

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

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

<sup>1</sup>Department of Computer and Information Science, University of Konstanz, Germany.

<sup>2</sup>Centre for the Advanced Study of Collective Behaviour, University of Konstanz, Germany.

<sup>3</sup>Department of Collective Behavior, Max Planck Institute of Animal Behavior, Konstanz, Germany.

<sup>4</sup>Department of Biology, University of Konstanz, Germany.

<sup>5</sup>Department of Ecology of Animal Societies, Max Planck Institute of Animal Behavior, Konstanz, Germany.

<sup>6</sup>Department of Biological Physics, Eötvös Loránd University, Budapest, Hungary.

<sup>7</sup>MTA-ELTE ‘Lendület’ Collective Behaviour Research Group, Hungarian Academy of Sciences, Budapest, Hungary.

\*Corresponding author(s). E-mail(s): [urs.waldmann@uni-konstanz.de](mailto:urs.waldmann@uni-konstanz.de);  
[hoi-hang.chan@uni-konstanz.de](mailto:hoi-hang.chan@uni-konstanz.de);

Contributing authors: [hnaik@ab.mpg.de](mailto:hnaik@ab.mpg.de); [nagymate@hal.elte.hu](mailto:nagymate@hal.elte.hu); [icouzin@ab.mpg.de](mailto:icouzin@ab.mpg.de);  
[Oliver.Deussen@uni-konstanz.de](mailto:Oliver.Deussen@uni-konstanz.de); [bastian.goldluecke@uni-konstanz.de](mailto:bastian.goldluecke@uni-konstanz.de);  
[fumihiro.kano@uni-konstanz.de](mailto:fumihiro.kano@uni-konstanz.de);

<sup>†</sup>These authors contributed equally to this work.

## Abstract

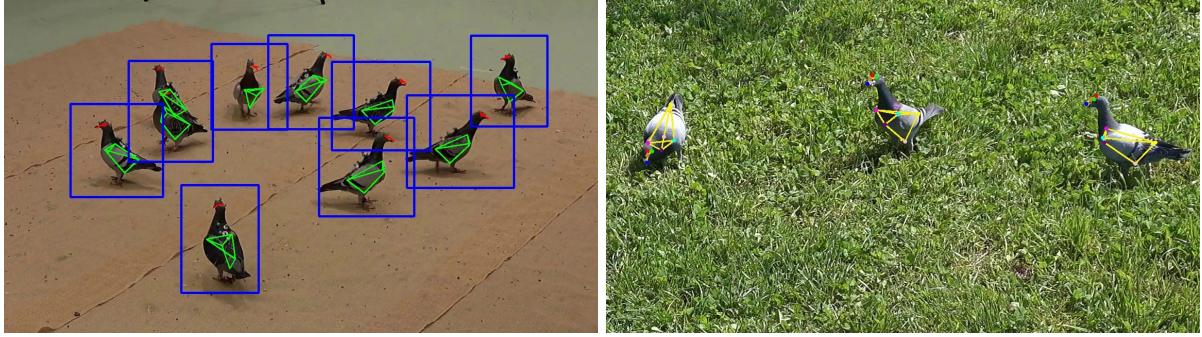
Markerless methods for animal posture tracking have been developing recently, but frameworks and benchmarks for tracking large animal groups in 3D are still lacking. To overcome this gap in the literature, we present 3D-MuPPET, a framework to estimate and track 3D poses of up to 10 pigeons at interactive speed using multiple-views. We train a pose estimator to infer 2D keypoints and bounding boxes of multiple pigeons, then triangulate the keypoints to 3D. For correspondence matching, we first dynamically match 2D detections to global identities in the first frame, then use a 2D tracker to maintain correspondences across views in subsequent frames. We achieve comparable accuracy to a state of the art 3D pose estimator for Root Mean Square Error (RMSE) and Percentage of Correct Keypoints (PCK). We also showcase a novel use case where our model trained with data of single pigeons provides comparable results on data containing multiple pigeons. This can simplify the domain shift to new species because annotating single animal data is less labour intensive than multi-animal data. Additionally, we benchmark the inference speed of 3D-MuPPET, with up to 10 fps in 2D and 1.5 fps in 3D, and perform quantitative tracking evaluation, which yields encouraging results. Finally, we show that 3D-MuPPET also works in natural environments without model fine-tuning on additional annotations. To the best of our knowledge we are the first to present a framework for 2D/3D posture and trajectory tracking that works in both indoor and outdoor environments.

## 1 Introduction

Pose estimation and tracking are among the fundamental problems in computer vision and a crucial task in many visual tracking applications ranging from sports in humans (Bridgeman, Volino, Guillemaut, & Hilton, 2019) to the study of collective behaviour in nonhuman animals (Couzin & Heins, 2022; Koger et al., 2023). For the latter, accurate quantification of behavior is critical to understand the underlying principles of social interaction and the neural and cognitive underpinnings of animal behaviour (Bernshtain, 1967; Altmann, 1974; Berman, 2018; Mathis et al., 2018; Kays, Crofoot, Jetz, & Wikelski, 2015). While researchers conventionally analyzed animal behaviour manually using a predefined catalogue of behaviours using ethograms, the recent advances in computer vision, as well as the increasing demands for large datasets involving the analysis of the fine-scaled and rapidly-changing behaviours of animals, encouraged the development of automated tracking methods (Dell et al., 2014; Gomez-Marin, Kampff, Costa, & Mainen, 2014; Anderson & Perona, 2014; Mathis et al., 2018). In such applications, multi-object pose estimation is essential to observe the dynamics of socially interacting individuals because individuals in a group tend to be partially occluded. Notably, with the recent advances in hardware and computer vision, marker-based motion capture systems have enabled posture tracking of single and multiple animals in controlled captive environments (Nagy et al., 2023; Kano, Naik, Keskin, Couzin, & Nagy, 2022; Itahara & Kano, 2022; Miñano, Golodetz, Cavallari, & Taylor, 2023; Itahara & Kano, 2023). Such marker-based motion capture systems also facilitated the curation of large-scale animal posture datasets (Naik et al., 2023; Marshall et al., 2021) to develop markerless methods for posture tracking of single (Mathis et al., 2018; Pereira et al., 2019; Dunn et al., 2021; Graving et al., 2019) and multiple animals (Lauer et al., 2022; Pereira et al., 2022; Waldmann, Naik, et al., 2022). A crucial advantage of markerless over marker-based methods is that individuals do not have to be equipped with

markers, thus opening possibilities for pose tracking of unhabituated animals even in the wild (i.e., natural habitat). Recently, with the success of 2D single animal markerless pose estimation methods like LEAP (Pereira et al., 2019) and DeepLabCut (DLC, Mathis et al. (2018)), this research area has received increased attention in method development for 2D tracking multiple animals (Lauer et al., 2022; Pereira et al., 2022; Graving et al., 2019; Waldmann, Naik, et al., 2022) and 3D postures (Günel et al., 2019; Joska et al., 2021; Dunn et al., 2021; Giebenhain, Waldmann, Johannsen, & Goldluecke, 2022; Han et al., 2023). This recent progress in markerless pose estimation also boosted the research area of computer vision for animals, as exemplified by the fact that the CVPR workshop on “Computer Vision for Animal Behavior Tracking and Modeling” (Zuffi et al., 2023) has been taking place every year since 2021. Topics of this workshop range from object detection (Duporge, Isupova, Reece, Macdonald, & Wang, 2021), behavior analysis (Nourizono et al., 2020; Bolaños et al., 2021), object segmentation (Chen et al., 2020; Waldmann, Bamberger, Johannsen, Deussen, & Goldlücke, 2022), 3D shape and pose fitting (Biggs, Roddick, Fitzgibbon, & Cipolla, 2019; Badger et al., 2020) to pose estimation (Labuguen et al., 2021; Gosztolai et al., 2021; Waldmann, Naik, et al., 2022) and tracking (Romero-Ferrero, Bergomi, Hinz, Heras, & de Polavieja, 2019; Pedersen, Haurum, Bengtson, & Moeslund, 2020; Waldmann, Naik, et al., 2022).

Despite recent progress in the field of computer vision for animals, reliable tracking of multiple moving animals in real-time and estimating their 3D pose to measure behaviours in a group remain an open challenge. While frameworks for multi-animal pose estimation and tracking in 2D (Lauer et al., 2022; Pereira et al., 2022; Waldmann, Naik, et al., 2022) are common, frameworks for 3D multi-animal pose estimation are generally lacking, with a few notable exceptions. We are aware of only two frameworks that estimate the 3D pose of more than one individual (two macaques (Bala et al., 2020) and two rats/parrots (Han et al., 2023)) in controlled captive environments, and finally one framework (Joska et al., 2021; Nath et al., 2019)



**Fig. 1** 3D Multi-Pigeon Pose Estimation and Tracking (3D-MuPPET) is a framework for multi-animal pose estimation and tracking for lab (left) and outdoor data (right). *Left:* Estimated complex pose (beak, nose, left and right eye, left and right shoulder, top and bottom keel and tail) of pigeons recorded in a captive environment. *Right:* The image shows an example with three pigeons recorded outdoors with estimated 3D keypoints reprojected to camera view (colored dots).

that estimates 3D poses of single Cheetahs in the wild.

One limiting factor for the development of animal pose estimation methods is the limited amount of annotated data as ground truth for training and evaluation, especially compared to human datasets (for example 3.6 million in Human 3.6M ([Ionescu, Papava, Olaru, & Sminchisescu, 2014](#))). Using birds as an example, we are aware of only four datasets for birds across different bird species ([Welinder et al., 2010](#); [Van Horn et al., 2015](#); [Badger et al., 2020](#); [Naik et al., 2023](#)). The lack of annotated datasets not only limits the ability to do thorough quantitative evaluation for new proposed methods, but biologists who want to make use of these methods also require a large amount of laborious manual annotations. DeepLabCut ([Mathis et al., 2018](#)), LEAP ([Pereira et al., 2019](#)) and DeepPoseKit ([Graving et al., 2019](#)) overcome this lack of training data using a human in loop approach where a small manually labelled dataset is used to train a neural network, then predict body parts (pre-labeling) of previously unlabeled material to generate larger training datasets. Creatures Great and SMAL ([Biggs et al., 2019](#)) instead creates synthetic silhouettes for training and extracts silhouettes with [J. Wang and Yuille \(2015\)](#); [P. Wang et al. \(2015\)](#) from real data for inference. Hence, one way to circumvent the lack of available annotated large-scale datasets for many animal species is to develop methods that exploit few training data in an efficient way. However, the drawback of this approach is that these methods cannot be evaluated quantitatively

in detail due to the few annotated data that they leverage.

We choose pigeons as an example species because of the recent introduction of a large scale multi-animal 2D/3D posture dataset in 3D-POP ([Naik et al., 2023](#)). This dataset opens up possibilities to propose and benchmark methods for 3D posture estimation and tracking due to its size. Here, we extend I-MuPPET ([Waldmann, Naik, et al., 2022](#)), a recent framework proposed for interactive 2D posture estimation and tracking of multiple pigeons, by incorporating multiple views to obtain 3D information. We will first evaluate the 2D framework proposed in I-MuPPET ([Waldmann, Naik, et al., 2022](#)) on the 3D-POP dataset, then introduce and evaluate our extension to 3D. We also highlight the applicability of the framework to data recorded in outdoor settings without any further annotations.

**Contributions.** In this paper, we present 3D-MuPPET, a framework for interactive tracking and 3D pose estimation of multiple pigeons that works for data recorded both in captivity and the wild. We obtain 3D poses by triangulating 2D poses of multiple views to 3D. 3D-MuPPET is comparable with a state of the art 3D pose estimation method in terms of Root Mean Square Error (RMSE) and Percentage of Correct Keypoints (PCK). We track up to ten pigeons (the upper limit in [Naik et al. \(2023\)](#)) with up to 10 fps in 2D and 1.5 fps in 3D, and report detailed results for speed and accuracy. We demonstrate that it is possible to train on an annotated dataset containing only a single pigeon to predict keypoints of a complex pose for multiple pigeons in

a stable and accurate way. This can simplify the domain shift to new species in applications in the wild because annotating single animal data is easier and less labour intensive than data with multiple animals. We also demonstrate the flexibility of our framework by estimating 3D poses of pigeons recorded outdoors, cf. Fig. 1, without any additional annotations. To evaluate pose estimation from data recorded outdoors, we also present Wild-MuPPET, a 3D posture dataset of 500 manually annotated frames from 4 camera views collected in the wild. To the best of our knowledge, we are the first to present a markerless 2D and 3D pose estimation framework for more than two animals that works with data recorded in both captivity and in the wild of the same species. Our approach is also not limited to pigeons and can be applied to other species, given 2D posture annotations are available. The code to reproduce the results of this work will be publicly available. We think that 3D-MuPPET offers a promising framework to make further developments for automated 3D multi-animal pose estimation and identity tracking, and ultimately promote fine-scaled methods for the study of animal collective behaviour.

## 2 Related Work

In this section, we explore existing work on both 2D and 3D posture estimation and multi-animal tracking, since 3D-MuPPET makes use of 2D detections and triangulation for 3D poses. We identify existing methods, then major gaps that we hope 3D-MuPPET can fill.

### 2.1 Animal Pose Estimation

**2D Single Animal Pose Estimation.** With the success of DeepLabCut (Mathis et al., 2018) and LEAP (Pereira et al., 2019), animal pose estimation has been developing into its own research branch parallel to human pose estimation. DeepLabCut and LEAP both introduce a method for labelling animal body parts and training a deep neural network for predicting 2D body part positions. DeepPoseKit (Graving et al., 2019) improved the inference speed by a factor of approximately two while maintaining the accuracy of DeepLabCut. 3D Bird Reconstruction (Badger

et al., 2020) predicts 2D keypoints and silhouettes to estimate the 3D shape of cowbirds from a single view. However, other than the extension of DeepLabCut in DeepLabCut-live (Kane, Lopes, Saunders, Mathis, & Mathis, 2020), most applications have focused on offline post-hoc analysis, which limits any application that might require posture estimation at interactive speeds to perform stimulus driven behavior experiments e.g. VR for animals (Naik, Bastien, Navab, & Couzin, 2020; Naik, 2021).

#### 2D Multi-Animal Pose Estimation.

DeepLabCut (Mathis et al., 2018) is extended in Lauer et al. (2022) to predict 2D body parts of multiple animals and maintain identity by temporal tracking. This extension uses training data with annotations of multiple animals. The authors released four datasets with annotations containing mice ( $n = 3$ ), mouse with pups ( $n = 2$ ), marmosets ( $n = 2$ ) and fish ( $n = 14$ ). Similarly SLEAP (Pereira et al., 2022) provides several architectures to estimate 2D body parts of multiple animals. These two approaches (Lauer et al., 2022; Pereira et al., 2022) can track the poses of multiple animals and are trained on multi-animal annotated data. However, manual annotations for multi-animal data is often challenging and time consuming to obtain, largely constraining the development of multi-animal methods.

**3D Animal Pose Estimation.** To infer 3D poses of single rodents from multi-view data, Dunn et al. (2021) developed a method similar to Iskakov, Burkov, Lempitsky, and Malkov (2019) by learning the triangulation process from multiple views using a 3D CNN. Similar to Iskakov et al. (2019), Dunn et al. (2021) has a cost of longer run times due to its 3D CNN architecture. Neural Puppeteer (Giebenhain et al., 2022) is a keypoint based neural rendering pipeline. By inverse rendering the authors estimate 3D keypoints from multi-view silhouettes. While this method is independent from variations in texture and lighting, most of their evaluation is performed using synthetic data, and thus its applicability to real-world animal data has not been extensively tested. J.J. Sun et al. (2023) proposes a self-supervised method for 3D keypoint discovery in animals filmed from multiple views without reliance on 2D/3D annotated data. This method uses joint length constraints and a similarity measure for spatio-temporal differences across

multiple views. While there is no need for annotated data, this method comes with a cost of lower accuracy. For Günel et al. (2019); Nath et al. (2019); Joska et al. (2021); Bala et al. (2020); Karashchuk et al. (2021); Ebrahimi et al. (2023); Han et al. (2023); Naik et al. (2023) the procedure to obtain 3D poses is to use a 2D pose estimator (e.g. Newell, Yang, and Deng (2016); Mathis et al. (2018)) and to triangulate to 3D using the 2D keypoint predictions of multiple views. Just like the proposed method, these 3D frameworks exploit 2D keypoints, while Ebrahimi et al. (2023) also post-processes the triangulated 3D keypoints.

All these methods are limited to the pose tracking of up to two individuals, and no framework has been shown to track the 3D poses of larger animal groups.

## 2.2 Multi-Animal Identity Tracking

Multiple animal tracking (Zhang, Gao, Xiao, & Fan, 2023), a variation of multi-object tracking (MOT, Dendorfer et al. (2021)), is important in order to maintain identities of animals throughout behavioural experiments.

Romero-Ferrero et al. (2019) and Heras, Romero-Ferrero, Hinz, and de Polavieja (2019) use the software idtracker.ai (Ferrero et al., 2017) to track up to 100 zebrafish in 2D at once. The software needs to know the number of individuals beforehand since it performs individual identification in each frame. TRex (Walter & Couzin, 2021) is capable of tracking up to 256 individuals while estimating the 2D head and rear positions of animals. It achieves real-time tracking using background subtraction. Zhang et al. (2023) provides a multi-animal tracking benchmark in the wild. The benchmark includes 58 sequences with around  $25K$  frames containing ten common animal categories with 33 target objects on average for tracking. Pedersen et al. (2020) provides a zebrafish tracking benchmark in 3D. The benchmark includes 3D data of up to ten zebrafish recorded in an aquarium.

**Frameworks for Animal Pose Estimation and Identity Tracking.** For applications in biological experiments of multiple individuals, the problem of posture estimation and tracking often goes hand in hand, because the posture of multiple individuals alone will not be meaningful if

the identities are not maintained. Existing posture estimation frameworks also provide identity tracking, but are often limited to 2D.

DeepLabCut (Lauer et al., 2022) splits the workflow in local and global animal tracking. For local animal tracking they build on SORT (Bewley, Ge, Ott, Ramos, & Upcroft, 2016), a simple online tracking approach. For animals that are closely interacting or in case of occlusions they introduce a global tracking method by optimizing the local tracklets with a global minimization problem using multiple cost functions on the basis of the animals' shape or motion. SLEAP (Pereira et al., 2022) also uses a tracker based on Kalman filter or flow shift inspired by Xiao, Wu, and Wei (2018) for candidate generation to track multiple individuals.

In contrast to the previous two works (Lauer et al., 2022; Pereira et al., 2022), we propose a posture estimation and tracking framework in 2D and 3D, that focuses on online tracking. We first initialize correspondences between cameras using the first frame and then use a 2D tracker from each view to maintain correspondences to reduce computation time. In addition, our framework works both on data recorded in captive and outdoor environments.

## 3 Technical Framework

We first discuss the datasets that we use for this study, describe the technical framework behind 3D-MuPPET, discuss ablation studies and network training, and finally explain how we extend the framework to data collected outdoors.

### 3.1 Datasets

We describe the dataset containing multi-pigeon annotations. In addition to pigeons, we also evaluate our framework on mouse and cowbird data in order to see how our framework performs compared to other animal pose estimation methods when applied to different species.

**3D-POP.** For this study, we use the 3D-POP dataset (Naik et al., 2023), a multi-view multi-individual dataset of freely-moving pigeons filmed by both RGB and motion-capture cameras. This dataset contains RGB video sequences from 4 views ( $4K$ ,  $3840 \times 2160$  px) of 1, 2, 5 and 10 pigeons. The ground truth provided by the dataset

for each individual is a bounding box, 9 distinct keypoints in 2D and 3D (beak, nose, left and right eye, left and right shoulder, top and bottom keel and tail), and individual identities. For more details on the curation and features of the dataset, we refer to [Naik et al. \(2023\)](#).

From this dataset, we adopt a 60/30/10 (training/validation/test) split based on 3D-POP ([Naik et al., 2023](#)), by sampling a total of 6036 random images as our training set from the sequences of 1, 2, 5 and 10 pigeons (25% for each type). We ensure that an equal number of frames was sampled from each sequence to avoid bias. As our validation and test set, we sample 3040 and 1004 frames separately from the training set following the same sampling method.

To test if training a model on 1 pigeon can be used to track multiple pigeons, we also sample a single pigeon training set, using the same sampling method but only from single pigeon sequences. The dataset contains 6006, 3012 and 1034 images for training, validation and test respectively.

Finally, to perform quantitative evaluation on multi-object tracking in 2D and 3D, we use the 5 separate test sequences containing 10 pigeons provided in 3D-POP ([Naik et al., 2023](#)), ranging between 1 to 1.5 minutes in length.

**Wild-MuPPET.** We also provide a novel dataset collected from pigeons foraging in an outdoor environment. The data is collected from 4 synchronized and calibrated cameras (4K, 30fps) mounted on tripods in a rectangular formation, similar to 3D-POP ([Naik et al., 2023](#)). The dataset consists of short sequences featuring between 1 to 3 pigeons under natural sunlight conditions. To provide a quantitative evaluation of pose estimation performance in the wild, we also sample and manually annotate 500 frames from a single individual sequence, taken from all 4 views. These annotated keypoints are then triangulated to obtain 3D ground truth data. To the best of our knowledge, this is the first calibrated multi-view video dataset of more than one animal that is captured in fully outdoor settings (cf. [Joska et al. \(2021\)](#) for a 3D single Cheetah dataset).

**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  $640 \times 480$  or  $800 \times 800$  since the data was recorded with two different monochromatic cameras.

**Cowbird Data.** This 2D data from [Badger et al. \(2020\)](#) contains single cowbirds. Their original images have a maximum resolution of  $1920 \times 1200$  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  $256 \times 256$ .

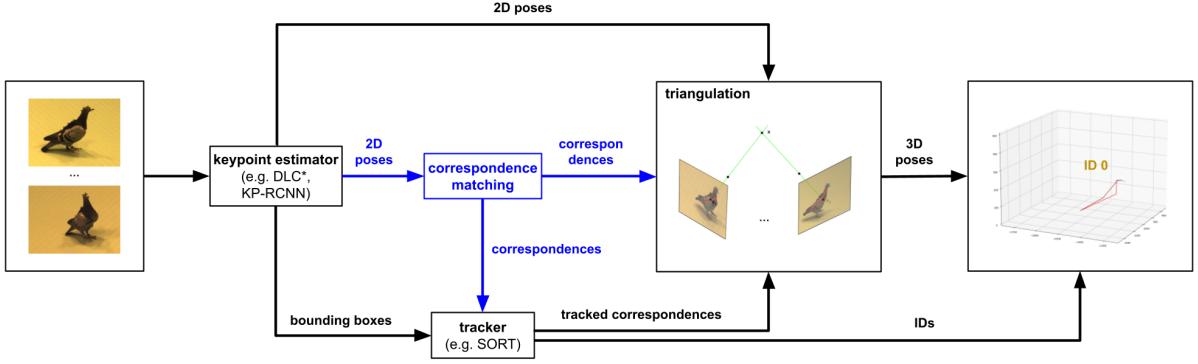
For more details on these two datasets we refer to [Mathis et al. \(2018\)](#); [Badger et al. \(2020\)](#).

### 3.2 Pose Estimation and Identity Tracking

This work extends upon I-MuPPET ([Waldmann, Naik, et al., 2022](#)) and thus the core components of our framework are a pose estimation module and a tracking module, into which we can readily slot any state of the art pose estimator or tracking method, see Fig. 2. In the pose estimation module we use two methods for comparison, i.e. a KeypointRCNN ([He, Gkioxari, Dollar, & Girshick, 2017](#)) and a modified DeepLabCut (DLC, [Mathis et al. \(2018\)](#)).

For the modified DLC, we adopt a top-down approach, by first using YOLOv8 ([Jocher, Chaurasia, & Qiu, 2023](#)) to detect the individual pigeons in each frame and then pass the cropped pigeon images into the single animal DLC ([Mathis et al., 2018](#)) pipeline. For details, we refer to [Mathis et al. \(2018\)](#). In the following, we denote this model by DLC\* (with an asterisk).

The KeypointRCNN is a PyTorch ([Paszke et al., 2019](#)) implementation of a Mask R-CNN ([He et al., 2017](#)), which is modified to output nine keypoints for each detected instance (individual), in addition to a confidence score (confidence of the model about its prediction), label (background vs. object) and bounding box. Like DLC ([Mathis et al., 2018](#)), this network has a ResNet-50-FPN ([He, Zhang, Ren, & Sun, 2016; Lin et al., 2017](#)) backbone that was pre-trained on ImageNet ([Deng et al., 2009](#)), similar to [He et al. \(2017\)](#). For details, we refer to [He et al. \(2017\)](#). The input to the KeypointRCNN are RGB images (cf. Fig. 2) normalized to mean and standard deviation of 0.5.



**Fig. 2** 3D-MuPPET. The framework consists of a pose estimation and tracking module, into which we can readily slot any state of the art pose estimator and tracking method. We perform correspondence matching (blue part) based on Huang et al. (2020) in the first frame only. In the subsequent frames we track the correspondences with SORT (Bewley et al., 2016). 3D-MuPPET predicts 3D poses together with IDs from multi-view image input using triangulation.

The KeypointRCNN is trained in a fully supervised manner using stochastic gradient descent with learning rate decay, momentum and weight decay.

We also implement data augmentation for training in order to avoid overfitting and to mimic other conditions than those present in the pigeon data. Specifically, our input data to the KeypointRCNN (He et al., 2017) has a specific probability to be flipped, scaled within a specified range, and changed in brightness or sharpness. For DLC (Mathis et al., 2018) we use their default augmentation parameters.

**3D Posture Estimation.** We use the 2D postures of all four camera views obtained from KeypointRCNN and DLC\* to acquire 3D keypoint estimates using triangulation with sparse bundle adjustment. For that we determine the individual ID of each keypoint detection by matching the detected and ground truth bounding box based on percentage overlap and Hungarian algorithm (Kuhn, 1955), allowing correspondence matching across all four views. Since correspondence matching errors during triangulation can lead to inflated error metrics in terms of RMSE which do not reflect the actual accuracy of the methods, we apply a simple filtering process. Specifically, we remove individuals with a mean keypoint error that exceeds a given threshold. We choose 100mm here, which is similar to the body width of a pigeon, and assuming any error bigger than that is due to correspondence matching errors instead of keypoint estimation error.

Since we use randomly sampled frames to evaluate pose estimation, we can only rely on an arbitrary threshold for filtering. An effective alternative is to use temporal information (e.g. sudden jumps of keypoints) for filtering and subsequent gap-filling. In the following, we denote the two 3D-MuPPET posture estimation modules by 3D-KeypointRCNN and 3D-DLC\*.

**3D Multi-Animal Identity Tracking.** For multi-animal tracking, we first use SORT (Bewley et al., 2016) to track the identity of individuals in each of the four camera views in 2D. We chose this method since we are primarily interested in online tracking and high inference speed, and SORT (Bewley et al., 2016) can run up to 260 fps. We use standard parameters and a maximum age of 10 frames (refer to Bewley et al. (2016) for details). To determine the correct correspondences between 2D tracks of each view, we use a dynamic matching algorithm based on Huang et al. (2020) in the first frame to assign each SORT ID from each view to a global ID (cf. blue part in Fig. 2). The dynamic matching algorithm first obtains 3D estimates of objects in a 3D subspace between each camera pair, then matches 3D poses based on Euclidian distance, until the pairwise distance threshold of 200mm is reached. Here, we choose a conservative threshold of 200mm to ensure all subjects are matched. Note that the algorithm prioritizes matches with lower distance, hence a larger threhold doesn't lead to worse performance (cf. Huang et al. (2020)). After

the dynamic matching is completed, we maintain 3D correspondences in subsequent frames and triangulate based on 2D tracklets. Finally, if a 2D tracklet in a certain camera view is lost or switched, we skip the detections of the given camera.

**Pigeons in the Wild.** Usually, the difference in the background between different datasets is one of the biggest hurdles for generalizing a keypoint detection model trained on an annotated dataset to other data of the same species. Here, we propose a methodology to eliminate the effect of the background to estimate postures of pigeons in the wild without further annotation and fine-tuning. For training, we make use of the same multi-animal training set sampled from 3D-POP cf. Sec. 3.1. We first remove the influence of the background by using the Segment-Anything-Model (SAM, Kirillov et al. (2023)), a model that allows any object in an image to be segmented based on a prompt of the object location. In this case, we use the ground truth bounding box as prompt to obtain masks of pigeons. We then train our framework to predict keypoints on masked images instead of bounding boxes containing both object and background. For inference in videos captured in the wild, we first use a pre-trained MaskRCNN (He et al., 2017) to find all pigeons in the frame and then pass them to the pose estimator.

For our results of pigeons in the wild we use DLC\* in the pose estimation module of 3D-MuPPET, cf. Fig. 2.

### 3.3 Network Training and Ablation Studies

**Data Augmentation for Pigeons.** In I-MuPPET (Waldmann, Naik, et al., 2022), we performed ablation studies on data augmentation for pigeons. These ablation studies can be found in our supplemental material. In this work, we use the same data augmentation parameters to train the KeypointRCNN model (cf. Sec. 3.2). They include changing the sharpness with a probability of 0.2, blurring the input image with a small probability of 0.2, randomly jittering the brightness by a factor chosen uniformly from [0.4, 1.6], a flipping probability of 0.5 and a small scaling range of  $\pm 5\%$ .

For DLC\* (cf. Sec. 3.2), we use their default augmentation parameters (Mathis et al., 2018; Jocher et al., 2023) that also include blurring and jittering.

**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. supplemental material).

**Training Hyperparameters.** To find out the best network configuration for the KeypointRCNN (cf. Sec. 3.2) we perform several experiments (see supplemental material). From this analysis we find that using a learning rate of 0.005 and reducing it by  $\gamma = 0.5$  every given step size to reach a final learning rate of 0.0003 at the end of training works best.

For DLC\* (cf. Sec. 3.2), we use a custom learning rate schedule from 0.0001 to 0.00001 over 30000 iterations for DLC, and default hyperparamters for all others (Mathis et al., 2018; Jocher et al., 2023).

**Training Procedure.** For all trained neural networks, we monitor the validation loss when training, with the final weights chosen based on the epoch with the lowest validation loss overall to ensure the best performance and least over-fitting. For DeepLabCut, we instead use RMSE accuracy provided by the package (Mathis et al., 2018).

## 4 Evaluation

First, we briefly discuss the evaluation metrics we use. Then, we evaluate the pose estimation module of 3D-MuPPET on captive and wild pigeon, mouse and cowbird datasets, cf. Sec. 4.2. In addition we evaluate the tracking module of 3D-MuPPET in Sec. 4.3 by performing quantitative tracking evaluation and benchmarking the inference speed. We also show qualitative results on all tasks.

Since the current framework extends the work of I-MuPPET (Waldmann, Naik, et al., 2022), we first re-evaluate the 2D performance with 3D-POP (Naik et al., 2023) and then evaluate the 3D

**Table 1** Quantitative Evaluation of 2D Pigeon Poses. We report the RMSE and its median (px), PCK05 (%) and PCK10 (%) for estimated 2D poses. Comparison between DeepLabCut (DLC) and KeypointRCNN (KP-RCNN), cf. Sec. 3.2.  $\dagger$ : These models are trained on single-pigeon data, instead of multi-pigeon data for a comparison. Best results per row in bold.

Metric / Method	KP-RCNN $\dagger$	KP-RCNN	DLC*
RMSE (px) $\downarrow$	136.1	40.0	<b>15.7</b>
Median (px) $\downarrow$	7.3	5.1	<b>4.4</b>
PCK05 (%) $\uparrow$	68.4	86.0	<b>91.0</b>
PCK10 (%) $\uparrow$	82.6	97.3	<b>98.7</b>

performance for all tasks. As 3D-MuPPET estimates 3D poses from multi-view 2D keypoints, comparing 2D and 3D performance can also provide insights into how errors propagate. For original evaluation results in I-MuPPET, we refer to [Waldmann, Naik, et al. \(2022\)](#).

The source code to reproduce the results of this paper will be publicly available at <https://github.com/alexhang212/3D-MuPPET>.

## 4.1 Metrics

**Pose Estimation.** Two widely used metrics, also in human pose estimation, are the Root Mean Square Error (RMSE), in human pose estimation better known as Mean Per Joint Position Error (MPJPE, cf. e.g. [Iskakov et al. \(2019\)](#)), and the Percentage of Correct Keypoints (PCK, cf. e.g. [Yang and Ramanan \(2013\)](#)). DeepLabCut ([Mathis et al., 2018](#)) uses the former, 3D Bird Reconstruction ([Badger et al., 2020](#)) the latter to evaluate their animal pose estimation, hence we use both here.

RMSE is calculated by taking the root mean squared of the Euclidean distance between each predicted point and the ground truth point, while PCK is the percentage of predicted keypoints that fall within a given threshold ([Badger et al., 2020](#)). We compute PCK05 and PCK10, where the threshold is a fraction (0.05 and 0.1) of the largest dimension of the ground truth bounding box for 2D and the maximum distance between any two ground truth keypoints in 3D. Compared to RMSE, the PCK takes into account the size and scale of the tracked animal, providing a more meaningful estimate of keypoint accuracy compared to the RMSE.

**Tracking.** There are three sets of tracking performance measures that are widely used in the

literature ([Dendorfer et al., 2021](#)): the CLEAR-MOT metrics introduced in [Bernardin and Stiefelhagen \(2008\)](#), the metrics introduced in [Li, Huang, and Nevatia \(2009\)](#) to measure track quality, and the trajectory-based metrics proposed in [Ristani, Solera, Zou, Cucchiara, and Tomasi \(2016\)](#). Here, we also report the novel Higher Order Tracking Accuracy (HOTA), introduced in [Luiten et al. \(2021\)](#) because the other metrics overemphasize the importance of either detection or association. HOTA measures how well the trajectories of matching detections align, and averages this over all matching detections, while also penalising detections that do not match ([Luiten et al., 2021](#)).

For further details on the tracking metrics we refer to [Dendorfer et al. \(2021\)](#); [Luiten et al. \(2021\)](#). A detailed description of each reported metric is also available in the supplementary material. For the evaluation, we use [Jonathon Luiten \(2020\)](#) and code provided by [Dendorfer](#).

## 4.2 Keypoint Estimation

In Sec. 4.2.1 we report quantitative and qualitative results of 2D and 3D poses on the pigeon data and compare 3D-MuPPET to a 3D baseline ([Iskakov et al., 2019](#)). Further, we compare the KeypointRCNN (cf. Sec. 3.2) to DLC ([Mathis et al., 2018](#)) on the 2D odor trail tracking data from [Mathis et al. \(2018\)](#) in Sec. 4.2.3. And in Sec. 4.2.4 we compare the performance of KeypointRCNN to 3D Bird Reconstruction ([Badger et al., 2020](#)), which uses an HRNet architecture ([K. Sun, Xiao, Liu, & Wang, 2019](#)) on a 2D cowbird keypoint dataset. In addition to the detailed evaluation on pigeon data, we also evaluate the mouse and cowbird data to see how our framework performs with respect to other animal pose estimation methods when applied to different species.

#### 4.2.1 Comparison on 3D-POP (Pigeons)

**Baseline.** For a 3D comparison, we train the [Iskakov et al. \(2019\)](#) framework (LToHP), a state of the art 3D pose estimation model for humans, on the same training dataset specified in Sec. 3.1. The framework predicts a 2D heatmap from each view that is projected into a 3D voxel grid. The model then learns to predict 3D keypoints using a 3D CNN architecture. We provide cropped images of pigeon individuals from all views and a 3D root point (top keel) from ground truth data, as well as camera calibration parameters during training and inference. We train this model for 782 epochs with default augmentation parameters.

**Results.** We train our framework on single and multi-pigeon data from [Naik et al. \(2023\)](#), cf. Sec. 3.1, and choose the best weights that have the lowest validation loss. We train the KeypointRCNN (cf. Sec. 3.2) for 44 epochs on multi-pigeon data, and 30 epochs on the single-pigeon dataset. In the case of DLC\* (cf. Sec. 3.2), we train YOLOv8 ([Jocher et al., 2023](#)) for 27 epochs and DLC ([Mathis et al., 2018](#)) for 86000 iterations. Since this is a two step approach, we train the two networks separately.

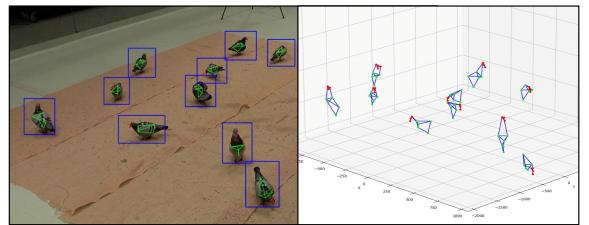
Quantitative results for 2D pose estimation are in Tab. 1. We find that DLC\* performs better across all metrics than KeypointRCNN, cf. Tab. 1. From Tab. 1 we also see that it is possible to train the KeypointRCNN (cf. Sec. 3.2) on single pigeon data and obtain good results on multi-pigeon data (median of 7.3px vs. 5.1px). This can simplify the domain shift to new species in applications in the wild because obtaining annotated single animal data is easier and less labour intensive than data with multi-animal annotations.

In 3D, the RMSE is inflated by predictions that have implausible values due to correspondence matching errors during triangulation (rather than inaccurate 2D predictions and/or triangulation errors). To avoid such implausible values, we remove individuals with a mean keypoint error exceeding 100mm (which is large compared to the size of a pigeon) prior to calculating the evaluation metrics. From this filtering process, we remove 0, 2, 24 (698†) individuals (out of 4486 individual instances within the

1034 test images, †: model is trained on single-pigeon dataset) for LToHP ([Iskakov et al., 2019](#)), 3D-DLC\*, and 3D-KeypointRCNN (cf. Sec. 3.2) respectively. Note that the quantitative 3D results in Tab. 2 undergo this filtering process. Like in 2D, we find that 3D-DLC\* performs better across all metrics than the 3D-KeypointRCNN. This is not surprising since DLC\* already performs better in 2D, cf. Tab. 1. Additionally, 3D-KeypointRCNN suffers from a higher number of outliers compared to 3D-DLC\* as seen above in the number of individuals removed by filtering. We conclude that in applications where a higher accuracy is needed, researchers should prefer 3D-DLC\* for the pose estimation module of 3D-MuPPET, cf. Fig. 2.

LToHP ([Iskakov et al., 2019](#)) has the best performance in 3D across all metrics, cf. Tab. 2. One of the reasons that LToHP performs best in 3D pose estimation is that bounding boxes of the subject and root point (top keel) are provided from the ground truth. In addition, the model can learn the general 3D structure of a pigeon instead of relying on simple triangulation. While the RMSE is about 10mm and the PCK10 over 90% for both LToHP ([Iskakov et al., 2019](#)) and 3D-MuPPET, for PCK05 3D-MuPPET falls behind by about 15%, cf. Tab. 2.

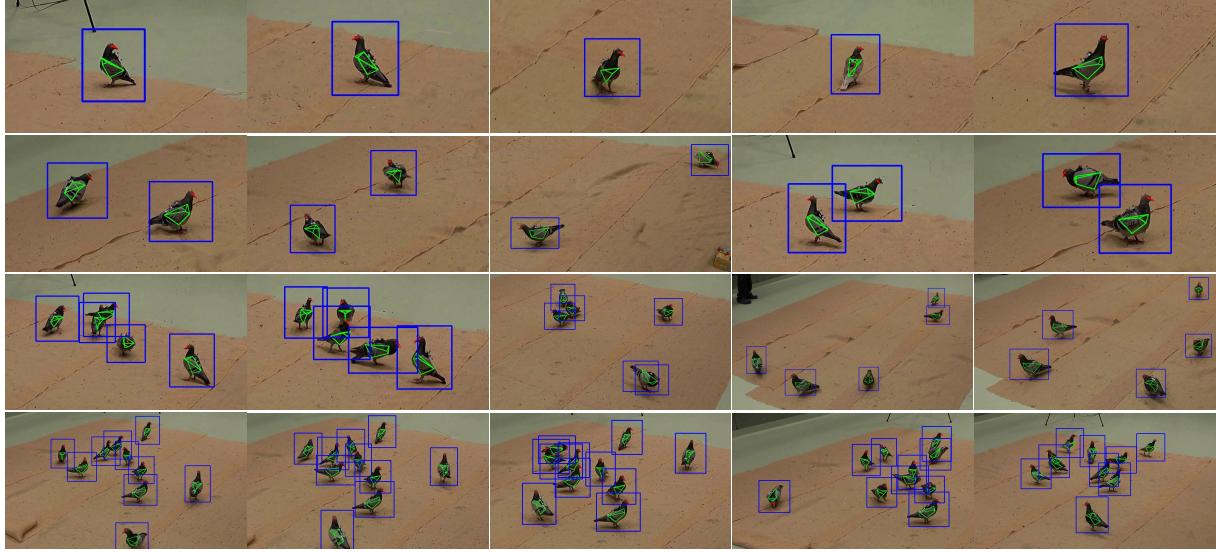
Overall, this comparison shows that 3D-MuPPET using simple triangulation achieves performance comparable to LToHP ([Iskakov et al., 2019](#)), see Figs. 3 and 4. From Tab. 2 we also see that it is possible to train on single pigeon data and predict poses on multi-pigeon data (marked with † in Tab. 2) with comparable results (median of 13.1mm vs. 8.6mm), even though the method suffers from higher amount of outliers as evident from the RMSE and large number of individuals filtered. This can simplify the domain shift to new



**Fig. 3 Qualitative 3D Results:** Example frame from [Naik et al. \(2023\)](#) using 3D-MuPPET. 2D (left side) and 3D (right side) pose estimates.

**Table 2** Quantitative Evaluation of 3D Pigeon Poses. We report the filtered (cf. Sec. 3.2) RMSE and its median (mm), PCK05 (%) and PCK10 (%) for the 3D poses. Comparison between LToHP (Iskakov et al., 2019) and 3D-MuPPET (highlighted in gray). †: These models are trained on single-pigeon data, instead of multi-pigeon data for a comparison. Best results per row in bold. See text for a discussion of the results.

Metric / Method	3D-KP-RCNN†	3D-KP-RCNN	3D-DLC*	LToHP (Iskakov et al., 2019)
RMSE (mm) ↓	40.4	15.1	11.5	<b>7.5</b>
Median (mm) ↓	13.1	8.6	6.9	<b>4.2</b>
PCK05 (%) ↑	33.5	59.3	72.3	<b>89.2</b>
PCK10 (%) ↑	56.7	89.1	95.2	<b>98.4</b>



**Fig. 4** Qualitative Results. Example frames from 3D-POP (Naik et al., 2023) for multi-pigeon pose estimation and tracking in 3D, reprojected to 2D view. Green lines connect the body, red lines the head keypoints. Some frames are cropped for a better view.

**Table 3** Quantitative Evaluation of 3D Pigeon Poses in the Wild. We report the filtered (cf. Sec. 3.2) RMSE and its median (mm), PCK05 (%) and PCK10 (%) for the 3D poses of pigeons in the wild, cf. Sec. 3.1. See text for a discussion of the results.

Metric / Method	3D-MuPPET in the Wild
RMSE (mm) ↓	56.5
Median (mm) ↓	24.7
PCK05 (%) ↑	8.15
PCK10 (%) ↑	31.0

species in applications in the wild because obtaining annotated single animal data is easier and less labour intensive than data with multi-animal annotations.

#### 4.2.2 Comparison on Wild-MuPPET (Pigeons)

In Tab. 3 we report quantitative results for 3D pose estimation of pigeons in the wild, cf. Sec. 3.1. We adopt the same filtering procedure for RMSE of  $> 100\text{mm}$  and remove 125 frames. Generally, the results suggest a large number of outliers, but we still show that 3D-MuPPET can work with data collected in outdoor settings, with models trained with data collected in captive environments cf. also Fig. 6. Further model fine-tuning with the data can further improve performance in the wild.

**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). We report precision within  $\pm 0.2$  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	$3.6 \pm 0.2$
DLC, 350K/600K iterations	<b><math>3.2 \pm 0.2</math></b>

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

In order to compare the KeypointRCNN (cf. Sec. 3.2) to DeepLabCut (which is our other choice for the pose estimation module of our framework, cf. Fig. 2) 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 Sec. 3.3. 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 (cf. Sec. 3.2). 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 Tab. 1 in Sec. 4.2.1. Please note that DLC in Mathis et al. (2018) in contrast to the KeypointRCNN (cf. Sec. 3.2) 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.

#### 4.2.4 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. 4.1) 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 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, cf. Sec. 3.2 and Fig. 2) to the modified HRNet (K. Sun et al., 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 Sec. 3.3. 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

**Table 5** 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	<b>0.46</b>	<b>0.64</b>

**Table 6** Quantitative Tracking Evaluation in 2D. We test 20 video sequences quantitatively with the metrics specified in Sec. 4.1 and Sec. 2.2 in supplementary materials. The threshold for the confidence score of DLC\* (cf. Sec. 3.2) is set to 0.5.

Test seq.	HOTA $\uparrow$	MOTA $\uparrow$	MOTP $\uparrow$	Rec $\uparrow$	Prcn $\uparrow$	MT $\uparrow$	ML $\downarrow$	FPF $\downarrow$	IDS $\downarrow$	Frag $\downarrow$	IDF1 $\uparrow$
11, view 1	0.83	0.92	0.90	0.96	0.96	0.90	0	0.39	2	14	0.92
11, view 2	0.85	0.92	0.89	0.96	0.96	0.90	0	0.41	0	7	0.96
11, view 3	0.85	0.92	0.89	0.96	0.96	0.90	0	0.41	0	11	0.96
11, view 4	0.86	0.94	0.91	0.97	0.97	1	0	0.26	3	29	0.95
19, view 1	0.91	0.99	0.92	0.99	1	1	0	0	2	13	0.97
19, view 2	0.93	1	0.92	1	1	1	0	0	0	1	1
19, view 3	0.93	1	0.92	1	1	1	0	0	0	4	1
19, view 4	0.89	0.99	0.93	0.99	1	1	0	0	4	11	0.94
30, view 1	0.84	0.96	0.93	0.97	1	1	0	0.03	9	25	0.88
30, view 2	0.90	0.99	0.93	0.99	1	1	0	0.03	7	14	0.95
30, view 3	0.89	0.99	0.89	0.99	1	1	0	0.03	2	7	0.99
30, view 4	0.87	0.99	0.91	0.99	1	1	0	0.02	6	13	0.95
48, view 1	0.88	0.99	0.90	1	1	1	0	0	1	14	0.96
48, view 2	0.91	1	0.90	1	1	1	0	0.02	0	8	1
48, view 3	0.92	1	0.91	1	1	1	0	0	0	4	1
48, view 4	0.92	1	0.91	1	1	1	0	0	0	6	1
59, view 1	0.77	0.98	0.89	0.98	1	1	0	0.02	8	33	0.82
59, view 2	0.80	0.97	0.90	0.97	1	1	0	0.02	12	40	0.84
59, view 3	0.79	0.98	0.89	0.98	1	1	0	0.02	8	28	0.87
59, view 4	0.80	0.97	0.89	0.97	1	1	0	0.02	8	40	0.89
<b>Combined</b>	<b>0.87</b>	<b>0.98</b>	<b>0.91</b>	<b>0.98</b>	<b>0.99</b>	<b>0.99</b>	<b>0</b>	<b>0.08</b>	<b>72</b>	<b>322</b>	<b>0.94</b>

the PCK results derived from KeypointRCNN with 45 and 60 epochs.

Tab. 5 compares results for 3D Bird Reconstruction from Badger et al. (2020) with the KeypointRCNN (cf. Sec. 3.2). 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).

### 4.3 Tracking Performance

Figs. 4 and 6 show results of the 3D pose estimation and tracking task for multiple pigeons in captive environments and the wild, respectively. Further qualitative results can be found in our supplementary video at [https://youtu.be/GZZ\\_u53UpfQ](https://youtu.be/GZZ_u53UpfQ).

**Quantitative Tracking Evaluation.** We test our framework quantitatively in 2D and 3D on five video sequences from 3D-POP, cf. Sec. 3.1. Each sequence contains ten pigeons (50 objects in total, 200 in 2D) and 10053 frames (40212 frames in 2D). Since the sequences contain small

gaps due to missed detections in motion capture (see Naik et al. (2023) for more details), we use linear interpolation to fill all gaps before evaluation. For evaluation we use DLC\* (cf. Sec. 3.2) and the metrics specified in Sec. 4.1. Note that for sequence 59, we remove the first 3 seconds (90 frames) since 2 pigeons are initially outside the tracking volume which causes the first frame matching (see Sec. 3.2) to fail.

Detailed 2D results for a detection confidence threshold of 0.5 are shown in Tab. 6. Overall, we achieve good results with our framework on the 2D video sequences (HOTA: 0.87, MOTA: 0.98, MOTP: 0.91, Recall: 0.98, Precision: 0.99, MT: 0.99, ML: 0, FPF: 0.08 and IDF1: 0.94; metrics specified in Sec. 4.1 and our supplemental material).

In Tab. 7 we report detailed 3D tracking results of the bottom keel joint for the five sequences where we set the maximum allowed distance between detections and ground truth positions in Dendorfer to 30mm. We choose 30mm as this threshold is well within the body size of a

**Table 7** Quantitative Tracking Evaluation in 3D. We test five sequences quantitatively with the metrics specified in Sec. 4.1. For detailed explanations on abbreviations and metrics, please refer to our supplemental material. See text for a discussion of the results.

Seq.	MOTA↑	MT↑	ML↓	IDS↓	Frag↓
11	0.92	1	0	0	192
19	0.89	0.90	0	0	272
30	0.89	0.80	0	0	337
48	0.90	1	0	0	328
59	0.59	0.60	0	4	490
<b>Comb.</b>	0.84	0.86	0	4	1619

**Table 8** 2D Inference Speed. Benchmark for the complete pipelines (including data loading, model loading, inference, data saving). We report the inference speed (fps) for the 2D models, cf. Sec. 3.2. Best results per column in bold. See text for a discussion of the results.

Method / Num. of Ind.	1	2	5	10
KP-RCNN	<b>6.06</b>	<b>6.20</b>	<b>6.10</b>	<b>5.9</b>
DLC*	2.80	2.58	2.13	1.64

pigeon, while taking into account the possible distance an individual can move within one frame. Overall, we achieve good 3D results with 3D-MuPPET (MOTA: 0.84, MT: 0.86, and ML: 0; metrics specified in Sec. 4.1 and our supplemental material).

**Inference Speed.** We also benchmark the inference speed of our framework in 2D and 3D with all 1004 frames in the test set sampled from 3D-POP (Naik et al., 2023), cf. Sec. 3.1. For this evaluation, we use a workstation with a 16GB Nvidia Geforce RTX 3070 GPU, 11th Gen Intel(R) Core(TM) i9-11900H @ 2.50GHz CPU, and Sandisk 2TB SSD.

Since each pose estimation module of 3D-MuPPET (cf. Fig. 2) has different data and model loading procedures, we include all processes (data loading, model loading, inference, data saving) to get a realistic comparison of the processing time. We loop three times over each inference script and report the average speed in frames per second (fps).

2D results are in Tab. 8. The KeypointRCNN has a faster inference speed than DLC\*. For the KeypointRCNN, we obtain an interactive speed of about 6 fps (cf. Tab. 8) for our full pipeline.

**Table 9** 2D Inference Speed. Benchmark for our in-memory pipeline using the KeypointRCNN, cf. Sec. 3.2. We benchmark our pipeline with our AVI video sequences preloaded in memory and report values for different batch sizes.

batch size	frame rate [fps]			
	1 pigeon	2 pigeons	5 pigeons	10 pigeons
1	9.10	9.32	9.14	8.68
2	9.26	9.25	9.07	8.76
4	9.54	9.49	9.44	8.88
8	9.56	9.64	9.43	8.97
16	<b>9.96</b>	<b>10.15</b>	<b>9.83</b>	<b>8.99</b>

Interestingly, speed is almost independent of the number of pigeons present in the video.

To push the inference speed of the KeypointRCNN even further, we also benchmark the scenario where we pre-load the video sequence in memory and are thus independent of disk I/O, with otherwise the same procedure, see Tab. 9 for results. We report values for batch sizes up to 16, restricted by the hardware that we use. The speed of our pipeline increases for a batch size of 1 by about 3 fps (comparing Tab. 8 with Tab. 9) if we preload the video to memory. The maximum speed is at a batch size of 16 with an interactive speed of about 9 – 10 fps depending on the number of pigeons present in the video sequence.

There is another framework that also performs 2D keypoint prediction of complex poses and tracking: SLEAP (Pereira et al., 2022). Their inference speed benchmark procedure and hardware are comparable to 3D-MuPPET. For details we refer to Pereira et al. (2022). A rough comparison yields that SLEAP (Pereira et al., 2022) is about an order of magnitude faster than the KeypointRCNN (numbers read off from Pereira et al. (2022), Figs. 2b, 3e and Extended Data Fig. 6c; considering the fact that the pigeon image resolution is higher than the one of the flies and mice (open field) and thus we process more data through the whole pipeline). While our framework solves the substantially harder task of a ‘generalist’ approach of training a single model that works on all datasets, SLEAP uses a ‘specialist’ paradigm where small, lightweight models have just enough representational capacity to generalize to the low variability typically found in scientific data (Pereira et al., 2022). The approach of our framework comes with an additional cost of computing resource requirements. However,

**Table 10** *3D Inference Speed.* Benchmark for the complete pipelines (including data loading, model loading, inference, data saving). We report the inference speed (fps) for the 3D models. Best results per column in bold, 3D-MuPPET versions highlighted in gray. See text for a discussion of the results.

Method / Num. of Ind.	1	2	5	10
3D-KP-RCNN	1.53	1.50	<b>1.45</b>	<b>1.37</b>
3D-DLC*	0.71	0.65	0.54	0.41
LToHP (cf. Sec. 4.2.1)	<b>3.95</b>	<b>2.12</b>	0.78	0.38

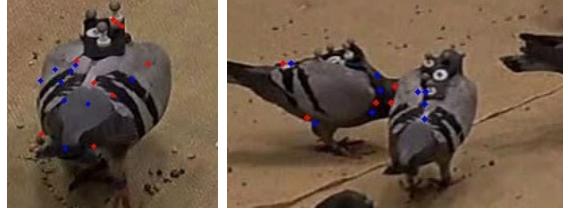
we want to offer a framework that works with both low and high variability data at the same time. Depending on the application, one can easily change the pose estimator of our framework (cf. Sec. 3.2 and Fig. 2) to achieve frame rates comparable to SLEAP.

Finally, we benchmark the inference speed in 3D and results are in Tab. 10. LToHP (Iskakov et al., 2019) has the fastest inference speed for up to two individuals, and 3D-MuPPET using 3D-KeypointRCNN is faster than LToHP for five and ten individuals.

This is likely because LToHP, like 3D-DLC\*, predicts postures from cropped images of individual pigeons, while 3D-KeypointRCNN predicts from the whole frame. The advantage for LToHP and 3D-DLC\* is the processing of smaller images, and thus less data is amortized. For groups with five or more individuals, these models are executed several times, compared to 3D-KeypointRCNN, which is only executed once. Also note that for applications with datasets that do not contain ground truth, the inference speed of LToHP will likely be slower since bounding box and root point of each subject will first need to be detected. Researchers that prioritize inference speed for multi-animal posture estimation and tracking may consider the KeypointRCNN for the pose estimation module in 3D-MuPPET. From the 2D inference speed evaluation we also see that the inference speed can be pushed even further by preloading the data in memory and processing batches, cf. Tab. 9.

#### 4.4 Limitations and Future Work

Keypoint detection can fail e.g. due to self-occlusions or occlusions from other individuals

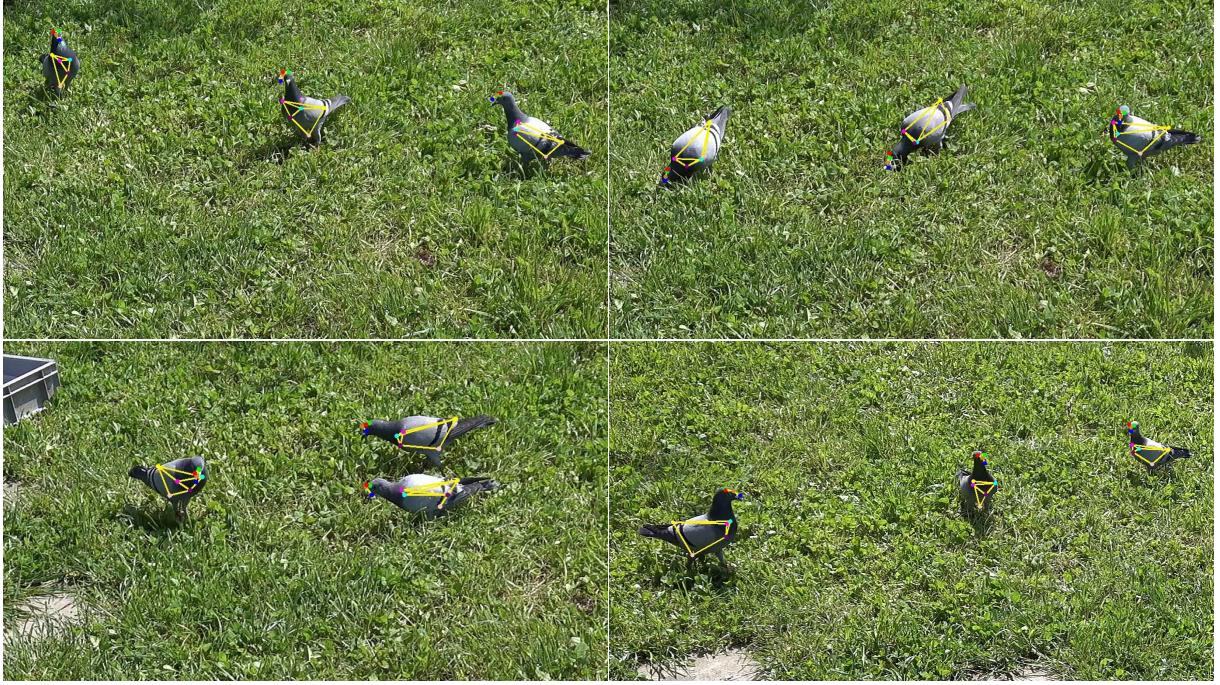


**Fig. 5 Limitations.** Cropped frames of failure cases from 3D-POP (Naik et al., 2023) data for 2D pose estimation using the KeypointRCNN (cf. Sec. 3.2), due to occlusions. Blue denotes the ground truth, red denotes the prediction.

(cf. Fig. 5), which can affect the triangulation procedure. This may have caused a large number of outliers present in 2D and 3D keypoint evaluation, as indicated by the higher RMSE values in contrast to the median error. Both drawbacks can be alleviated by methods that consider temporal consistency, by smoothing posture predictions in a temporal sequence by post-processing results in an offline fashion (e.g Lauer et al. (2022); Joska et al. (2021)). For pigeons in the wild we obtain promising pose estimation results (cf. Tab. 3) without fine-tuning the model trained on captive data with data recorded in the wild. While the quantitative results show that the accuracy is not high, the qualitative results (cf. Fig. 6) highlight the potential of the approach for generalizing models to completely new environments. The model can be further improved by fine-tuning, using a small amount of annotated data recorded in outdoor settings. Finally, our current tracking approach relies on all subjects being present in the first frame for first frame correspondence matching, as well as all subjects staying in frame for the whole sequence. Future work can improve upon the tracking algorithm e.g by using visual features for re-identification (Wojke & Bewley, 2018; Ferreira et al., 2020).

## 5 Conclusion

In this work we present 3D-MuPPET, a framework to estimate 3D poses of multiple pigeons from a multi-view setup. We show that our framework allows complex poses and trajectories of multiple pigeons to be tracked reliably in 2D and 3D (cf. Tabs. 1 and 2) at interactive speeds with up to 10 fps in 2D and 1.5 fps in 3D. While our results



**Fig. 6** *Qualitative Results of Pigeons in the Wild.* Example frames for 3D multi-pigeon pose estimation and tracking in the wild, reprojected to 2D view. Notably, we did not fine-tune 3D-MuPPET on the data recorded in the wild, cf. Sec. 3.2.

are comparable to a state of the art 3D pose estimator in terms of RMSE and PCK, 3D-MuPPET achieves a faster inference speed for groups with five or more individuals. We also demonstrate that training a pose estimation module on single pigeon training data yields comparable results compared to a model trained on multi-pigeon data, highlighting the potential of a domain shift to new species in applications in the wild where using data that can be easier for researchers to annotate. Finally, we perform the first quantitative tracking evaluation on 3D-POP and obtain good results, cf. Tabs. 6 and 7. 3D-MuPPET is the first 3D pose estimation framework for more than two animals that also works with data recorded in the wild, cf. Sec. 3.2. While previous work (Bala et al., 2020; Han et al., 2023) has demonstrated 3D pose estimation for up to two animals, 3D-MuPPET shows that it is possible to track the 3D poses of up to 10 pigeons. We hope that our work thus leads to further systematic progress in developing automated quantitative methods in the study of animal collective behaviour.

## 6 Declarations

**Data availability.** Upon acceptance, the datasets generated during and/or analysed during the current study are available in the GitHub repository, <https://github.com/alexhang212/3D-MuPPET>.

### Funding and Competing Interests.

Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany’s Excellence Strategy – EXC 2117 – 422037984. All authors certify that they have no affiliations with or involvement in any organization or entity with any financial interest or non-financial interest in the subject matter or materials discussed in this manuscript.

**Acknowledgements.** We thank Lili Karashchuk for the valuable feedback and suggestions in CVPR 2023.

## References

- Altmann, J. (1974). Observational study of behavior: Sampling methods. *Behaviour*, 49(3-4), 227 - 266,
- Anderson, D., & Perona, P. (2014). Toward a science of computational ethology. *Neuron*, 84(1), 18-31,
- 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).
- Bala, P.C., Eisenreich, B.R., Yoo, S.B.M., Hayden, B.Y., Park, H.S., Zimmermann, J. (2020). Automated markerless pose estimation in freely moving macaques with openmonkeystudio. *Nat. Commun.*, 11, 4560,
- Berman, G.J. (2018). Measuring behavior across scales. *BMC Biol.*, 16(23), ,
- Bernardin, K., & Stiefelhagen, R. (2008). Evaluating multiple object tracking performance: the clear mot metrics. *EURASIP Journal on Image and Video Processing*, 2008, 1–10,
- Bernshtain, N. (1967). *The co-ordination and regulation of movements*. Pergamon Press.
- Bewley, A., Ge, Z., Ott, L., Ramos, F., Upcroft, B. (2016). Simple online and realtime tracking. *Ieee int. conf. image process.* (p. 3464-3468).
- Biggs, B., Roddick, T., Fitzgibbon, A., Cipolla, R. (2019). Creatures great and smal: Recovering the shape and motion of animals from video. *Proceedings of the asian conference on computer vision* (pp. 3–19).
- Bolaños, L.A., Xiao, D., Ford, N.L., LeDue, J.M., Gupta, P.K., Doeblei, C., ... Murphy, T.H. (2021). A three-dimensional virtual mouse generates synthetic training data for behavioral analysis. *Nat. Methods*, 18, 378–381,
- Bridgeman, L., Volino, M., Guillemaut, J.-Y., Hilton, A. (2019, June). Multi-person 3d pose estimation and tracking in sports. *Ieee conf. comput. vis. pattern recog. worksh.*
- Chen, X., Zhai, H., Liu, D., Li, W., Ding, C., Xie, Q., Han, H. (2020). Siambomb: A real-time ai-based system for home-cage animal tracking, segmentation and behavioral analysis. *Proceedings of the twenty-ninth international conference on international joint conferences on artificial intelligence* (pp. 5300–5302).
- Couzin, I.D., & Heins, C. (2022). Emerging technologies for behavioral research in changing environments. *Trends in Ecology & Evolution*, ,
- Dell, A.I., Bender, J.A., Branson, K., Couzin, I.D., de Polavieja, G.G., Noldus, L.P., ... Brose, U. (2014). Automated image-based tracking and its application in ecology. *Trends in Ecology & Evolution*, 29(7), 417-428,
- Dendorfer, P. (n.d.). *Motchallengeevalkit*. <https://github.com/dendorferpatrick/MOTChallengeEvalKit>.
- 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,
- Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. *Ieee conf. comput. vis. pattern recog.* (p. 248-255).
- Dunn, T.W., Marshall, J.D., Severson, K.S., Aldarondo, D.E., Hildebrand, D.G., Chettih, S.N., ... others (2021). Geometric deep learning enables 3d kinematic profiling across species and environments. *Nat. Methods*, 18(5), 564–573,
- Duporge, I., Isupova, O., Reece, S., Macdonald, D.W., Wang, T. (2021). Using very-high-resolution satellite imagery and deep learning to detect and count african elephants in heterogeneous landscapes. *Remote Sensing in Ecology and Conservation*, 7(3), 369-381,
- Ebrahimi, A.S., Orlowska-Feuer, P., Huang, Q., Zippo, A.G., Martial, F.P., Petersen, R.S., Storchi, R. (2023). Three-dimensional

- unsupervised probabilistic pose reconstruction (3d-upper) for freely moving animals. *Scientific Reports*, 13(1), 155,
- Ferreira, A.C., Silva, L.R., Renna, F., Brandl, H.B., Renault, J.P., Farine, D.R., ... Doutrelant, C. (2020). Deep learning-based methods for individual recognition in small birds. *Methods in Ecology and Evolution*, 11(9), 1072–1085,
- Ferrero, F.R., Bergomi, M.G., Heras, F.J., Hinz, R., de Polavieja, G.G., the Champalimaud Foundation. (2017). *idtracker.ai*. (<https://idtrackerai.readthedocs.io/en/latest>)
- Giebenhain, S., Waldmann, U., Johannsen, O., Goldluecke, B. (2022, December). Neural puppeteer: Keypoint-based neural rendering of dynamic shapes. *Proceedings of the asian conference on computer vision (accv)* (p. 2830-2847).
- Gomez-Marin, J.J., Alex an Paton, Kampff, A.R., Costa, R.M., Mainen, Z.F. (2014). Big behavioral data: psychology, ethology and the foundations of neuroscience. *Nat. Neurosci.*, 17, 1455–1462,
- Gosztolai, A., Günel, S., Lobato-Ríos, V., Pietro Abrate, M., Morales, D., Rhodin, H., ... Ramdy, P. (2021). Liftpose3d, a deep learning-based approach for transforming two-dimensional to three-dimensional poses in laboratory animals. *Nat. Methods*, 18, 975–981,
- Graving, J.M., Chae, D., Naik, H., Li, L., Koger, B., Costelloe, B.R., Couzin, I.D. (2019, oct). Deeposekit, a software toolkit for fast and robust animal pose estimation using deep learning. *eLife*, 8, e47994, doi: <https://doi.org/10.7554/eLife.47994>
- Günel, S., Rhodin, H., Morales, D., Campagnolo, J., Ramdy, P., Fua, P. (2019). Deepfly3d, a deep learning-based approach for 3d limb and appendage tracking in tethered, adult *Drosophila*. *eLife*, 8, e48571,
- Han, Y., Chen, K., Wang, Y., Liu, W., Wang, X., Liao, J., ... others (2023). Social behavior atlas: A computational framework for tracking and mapping 3d close interactions of free-moving animals. *bioRxiv*, 2023–03,
- He, K., Gkioxari, G., Dollar, P., Girshick, R. (2017). Mask r-cnn. *Int. conf. comput. vis.*
- He, K., Zhang, X., Ren, S., Sun, J. (2016). Deep residual learning for image recognition. *Ieee conf. comput. vis. pattern recog.*
- Heras, F.J.H., Romero-Ferrero, F., Hinz, R.C., de Polavieja, G.G. (2019). Deep attention networks reveal the rules of collective motion in zebrafish. *PLOS Computational Biology*, 15(9), 1-23,
- 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).
- Ionescu, C., Papava, D., Olaru, V., Sminchisescu, C. (2014). Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE Trans. Pattern Anal. Mach. Intell.*, 36(7), 1325–1339,
- Iskakov, K., Burkov, E., Lempitsky, V., Malkov, Y. (2019). Learnable triangulation of human pose. *Int. conf. comput. vis.*
- Itahara, A., & Kano, F. (2022). “corvid tracking studio”: A custom-built motion capture system to track head movements of corvids. *Japanese Journal of Animal Psychology*, 72(1), 1–16,
- Itahara, A., & Kano, F. (2023). Gaze tracking of large-billed crows (*corvus macrorhynchos*) in a motion-capture system. *bioRxiv*, , doi: <https://doi.org/10.1101/2023.08.10.552747>
- Jocher, G., Chaurasia, A., Qiu, J. (2023, jan). *Yolo by ultralytics*. Retrieved from <https://github.com/ultralytics/ultralytics>
- Jonathon Luiten, A.H. (2020). *Trackeval*. (<https://github.com/JonathonLuiten/TrackEval>.

- Joska, D., Clark, L., Muramatsu, N., Jericevich, R., Nicolls, F., Mathis, A., ... Patel, A. (2021). Acinosest: A 3d pose estimation dataset and baseline models for cheetahs in the wild. *2021 ieee international conference on robotics and automation (icra)* (p. 13901-13908).
- Kane, G.A., Lopes, G., Saunders, J.L., Mathis, A., Mathis, M.W. (2020). Real-time, low-latency closed-loop feedback using markerless posture tracking. *Elife*, 9, e61909,
- Kano, F., Naik, H., Keskin, G., Couzin, I.D., Nagy, M. (2022). Head-tracking of freely-behaving pigeons in a motion-capture system reveals the selective use of visual field regions. *Scientific Reports*, 12(1), 19113,
- Karashchuk, P., Rupp, K.L., Dickinson, E.S., Walling-Bell, S., Sanders, E., Azim, E., ... Tuthill, J.C. (2021). Anipose: A toolkit for robust markerless 3d pose estimation. *Cell Reports*, 36(13), 109730,
- Kays, R., Crofoot, M.C., Jetz, W., Wikelski, M. (2015). Terrestrial animal tracking as an eye on life and planet. *Science*, 348(6240), aaa2478,
- Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., ... others (2023). Segment anything. *arXiv preprint arXiv:2304.02643*, ,
- Koger, B., Deshpande, A., Kerby, J.T., Graving, J.M., Costelloe, B.R., Couzin, I.D. (2023). Quantifying the movement, behaviour and environmental context of group-living animals using drones and computer vision. *Journal of Animal Ecology*, ,
- Kuhn, H.W. (1955). The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2), 83–97,
- Labuguen, R., Matsumoto, J., Negrete, S.B., Nishimaru, H., Nishijo, H., Takada, M., ... Shibata, T. (2021). Macaquepose: A novel “in the wild” macaque monkey pose dataset for markerless motion capture. *Frontiers in Behavioral Neuroscience*, 14, 268,
- Lauer, J., Zhou, M., Ye, S., Menegas, W., Schneider, S., Nath, T., ... Mathis, A. (2022). Multi-animal pose estimation, identification and tracking with deeplabcut. *Nat. Methods*, 19, 496–504,
- 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).
- Lin, T.-Y., Dollar, P., Girshick, R., He, K., Hariharan, B., Belongie, S. (2017). Feature pyramid networks for object detection. *Ieee conf. comput. vis. pattern recog.*
- Luiten, J., Osep, A., Dendorfer, P., Torr, P., Geiger, A., Leal-Taixé, L., Leibe, B. (2021). Hota: A higher order metric for evaluating multi-object tracking. *Int. J. Comput. Vis.*, 129(2), 548–578,
- Marshall, J.D., Klibaite, U., Gellis, A., Aldarondo, D.E., Ölveczky, B.P., Dunn, T.W. (2021). The pair-r24m dataset for multi-animal 3d pose estimation. *bioRxiv*, 2021–11,
- 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,
- Miñano, S., Golodetz, S., Cavallari, T., Taylor, G.K. (2023). Through hawks’ eyes: synthetically reconstructing the visual field of a bird in flight. *International Journal of Computer Vision*, 131(6), 1497–1531,
- 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. (2021). *Xr for all: Closed-loop visual stimulation techniques for human and non-human animals* (Dissertation). Technische Universität München, München.

- Naik, H., Bastien, R., Navab, N., Couzin, I.D. (2020). Animals in virtual environments. *IEEE Transactions on Visualization and Computer Graphics*, 26(5), 2073–2083,
- 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).
- Nath, T., Mathis, A., Chen, A.C., Patel, A., Bethge, M., Mathis, M.W. (2019). Using deeplabcut for 3d markerless pose estimation across species and behaviors. *Nat. Protoc.*, 14, 2152–2176,
- Newell, A., Yang, K., Deng, J. (2016). Stacked hourglass networks for human pose estimation. *Computer vision-eccv 2016: 14th european conference, amsterdam, the netherlands, october 11-14, 2016, proceedings, part viii* 14 (pp. 483–499).
- Nourizonoz, A., Zimmermann, R., Ho, C.L.A., Pellat, S., Ormen, Y., Prévost-Solié, C., ... Huber, D. (2020). Etholoop: automated closed-loop neuroethology in naturalistic environments. *Nat. Methods*, 17, 1052–1059,
- Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., ... Chintala, S. (2019). Pytorch: An imperative style, high-performance deep learning library. *Adv. neural inform. process. syst.*
- Pedersen, M., Haurum, J.B., Bengtson, S.H., Moeslund, T.B. (2020). 3d-zef: A 3d zebrafish tracking benchmark dataset. *Ieee conf. comput. vis. pattern recog.*
- Pereira, T.D., Aldarondo, D.E., Willmore, L., Kislin, M., Wang, S.S.-H., Murthy, M., Shaeivitz, J.W. (2019). Fast animal pose estimation using deep neural networks. *Nat. Methods*, 16, 117–125,
- Pereira, T.D., Tabris, N., Matsliah, A., Turner, D.M., Li, J., Ravindranath, S., ... Murthy, M. (2022). Sleap: A deep learning system for multi-animal pose tracking. *Nat. Methods*, 19, 486–495,
- 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).
- Romero-Ferrero, F., Bergomi, M.G., Hinz, R.C., Heras, F.J.H., de Polavieja, G.G. (2019). idtracker.ai: tracking all individuals in small or large collectives of unmarked animals. *Nat. Methods*, 16, 179–182,
- Sun, J.J., Karashchuk, L., Dravid, A., Ryoo, S., Fereidooni, S., Tuthill, J.C., ... others (2023). Bkind-3d: Self-supervised 3d keypoint discovery from multi-view videos. *Ieee conf. comput. vis. pattern recog.* (pp. 9001–9010).
- Sun, K., Xiao, B., Liu, D., Wang, J. (2019). Deep high-resolution representation learning for human pose estimation. *Ieee conf. comput. vis. pattern recog.*
- Van Horn, G., Branson, S., Farrell, R., Haber, S., Barry, J., Ipeirotis, P., ... Belongie, S. (2015). Building a bird recognition app and large scale dataset with citizen scientists: The fine print in fine-grained dataset collection. *Ieee conf. comput. vis. pattern recog.*
- Waldmann, U., Bamberger, J., Johannsen, O., Deussen, O., Goldlücke, B. (2022). Improving unsupervised label propagation for pose tracking and video object segmentation. B. Andres, F. Bernard, D. Cremers, S. Frintrop, B. Goldlücke, & I. Ihrke (Eds.), *Pattern recognition* (pp. 230–245). Cham: Springer International Publishing.
- Waldmann, U., Naik, H., Máté, N., Kano, F., Couzin, I.D., Deussen, O., Goldlücke, B. (2022). I-muppet: Interactive multi-pigeon pose estimation and tracking. B. Andres, F. Bernard, D. Cremers, S. Frintrop, B. Goldlücke, & I. Ihrke (Eds.), *Pattern recognition* (pp. 513–528). Cham: Springer International Publishing.
- Walter, T., & Couzin, I.D. (2021). Trex, a fast multi-animal tracking system with markerless identification, and 2d estimation of posture and visual fields. *eLife*, 10, e64000,

- Wang, J., & Yuille, A.L. (2015). Semantic part segmentation using compositional model combining shape and appearance. *Ieee conf. comput. vis. pattern recog.*
- Wang, P., Shen, X., Lin, Z., Cohen, S., Price, B., Yuille, A.L. (2015). Joint object and part segmentation using deep learned potentials. *Int. conf. comput. vis.*
- Welinder, P., Branson, S., Mita, T., Wah, C., Schroff, F., Belongie, S., Perona, P. (2010). *Caltech-UCSD Birds 200* (Tech. Rep. No. CNS-TR-2010-001). California Institute of Technology.
- Wojke, N., & Bewley, A. (2018). Deep cosine metric learning for person re-identification. *2018 ieee winter conference on applications of computer vision (wacv)* (pp. 748–756).
- Xiao, B., Wu, H., Wei, Y. (2018). Simple baselines for human pose estimation and tracking. *Eur. conf. comput. vis.*
- Yang, Y., & Ramanan, D. (2013). Articulated human detection with flexible mixtures of parts. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(12), 2878-2890,
- Zhang, L., Gao, J., Xiao, Z., Fan, H. (2023). Animaltrack: A benchmark for multi-animal tracking in the wild. *International Journal of Computer Vision*, 131(2), 496–513,
- Zuffi, S., Rhodin, H., Park, H.S., Beery, S., Kanazawa, A., Nobuhara, S., Zamansky, A. (2023). *Cv4animals: Computer vision for animal behavior tracking and modeling.* (<https://www.cv4animals.com/>)