3D-MuPPET: 3D Multi-Pigeon Pose Estimation and Tracking

Supplemental Material

Urs Waldmann[1,2*†], Alex Hoi Hang Chan[2,3,4*†], Hemal Naik[2,4,5], Máté Nagy[2,3,4,6,7], Iain D. Couzin[2,3,4], Oliver Deussen[1,2], Bastian Goldluecke[1,2], Fumihiro Kano[2,3,4]

[1]Department of Computer and Information Science, University of Konstanz, Germany.
[2]Centre for the Advanced Study of Collective Behaviour, University of Konstanz, Germany.
[3]Department of Collective Behavior, Max Planck Institute of Animal Behavior, Konstanz, Germany.
[4]Department of Biology, University of Konstanz, Germany.
[5]Department of Ecology of Animal Societies, Max Planck Institute of Animal Behavior, Konstanz, Germany.
[6]Department of Biological Physics, Eötvös Loránd University, Budapest, Hungary.
[7]MTA-ELTE 'Lendület' Collective Behaviour Research Group, Hungarian Academy of Sciences, Budapest, Hungary.

*Corresponding author(s). E-mail(s): urs.waldmann@uni-konstanz.de; hoi-hang.chan@uni-konstanz.de;
Contributing authors: hnaik@ab.mpg.de; nagymate@hal.elte.hu; icouzin@ab.mpg.de; Oliver.Deussen@uni-konstanz.de; bastian.goldluecke@uni-konstanz.de; fumihiro.kano@uni-konstanz.de;
†These authors contributed equally to this work.

Abstract

In the supplemental material, we first provide more details on the novel Wild-MuPPET dataset in Sec. 1, then provide more details on the dynamic matching algorithm. Next, we report detailed results on our network training and ablation studies in Sec. 3. Also, we briefly explain the metrics used in our main paper in Sec. 4. Finally, in Sec. 5 we report pose estimation results on the odor trail tracking dataset from Mathis et al. (2018) while in Sec. 6 we report results on the cowbird data from Badger et al. (2020).

1 Wild-MuPPET Dataset

Here, we provide more description on the novel Wild-MuPPET dataset.

Experimental setup. The dataset is collected in a private pigeon breeder in Singen, Germany, where pigeons are released everyday to

freely fly around the area. Similar to 3D-POP (Naik et al., 2023), we use 4 sony action cameras (rx0-ii, 30Hz, 3840 × 2160 px) cameras connected to a camera control box (CCB-WD1) to allow for synchronization. The cameras are mounted on 4 tripods to create a rectangular formation. Data was then collected opportunistically when pigeons land on the ground to return to their loft.

Calibration Procedure. For intrinsic calibration, we use a standard Charuco checkerboard. For extrinsics calibration, we used a Vicon IR active wand, which is a calibration wand made for a Vicon motion capture system, containing 5 unique RGB red light in a fixed position. To determine the exact 3D object coordinate of each of the 5 lights, we use SMART-BARN (Nagy et al., 2023), a large scaled motion capture facility, where the exact coordinates of the wand are measured. During calibration in Wild-MuPPET, we present the wand to all 4 cameras for a short calibration sequence, and the 5 unique points are detected each frame using a simple blob detection algorithm. All detections are finally combined with the 3D object definition for extrinsic calibration of the all 4 cameras.

Dataset Description. The Wild-MuPPET dataset contains a total of 2000 frames of manual 2D annotations of 9 keypoints, identical to 3D-POP (beak, nose, left/right eye, left/right shoulder, top/bottom keel, tail). These 2D detections are then triangulated into 3D using triangulation with bundle adjustment, to obtain 500 frames of 3D ground truth. All annotations contain a single pigeon individual.

For additional qualitative evaluation, we also provide 3600 frames of continuous sequence with a single pigeon (same sequence where annotated set was sampled from) in the tracking area and 900 frames of continuous sequence with 3 pigeons in the tracking area.

2 Details on the Dynamic Matching Algorithm

The dynamic matching algorithm based on Huang et al. (2020) first generates 3D pose estimates for each possible pair of 2D poses, creating a large 3D pose subspace. This 3D pose subspace contains only a small amount of correct 3D poses. We then pick the correct poses in the 3D subspace by assuming that the correct 3D poses are calculated from 2D poses belonging to the same individual. Thus, if the Euclidean distance between a pair of 3D poses from the 3D subspace is sufficiently small, we consider the 2D poses that belong to these two 3D poses a match (Huang et al., 2020). We match until the pairwise distance threshold of 200mm is reached. Since the algorithm does not know the number of individuals in the scene, we choose a conservative threshold of 200mm to ensure all individuals are matched. Note that the algorithm prioritizes matches with lower distance,

hence a larger threshold doesn't lead to worse performance while a lower threshold could lead to individuals that are not matched. For more details we refer to Huang et al. (2020).

3 Results on Network Training and Ablation Studies

In this section of our supplemental material, we give more detailed results on experiments on network training and ablation studies.

3.1 Data Augmentation for Pigeons

For data augmentation for the KeypointR-CNN (He, Gkioxari, Dollar, & Girshick, 2017) we find that changing brightness, flipping or scaling do not enhance performance, but changing sharpness with a probability of 0.2 results in the best performance in terms of RMSE (for numbers cf. Tabs. 1 and 2). This is intuitive since we train on the single pigeon data where the training data already contains a wide range of different pigeon positions and lighting conditions and thus covers most of the scaling and brightness. Also, the training data already include most body orientations (with respect to the camera), thus flipping does not improve test accuracy. Since the depth of field of the cameras is limited the pigeons are sometimes slightly out of focus and therefore blurring the input image with a small probability of 0.2 improves the accuracy of the test set.

In the case of multi-pigeon video sequences, however, we find that the best data augmentation parameters are not the same as for the single pigeon data. We keep the parameters from the single pigeon analysis but find that randomly jittering brightness by a factor chosen uniformly from $[0.4, 1.6]$ and a flipping probability of 0.5 produces the best outcome. This is intuitive because the single pigeon data does not cover the range of brightness found in the multi-pigeon data and the flipping makes the pose estimation in new situations more robust. A small scaling range of $\pm 5\%$ is sufficient since the single pigeon data covers already a large range of pigeon sizes. Also, if the scaling range is too large, we find multiple (mis-)detections if pigeons are nearby. This is also the case in situations where the pigeons occlude or are close to each other even if we do not apply scaling.

3.2 Training Hyperparameters

In Tab. 3 you find detailed results on experiments for hyperparameter tuning for the KeypointR-CNN (He et al., 2017). A step size of 50 and a multiplicative factor of learning rate decay $\gamma = 0.5$ yield the best result (cf. Tab. 3).

Table 1 *Ablation Study.* Data augmentation ablation study (single pigeon data) for the parameters brightness (b) and sharpness probability (sp). Framework trained on whole session four (s4) with batch size 20, learning rate 0.005, step size 10, gamma 0.5, number of epochs 100, no flipping and no scaling. Results are given as RMSE [px] for predictions where confidence score exceeds 0.999. s1, s2 and s3 denote the different recording sessions. *: No change in brightness.

config	s1	s2	s3
b = $[1, 1]^*$, sp = 0	25.1	6.4	9.7
b = $[0.7, 1.3]$, sp = 0.1	14.3	4.4	6.9
b = $[0.4, 1.6]$, sp = 0.1	12.7	4.5	6.6
b = $[0.7, 1.3]$, sp = 0.2	13.0	4.6	6.8
b = $[0.4, 1.6]$, sp = 0.2	13.3	4.6	6.9
b = $[0.4, 1.6]$, sp = 0	13.5	4.7	7.1
b= $[1, 1]^*$, sp= 0.2	17.0	**3.9**	6.7

Table 2 *Ablation Study.* Data augmentation ablation study (single pigeon data) for the parameters flip probability (fp) and scale range (sr). Framework trained on whole session four (s4) with batch size 40, learning rate 0.005, step size 77, gamma 0.7, number of epochs 500, brightness 0.6 and sharpness probability 0.2. Results are given as RMSE [px] for predictions where confidence score exceeds 0.999. s1, s2 and s3 denote the different recording sessions. No significant improvement within sessions.

config	s1	s2	s3
fp = 0, sr = $[50, 200]$	14.3	4.7	7.5
fp = 0.5, sr = $[75, 150]$	12.3	4.6	7.0
fp = 0.5, sr = $[90, 110]$	11.9	4.6	7.0
fp = 0.5, sr = $[78, 125]$	11.8	4.7	6.8

4 Metrics

In this section of our supplemental material, we briefly explain the metrics used in our main paper.

4.1 Pose Estimation

The RMSE is the L2 distance between the predicted and ground truth positions of keypoints.

Table 3 *Ablation Study.* Hyperparameter ablation study related to training for the parameters step size (sz) and γ. Framework trained on entire session four (s4) of the single pigeon data with batch size 20, learning rate 0.005, number of epochs 250, no change in brightness, sharpness probability 0.2, no flipping and no scaling. Results evaluated on 200 randomly sampled frames from session two (s2) for predictions where confidence score exceeds 0.999.

config	RMSE [px]
sz = 10, γ = 0.5	5.5
sz = 25, γ = 0.5	4.6
sz= 50, γ = 0.5	**3.8**
sz = 75, γ = 0.5	4.3
sz = 25, γ = 0.7	4.5
sz = 50, γ = 0.7	4.3
sz = 75, γ = 0.7	4.4
sz = 25, γ = 0.95	4.6
sz = 50, γ = 0.95	4.7
sz = 75, γ = 0.95	4.4

We average over samples and keypoints like Mathis et al. (2018).

The PCK is the percentage of predicted keypoints that fall within a normalized distance of the ground truth. This normalized distance in 3D Bird Reconstruction (Badger et al., 2020) is a fraction (0.05 and 0.1) of the largest dimension of the ground truth bounding box containing the bird and so do we use this, too, in our comparison on the cowbird data. For our comparison on the pigeon data instead, the normalized distance is again a fraction (0.05 and 0.1) of the largest dimension of the ground truth bounding box for the 2D evaluation and the maximum distance between any two ground truth keypoints for each individual in 3D.

4.2 Tracking

The CLEAR-MOT metrics are the Multi Object Tracking Accuracy (MOTA) and the Multi Object Tracking Precision (MOTP). MOTP is the total error in estimated position for matched object-hypothesis pairs over all frames, averaged by the total number of matches made (Bernardin & Stiefelhagen, 2008). MOTA summarizes three sources of errors with a single performance measure, i.e. the ratio of misses in the sequence, computed over the total number of objects present in all frames, the ratio of false positives and the ratio of mismatches (Bernardin & Stiefelhagen, 2008; Dendorfer et al., 2021). The track quality measures are classified as Recall, Precision, false positives per frame (FPF) mostly tracked (MT), partially tracked (PT) mostly lost (ML), fragments (Frag) and ID switches (IDS). Recall and Precision are the frame-based correctly matched objects divided by total ground truth objects and total output objects respectively (Li, Huang, & Nevatia, 2009). MT and ML are the percentage of ground truth trajectories which are covered by tracker output for more than 80% and less than 20% in length (Li et al., 2009). Frag is the number of fragmentations where a track is interrupted by miss detection (Bewley, Ge, Ott, Ramos, & Upcroft, 2016). The trajectory-based metric IDF1 is the ratio of correctly identified detections over the average number of ground-truth and computed detections (Ristani, Solera, Zou, Cucchiara, & Tomasi, 2016).

5 2D Mouse Pose Estimation

5.1 Odor Trail Tracking Data

This 2D data from Mathis et al. (2018) contains single mice following an odor and contains 1080 manually annotated samples. The samples are random, distinct frames from multiple sessions observing seven different mice (Mathis et al., 2018) and the resolution of the images is 640×480 or 800×800 since the data was recorded with two different monochromatic cameras. On average the mice cover an area of 256×256 pixels of the frame.

For more details on this dataset we refer to Mathis et al. (2018).

5.2 Comparison on the Odor Trail Tracking Data (Mouse)

In the original DLC article (Mathis et al., 2018) the authors evaluate and report numbers in terms of RMSE on their odor trail tracking data where they estimate the pose (snout, left and right ear and tail base) of single mice. We thus report only RMSE in this section for the purpose of comparison. In the DLC article, the networks are trained for a total of 650K iterations with batch size 1 for three splits of 0.8/0.2 (training/test) and evaluated every 50K iterations. The authors also report

Table 4 *Comparison on the Odor Trail Tracking Data (Mouse).* RMSE on the odor trail tracking test set from Mathis et al. (2018). Values for DLC from Mathis et al. (2018). Values of KP-RCNN is from our analysis, coloured in grey. We report precision within ±0.2 because we read values from Fig. 2c in Mathis et al. (2018).

Model, iterations	RMSE [px]
KP-RCNN, 200K iterations	4.2
DLC, 200K iterations	3.6 ± 0.2
DLC, 350K/600K iterations	$\mathbf{3.2 \pm 0.2}$

the average of the three splits. For more details see Mathis et al. (2018).

In order to compare the KeypointRCNN to DeepLabCut (which is our other choice for the pose estimation module of our framework) on the mouse data, we train their odor trail tracking data set with the KeypointRCNN. We train the KeypointRCNN on the DeepLabCut data with the configuration that we report in the main paper. We train for 250 epochs with a batch size of 20 instead of 1 to exploit our hardware and fine-tune twice for another 250 epochs with training configurations that lower the learning rate further to compare our results to those of DeepLabCut after 200K, 400K and 600K iterations.

Tab. 4 compares results for DeepLabCut from Mathis et al. (2018) with the KeypointRCNN. We obtain the results for DeepLabCut from Fig. 2c in Mathis et al. (2018). These results were achieved with a network based on ResNet-50. We report their values for 200K iterations and their absolute lowest RMSE on the test set averaged over the three 0.8/0.2 splits. For the KeypointRCNN we report numbers with the same precision as we are able to read for DeepLabCut. We report numbers only for 200K iterations because the KeypointRCNN does not improve the accuracy of pose estimation in the test set when trained for more iterations.

Overall, this comparison shows the same trend as the 2D pigeon results in our main paper. Please note that DLC in Mathis et al. (2018) in contrast to the KeypointRCNN is optimized on the odor trail tracking data. Thus we conclude that the KeypointRCNN is comparable with DeepLabCut in terms of RMSE on the mouse data meaning that the KeypointRCNN also achieves a RMSE of about 4 px on the odor trail tracking test set.

6 2D Cowbird Pose Estimation

6.1 Cowbird Data

This 2D data from Badger et al. (2020) contains single cowbirds. Their original images have a maximum resolution of 1920×1200 containing multiple birds. For 2D pose estimation they use 1000 cropped samples of single individuals from a subset of 18 moments across 6 of the 10 days (Badger et al., 2020) with a resolution of 256×256.

Table 5 *Ablation Study.* Data augmentation ablation study (cowbird data from Badger et al. (2020)) for the parameters brightness (b), sharpness probability (sp), contrast (c), saturation (s) and hue (h). Framework trained on their training split with batch size 20, learning rate 0.005, step size 9, gamma 0.5, number of epochs 45, no flipping and no scaling. Results are given as PCK and evaluated on their test split. *: No change in brightness.

config	@0.05	@0.1
b = $[1, 1]^*$, sp = 0, c = 0, s = 0, h = 0	0.37	0.55
b = $[1, 1]^*$, sp = 0.1, c = 0, s = 0, h = 0	0.35	0.54
b = $[1, 1]^*$, sp = 0.2, c = 0, s = 0, h = 0	0.38	0.55
b= [0.7, 1.3], sp= 0.1, c= 0, s= 0, h= 0	**0.39**	**0.56**
b = [0.4, 1.6], sp = 0.2, c = 0, s = 0, h = 0	0.36	0.52
b = [0.7, 1.3], sp = 0, c = 0, s = 0, h = 0	0.37	0.56
b = [0.4, 1.6], sp = 0, c = 0, s = 0, h = 0	0.37	0.55
b = [0.7, 1.3], sp = 0.1, c = 0.2, s = 0, h = 0	0.38	0.55
b = [0.7, 1.3], sp = 0.1, c = 0.4, s = 0, h = 0	0.37	0.53
b = [0.7, 1.3], sp = 0.1, c = 0.6, s = 0, h = 0	0.37	0.55
b = [0.7, 1.3], sp = 0.1, c = 0.8, s = 0, h = 0	0.38	0.55
b = [0.7, 1.3], sp = 0.1, c = 0, s = 0.2, h = 0	0.38	0.54
b = [0.7, 1.3], sp = 0.1, c = 0, s = 0.4, h = 0	0.37	0.55
b = [0.7, 1.3], sp = 0.1, c = 0, s = 0.6, h = 0	0.38	0.54
b = [0.7, 1.3], sp = 0.1, c = 0, s = 0.8, h = 0	0.37	0.56
b = [0.7, 1.3], sp = 0.1, c = 0, s = 0, h = 0.1	0.38	0.55
b = [0.7, 1.3], sp = 0.1, c = 0, s = 0, h = 0.2	0.38	0.56
b = [0.7, 1.3], sp = 0.1, c = 0, s = 0, h = 0.3	0.37	0.56
b = [0.7, 1.3], sp = 0.1, c = 0, s = 0, h = 0.4	0.37	0.53
b = [0.7, 1.3], sp = 0.1, c = 0.2, s = 0.8, h = 0.2	0.38	0.55
b = [0.7, 1.3], sp = 0.1, c = 0.8, s = 0.8, h = 0.2	0.37	0.55

For more details on this dataset we refer to Badger et al. (2020).

6.2 Data Augmentation for Cowbirds

The cowbird data set is recorded in outdoor aviaries (Badger et al., 2020). Thus different daylight and season conditions are present. To consider these different conditions inherent in the data, we use different data augmentation parameters. We find that randomly changing brightness by a factor chosen uniformly from $[0.7, 1.3]$, and a sharpness probability of 0.1, works best (for numbers cf. Tab. 5).

6.3 Comparison on the Cowbird Data

3D Bird Reconstruction (Badger et al., 2020) is state of the art for 3D bird shape recovery, and they also report on the accuracy of 2D bird pose estimation. The authors evaluate and report numbers in terms of PCK (cf. Sec. 4) on their cowbird data, where they estimate the pose (bill tip, right and left eyes, neck, nape, right and left wrists, right and left wing tips, right and left feet and the tail tip) of single cowbirds. Their network is trained for 60 epochs (personal e-mail communication with the authors) with a train/test split

Table 6 *Comparison on the Cowbird Data.* PCK on the cowbird test set from Badger et al. (2020). Values for 3DBR from Badger et al. (2020).

Model, epochs	@0.05	@0.1
KP-RCNN, 45 epochs	0.39	0.56
KP-RCNN, 60 epochs	0.36	0.54
3DBR, 60 epochs	**0.46**	**0.64**

of 0.75/0.25. For more details see Badger et al. (2020). In order to compare the KeypointRCNN (one of our choices for the pose estimation module in our framework) to the modified HRNet (Sun, Xiao, Liu, & Wang, 2019; Badger et al., 2020) used in 3D Bird Reconstruction on the cowbird data, we train their single cowbird data with the KeypointRCNN. We train the KeypointRCNN on the cowbird data with the configuration that we report in our main paper. We train for 60 epochs with a batch size of 20 to compare our results to those of 3D Bird Reconstruction. The KeypointRCNN achieves the best performance on the cowbird data after 45 epochs. We thus report the PCK results derived from KeypointRCNN with 45 and 60 epochs.

Tab. 6 compares results for 3D Bird Reconstruction from Badger et al. (2020) with the KeypointRCNN. While the KeypointRCNN achieves lower accuracy by 7% (PCK@0.05) and 8% (PCK@0.1) on the cowbird data set than 3D Bird reconstruction, the KeypointRCNN converges faster (45 epochs vs. 60 epochs).

References

Badger, M., Wang, Y., Modh, A., Perkes, A., Kolotouros, N., Pfrommer, B.G., ... Daniilidis, K. (2020). 3d bird reconstruction: A dataset, model, and shape recovery from a single view. *Eur. conf. comput. vis.* (pp. 1–17).

Bernardin, K., & Stiefelhagen, R. (2008). Evaluating multiple object tracking performance: the clear mot metrics. *EURASIP Journal on Image and Video Processing*, *2008*, 1–10,

Bewley, A., Ge, Z., Ott, L., Ramos, F., Upcroft, B. (2016). Simple online and realtime tracking. *Ieee int. conf. image process.* (p. 3464-3468).

Dendorfer, P., Osep, A., Milan, A., Schindler, K., Cremers, D., Reid, I., ... Leal-Taixé, L. (2021). Motchallenge: A benchmark for single-camera multiple target tracking. *Int. J. Comput. Vis.*, *129*(4), 845–881,

He, K., Gkioxari, G., Dollar, P., Girshick, R. (2017). Mask r-cnn. *Int. conf. comput. vis.*

Huang, C., Jiang, S., Li, Y., Zhang, Z., Traish, J., Deng, C., ... Da Xu, R.Y. (2020). End-to-end dynamic matching network for multi-view multi-person 3d pose estimation. *Computer vision–eccv 2020: 16th european conference, glasgow, uk, august 23–28, 2020, proceedings, part xxviii 16* (pp. 477–493).

Li, Y., Huang, C., Nevatia, R. (2009). Learning to associate: Hybridboosted multi-target tracker for crowded scene. *Ieee conf. comput. vis. pattern recog.* (p. 2953-2960).

Mathis, A., Mamidanna, P., Cury, K.M., Abe, T., Murthy, V.N., Mathis, M.W., Bethge, M. (2018). Deeplabcut: markerless pose estimation of user-defined body parts with deep learning. *Nat. Neurosci.*, *21*, 1281–1289,

Nagy, M., Naik, H., Fumihiro, K., Nora, C.V., Koblitz, J.C., Wikelski, M., Couzin, I.D. (2023). Smart-barn: Scalable multimodal arena for real-time tracking behavior of animals in large numbers. *(in press) Science Advances*, ,

Naik, H., Chan, A.H.H., Yang, J., Delacoux, M., Couzin, I.D., Kano, F., Nagy, M. (2023, June). 3d-pop - an automated annotation approach to facilitate markerless 2d-3d tracking of freely moving birds with marker-based motion capture. *Ieee conf. comput. vis. pattern recog.* (p. 21274-21284).

Ristani, E., Solera, F., Zou, R., Cucchiara, R., Tomasi, C. (2016). Performance measures and a data set for multi-target, multi-camera tracking. *Eur. conf. comput. vis.* (pp. 17–35).

Sun, K., Xiao, B., Liu, D., Wang, J. (2019). Deep high-resolution representation learning for

human pose estimation. *Ieee conf. comput. vis. pattern recog.*