The dataset includes 11000 images for training and 1000 images for testing where 14 keypoints in the entire body are labeled.



(a) Zebra

(b) Fly

Fig. S4. Examples of the estimated poses on the Zebra and Fly datasets. The proposed method can achieve accurate localisation on the both datasets.

SUPPLEMENTARY G

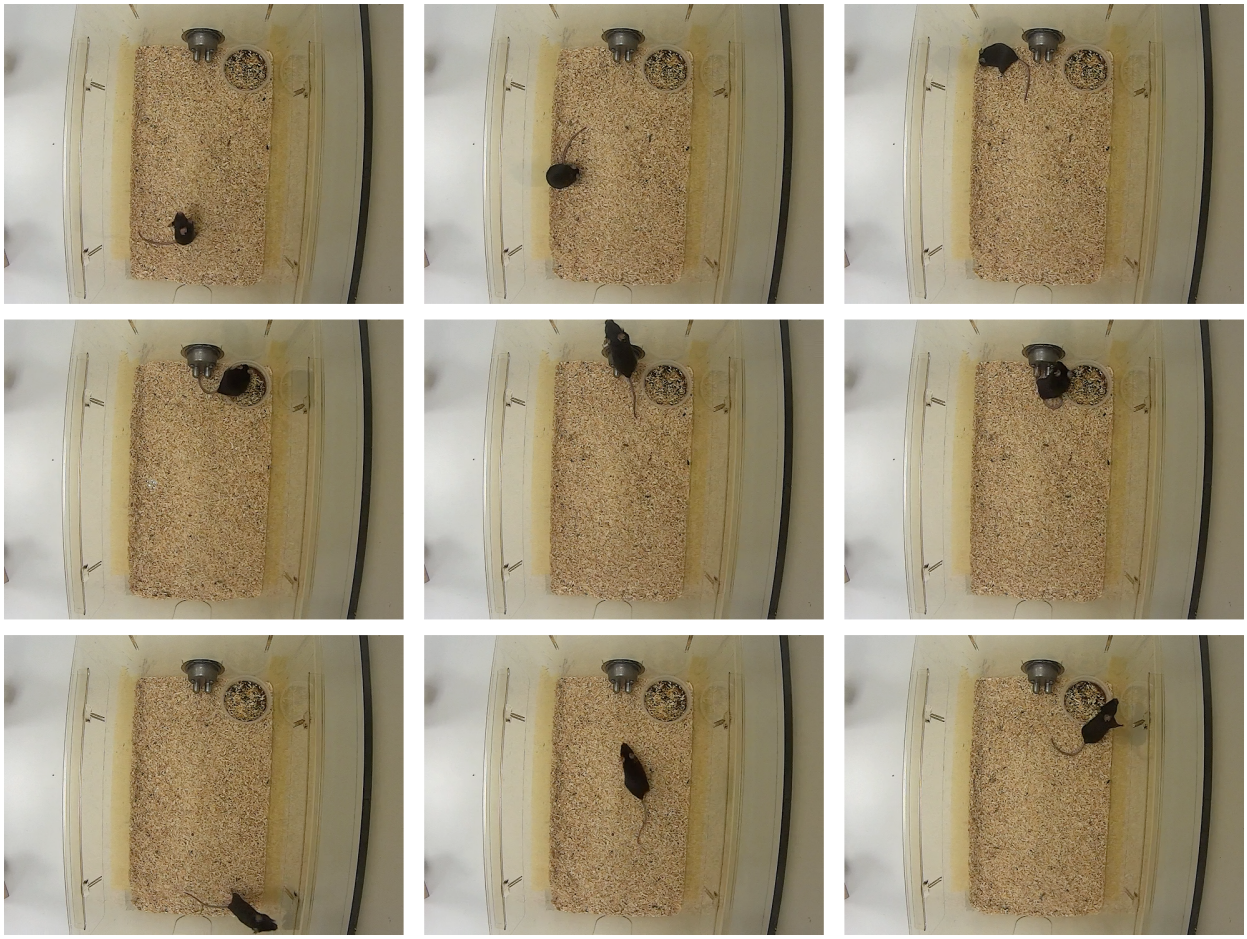


Fig. S5. Examples of our PDMB dataset.

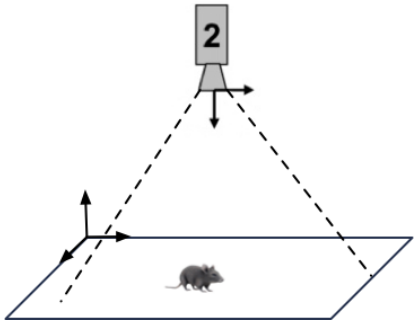


Fig. S6. The location of the camera used in our mouse pose PDMB dataset. The camera is about 50cm from the bottom of the cage. The location of the camera in all experiments is fixed.

Pose annotation. All extracted video frames were annotated using a freeware DeepLabCut (available at https://github.com/DeepLabCut/DeepLabCut). A team of six professionals were trained to annotate home-cage mouse poses. Following the experimental setting of DeepLabcut [17], we also annotated the locations of four body parts, i.e., snout, left ear, right ear and tail base. Additionally, we annotated the visibility of mouse parts, as shown in Fig S7. After data annotation, we performed secondary screening to remove ambiguous frames, leaving 9248 frames to establish our mouse pose PDMB dataset.

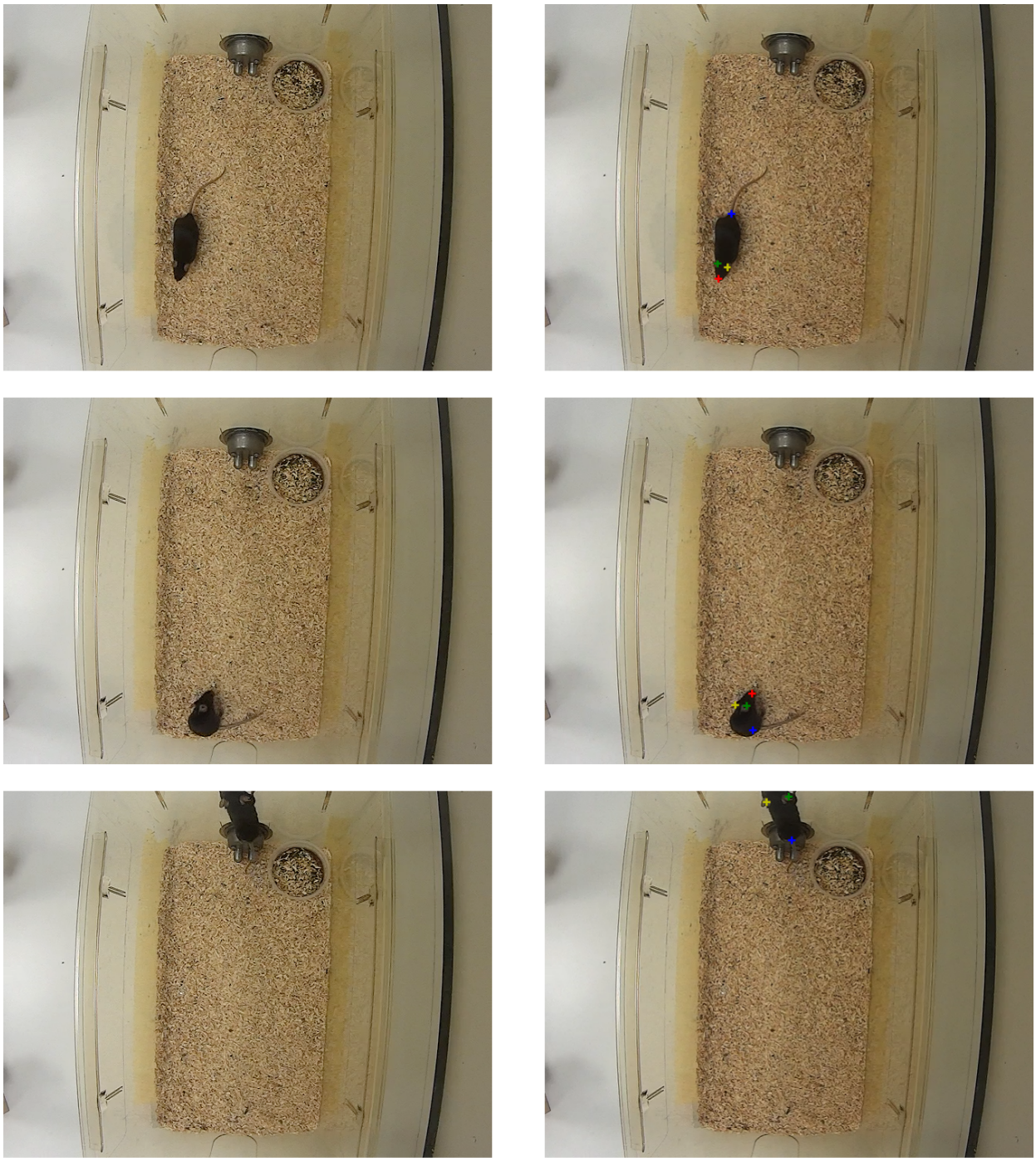


Fig. S7. Snapshots taken from top-view camera for different mouse poses. (a) and (b) show the original images and manual annotation results. In the first row of (b), all keypoints are visible, and the flag of the visibility of mouse parts is set to 1. The second row of (b) shows the case where a part, i.e., tail base, may be occluded. The tail base is not visible but its position is apparent given the context of the image. Our dataset provides ground truth locations for these keypoints and an additional annotation indicates their lack of visibility. The last row of (b) shows that the snout is truncated. In such case, Our dataset will not provide a ground-truth annotation.

SUPPLEMENTARY H

Reference

[1] I. Lifshitz, E. Fetaya, and S. Ullman, "Human pose estimation using deep consensus voting," in European Conference on Computer Vision. Springer, 2016, pp. 246–260.

[2] A. Bulat and G. Tzimiropoulos, "Human pose estimation via convolutional part heatmap regression," in European Conference on Computer Vision. Springer, 2016, pp. 717–732.

[3] K. Sun, C. Lan, J. Xing, W. Zeng, D. Liu, and J. Wang, "Human pose estimation using global and local normalization," in Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 5599– 5607.

[4] W. Yang, S. Li, W. Ouyang, H. Li, and X. Wang, "Learning feature pyramids for human pose estimation," in proceedings of the IEEE international conference on computer vision, 2017, pp. 1281–1290.

[5] H. Zhang, H. Ouyang, S. Liu, X. Qi, X. Shen, R. Yang, and J. Jia, "Human pose estimation with spatial contextual information," arXiv preprint arXiv:1901.01760, 2019.

[6] Z. Cao, R. Wang, X. Wang, Z. Liu, and X. Zhu, "Improving human pose estimation with self-attention generative adversarial networks," in 2019 IEEE International Conference on Multimedia & Expo Workshops (ICMEW). IEEE, 2019, pp. 567–572.

[7] F. Zhang, X. Zhu, and M. Ye, "Fast human pose estimation," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 3517–3526.

[8] W. Tang, P. Yu, and Y. Wu, "Deeply learned compositional models for human pose estimation," in Proceedings of the European conference on computer vision (ECCV), 2018, pp. 190–206.

[9] Y. Xiao, D. Yu, X. Wang, T. Lv, Y. Fan, and L. Wu, "Spcnet: Spatial preserve and content-aware network for human pose estimation," arXiv preprint arXiv:2004.05834, 2020.