International Journal of Computer Vision (2024) 132:4235–4252
https://doi.org/10.1007/s11263-024-02074-y

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

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]

Received: 22 August 2023 / Accepted: 28 March 2024 / Published online: 7 May 2024

Abstract
Markerless methods for animal posture tracking have been rapidly 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 camera views. We train a pose estimator to infer 2D keypoints and bounding boxes of multiple pigeons, then triangulate the keypoints to 3D. For identity matching of individuals in all views, we first dynamically match 2D detections to global identities in the first frame, then use a 2D tracker to maintain IDs across views in subsequent frames. We achieve comparable accuracy to a state of the art 3D pose estimator in terms of median error and Percentage of Correct Keypoints. Additionally, we benchmark the inference speed of 3D-MuPPET, with up to 9.45 fps in 2D and 1.89 fps in 3D, and perform quantitative tracking evaluation, which yields encouraging results. Finally, we showcase two novel applications for 3D-MuPPET. First, we train a model with data of single pigeons and achieve comparable results in 2D and 3D posture estimation for up to 5 pigeons. Second, we show that 3D-MuPPET also works in outdoors without additional annotations from natural environments. Both use cases simplify the domain shift to new species and environments, largely reducing annotation effort needed for 3D posture tracking. To the best of our knowledge we are the first to present a framework for 2D/3D animal posture and trajectory tracking that works in both indoor and outdoor environments for up to 10 individuals. We hope that the framework can open up new opportunities in studying animal collective behaviour and encourages further developments in 3D multi-animal posture tracking.

Keywords 3D pose estimation · Multi-object tracking · Online methods · Animals · Lab data · Field data

Communicated by Anna Zamansky.

Urs Waldmann and Alex Hoi Hang Chan contributed equally to this work

✉ Urs Waldmann
urs.waldmann@uni-konstanz.de

✉ Alex Hoi Hang Chan
hoi-hang.chan@uni-konstanz.de

Hemal Naik
hnaik@ab.mpg.de

Máté Nagy
nagymate@hal.elte.hu

Iain D. Couzin
icouzin@ab.mpg.de

Oliver Deussen
Oliver.Deussen@uni-konstanz.de

Bastian Goldluecke
bastian.goldluecke@uni-konstanz.de

Fumihiro Kano
fumihiro.kano@uni-konstanz.de

[1] Department of Computer and Information Science, University of Konstanz, Konstanz, Germany

[2] Centre for the Advanced Study of Collective Behaviour, University of Konstanz, Konstanz, Germany

[3] Department of Collective Behavior, Max Planck Institute of Animal Behavior, Konstanz, Germany

[4] Department of Biology, University of Konstanz, 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

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 et al., 2019) to the study of collective behaviour in nonhuman animals (Couzin & Heins, 2023; 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 (Bernshtein, 1967; Altmann, 1974; Berman, 2018; Mathis et al., 2018; Kays et al., 2015). While researchers conventionally analyzed animal behaviour manually using a predefined catalogue of behaviours using ethograms, 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 et al., 2014; Anderson & Perona, 2014; Mathis et al., 2018). In such applications, multi-object pose estimation is essential to observe the dynamics of socially interacting animals 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 et al., 2022; Itahara & Kano, 2022; Miñano et al., 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 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 posture tracking and behavioural quantification 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 et al., 2022) and 3D postures (Günel et al., 2019; Joska et al., 2021; Dunn et al., 2021; Giebenhain et al., 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 et al., 2021), behavior analysis (Nourizonoz et al., 2020; Bolaños et al., 2021), object segmentation (Chen et al., 2020; Waldmann et al., 2022), 3D shape and pose fitting (Biggs et al., 2019; Badger et al., 2020) to pose estimation (Labuguen et al., 2021; Gosztolai et al., 2021; Waldmann et al., 2022) and tracking (Romero-Ferrero et al., 2019; Pedersen et al., 2020; Waldmann 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 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 three frameworks that estimate the 3D pose of more than one individual (two macaques Bala et al. (2020), two rats/parrots Han et al. (2023), and four/two pigs/dogs An et al. (2023)) in controlled captive environments, and finally one framework (Joska et al., 2021; Nath et al., 2019) 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 et al., 2014)), cf. also Sanakoyeu et al. (2020). 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 Wang and Yuille (2015), 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 not only because it is a common model species for animal collective behaviour (e.g Yomosa et al., 2015; Nagy et al., 2010; Nagy et al., 2013; Papadopoulou et al., 2022; Sasaki & Biro, 2017), but also because of the recent introduction of a large scale multi-

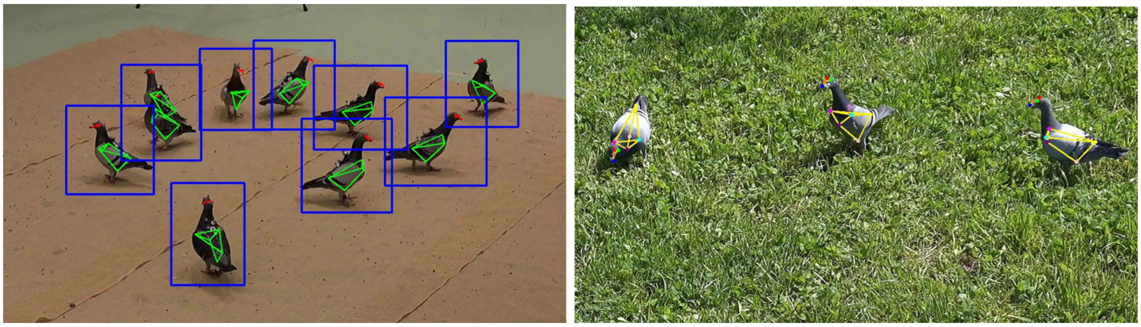

Fig. 1 3D Multi-Pigeon Pose Estimation and Tracking (3D-MuPPET) is a framework for multi-animal pose estimation and tracking for lab (left) and outdoor data (right). *Left*: estimated complex pose (beak, nose, left and right eye, left and right shoulder, top and bottom keel and tail) of pigeons recorded 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)

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 et al., 2022), a recent framework proposed for interactive 2D posture estimation and tracking of multiple pigeons, by incoporating multiple views to obtain 3D information. We will first evaluate the 2D framework proposed in I-MuPPET (Waldmann 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 flexible framework for interactive tracking and 3D pose estimation of multiple pigeons that works for data recorded both in captivity and the wild. The framework is based on triangulating 2D poses from multiple views to 3D, allowing 3D reconstruction if a 2D posture estimation model and a multiview setup is available. Compared to a state of the art 3D pose estimation method (Learnable Triangulation of Human Pose; LToHP, Iskakov et al. (2019)) that requires ground truth in 3D for training, 3D-MuPPET is less accurate (Root Mean Square Error; RMSE of 24.0 mm vs. 14.8 mm, and Percentage of Correct Keypoints; PCK05 of 71.0% vs. 76.7% for ours and LToHP respectively), but comparable in terms of median error (7.0 mm vs. 5.8 mm for LToHP) and Percentage of Correct Keypoints (PCK10 of 92.5% vs. 94.3% for LToHP). We track up to ten pigeons (the upper limit in Naik et al. (2023)) with up to 9.45 fps in 2D and 1.89 fps in 3D, and report detailed results for speed and accuracy. Additionally, we highlight two use cases that showcases the flexibility of our framework.

1. 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.
2. We demonstrate the ability to estimate 3D poses of pigeons recorded outdoors, cf. Fig. 1, without any additional annotations.

Both applications provide alternatives for the domain shift to other species or applications in the wild by reducing annotation effort required for multi-animal posture estimation.

Finally, to evaluate pose estimation from data recorded outdoors, we also present Wild-MuPPET, a novel 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 animal pose estimation framework for more than four individuals. Our approach is also not limited to pigeons and can be applied to any other species, given 2D posture annotations and a calibrated multi-camera system are available. In our supplemental material we also showcase the applicability to other species like mice from Mathis et al. (2018) and cowbirds from Badger et al. (2020) where 2D posture annotations from one camera view are available. The source code and data to reproduce the results of this paper are publicly available at https://alexhang212.github.io/3D-MuPPET/. We think that 3D-MuPPET offers a promising framework for automated 3D multi-animal pose estimation and identity tracking, opening up new ways for biologists to study animal collective behaviour in a fine-scaled way.

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 et al., 2020), most applications have focused on offline post-hoc analysis, which limits any application that might require posture estimation at interactive speeds ($\geq$ 1 fps) to perform stimulus driven behavior experiments e.g. VR for animals (Naik et al., 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 DANNCE, a method similar to Iskakov et al. (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. 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. An et al. (2023) fits a mesh model, which must be prepared for each species, to 10 camera views for 3D pose estimation of four pigs, two dogs and one mouse captured in indoor environments. 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 et al. (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 and trigulation from multiple views.

All these methods are limited to the pose tracking of up to four 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 et al., 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 et al. (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 et al., 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 et al. (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, explain how we extend the framework to two further use cases, and finally discuss ablation studies and network training.

3.1 Datasets

We describe the indoor dataset (Naik et al., 2023) and the additional datasets that we use for our two domain shifts including our novel outdoor pigeon dataset.

3.1.1 3D-POP

For this study, we use the 3D-POP dataset (Naik et al., 2023), a multi-view multi-individual dataset of freely-moving (i.e. walking) pigeons filmed by both RGB and motion-capture cameras. This dataset contains RGB video sequences from 4 views (4K, 3840 × 2160 px) of 1, 2, 5 and 10 pigeons. The ground truth provided by the dataset for each individual is a bounding box (on average 215 px wide and 218 px high in 2D), 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 were sampled from each sequence to avoid bias. As our validation set, we sample 3040 frames separately from the training set following the same sampling method. As our test set for posture estimation, we use 1000 frames, across four test sequences of different individual numbers (1, 2, 5, 10 pigeons), each 250 frames long. We choose temporal sequences as the test set to evaluate the complete 3D-MuPPET pipeline (cf. Fig. 2 and Sect. 3.2).

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

3.1.2 Additional Datasets

We also extend 3D-MuPPET in two applications of domain shifts of training a single individual model and tracking outdoors, which corresponds to two additional datasets. For discussion of the implementation of the two use cases, we refer to Sect. 3.3.

Single Pigeon Dataset To test if training a model on 1 pigeon can be used to track multiple pigeons, we sample a single pigeon training set from 3D-POP, using the same sampling method as the multiple pigeon dataset (cf. Sect. 3.1.1) but only from single pigeon sequences. The dataset contains 6006 and 3012 images for training and validation respectively. We use the same 1000 frames of test sequences (cf. Sect. 3.1.1) that contains both single and multi-individual data for quantitative evaluation.

Wild-MuPPET To evaluate tracking in the wild, we 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). We hope to mirror the 3D-POP setup to minimize the differences between the indoor and outdoor datasets, with the only difference being the outdoor environment.

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 (2000 frames in 2D). 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).

Finally, for additional network training and fine-tuning (cf. Sect. 5.2), we further separated the dataset into an 80%/20% train-test split, resulting in 100 3D test frames for evaluation.

For more details on data collection and calibration procedure used for the dataset, we refer to the supplementary information.

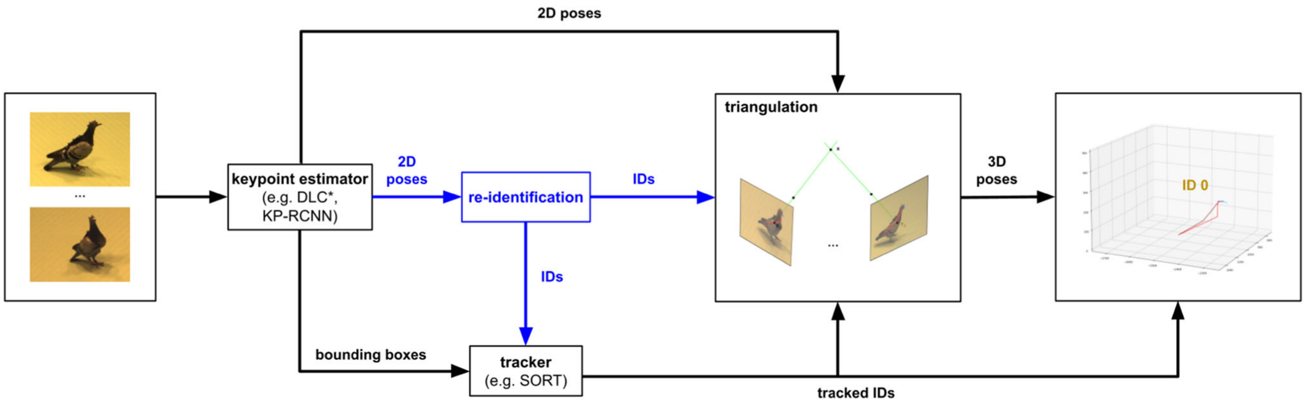

Fig. 2 3D-MuPPET. The framework consists of a pose estimation and tracking module, into which we can readily slot any state of the art pose estimator and tracking method. We identify all individuals in all views (blue part) based on Huang et al. (2020) in the first frame only. In the subsequent frames we track the identities (IDs) with SORT (Bewley et al., 2016)). 3D-MuPPET predicts 3D poses together with IDs from multi-view image input using triangulation. For details we refer to Sect. 3.2

3.2 Pose Estimation and Identity Tracking

This work extends upon I-MuPPET (Waldmann 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 three methods for comparison, i.e. a KeypointRCNN (He et al., 2017), a modified DeepLabCut (DLC, Mathis et al. (2018)) and a modified ViTPose (Xu et al., 2022). We choose DLC and ViTPose because they are state of the art frameworks for animal and human pose estimation respectively. The choice of the KeypointRCNN allows for the domain shift from single to multiple individuals, cf. Sect. 5.1. In addition KeypointRCNN achieves the fastest inference speed for multiple individuals (on average 7.5 fps and 1.76 fps for 2D and 3D poses respectively, cf. Tables 1 and 2 respectively). In this way we present options for the pose estimator module in terms of accuracy and speed, allowing researchers to choose based on their application.

For the modified DLC and ViTPose, we adopt a top-down approach, by first using YOLOv8 (Jocher et al., 2023) to detect the individual pigeons in each frame and then pass the cropped pigeon images into the single individual DLC (Mathis et al., 2018) and ViTPose (Xu et al., 2022) pipeline. For details, we refer to Mathis et al. (2018) and Xu et al. (2022). In the following, we denote these models by DLC* and ViTPose* (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 et al., 2016; Lin et al., 2017) backbone that was pre-trained on ImageNet (Deng et al., 2009). For details, we refer to He et al. (2017). The input to the KeypointRCNN are RGB images (cf. Figure 2) normalized to mean and standard deviation of 0.5.

3D Posture Estimation We use the 2D postures of all four camera views obtained from KeypointRCNN, DLC* and ViTPose* to acquire 3D keypoint estimates using triangulation with sparse bundle adjustment. 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 Kalman filter (Kalman, 1960) to smooth our pose estimates. In the following, we denote the three 3D-MuPPET posture estimation modules by 3D-KeypointRCNN, 3D-DLC* and 3D-ViTPose*.

3D Mutli-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 match each individual across views, 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). After the assignment, we maintain the identities based on SORT tracking in 2D. We choose to do identity matching in the first frame only to allow the whole framework to be used in an online manner.

The dynamic matching algorithm first generates 3D pose estimates for each possible pair of 2D poses, creating a large 3D pose subspace. Within the 3D pose subspace, we match 3D poses that are close together based on the Euclidean distance, and assign 2D poses that contribute to the matched 3D poses as the same individual. We match until the pairwise distance threshold of 200 mm is reached. Since the algorithm does not know the number of individuals in the scene, we choose a conservative threshold of 200 mm 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. For more details we refer to Huang et al. (2020). After the dynamic matching is completed, we maintain the global ID in subsequent frames and triangulate based on the 2D tracklets from SORT. Finally, if a 2D tracklet in a certain camera view is lost or switched, we skip the detections of the given camera.

3.3 Further Applications

Here, we discuss how we adapt our framework for the two use cases of training a single pigeon model and posture tracking outdoors.

Single to Multi-Animal Domain Shift Annotating frames of multiple individuals is often more labour intensive than labelling frames with a single animal. Here, we explore this idea by training a model using our single pigeon dataset (cf. Sect. 3.1.2). For trianing and evaluation, we use the same framework as for indoor posture tracking from Fig. 2 and Sect. 3.2. However, in our pose estimation module we use the KeypointRCNN because the YOLOv8 object detection model in DLC* and ViTPose* cannot reliably generalize to multiple pigeons when only trained on images of a single pigeon.

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. Sect. 3.1.1. But as an extra processing step, we remove the influence of the background by using the Segment-Anything-Model (SAM, Kirillov et al. (2023)), a model that allows objects in an image to be segmented based on a prompt of the object location (ground truth bounding box). We then train our framework to predict keypoints on masked images instead of crops from bounding boxes that contains both the object and background.

For the choice of pose estimator module, we train both ViTPose* and DLC* on the masked images because they perform similarly well on the 3D-POP dataset (cf. Table 2). To remove confusion from the 3D-POP benchmarking results, we refer to these 2 models as Wild-VitPose and Wild-DLC.

Finally, we evaluate the models on the 100 test frames of our novel Wild-MuPPET dataset, cf. Sect. 3.1.2. We first use a pre-trained MaskRCNN (He et al., 2017) to localize and segment all objects with the "bird" class in the frame and then pass them to the pose estimator. We do not use SAM during inference because it does not provide category labels. Unlike the evaluation on 3D-POP, we also do not perform any temporal filtering since the Wild-MuPPET test set only contains individually sampled frames.

3.4 Network Training and Ablation Studies

Data Augmentation In I-MuPPET (Waldmann 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. Sect. 3.2). This includes 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 ±5%.

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

And finally for ViTPose*, we also use the default augmentation implementation (Xu et al., 2022) for our animal posture tracking.

Training Hyperparameters To find out the best network configuration for KeypointRCNN (cf. Sect. 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. Sect. 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).

For ViTPose*, we use default hyperparamters and training configuration (Xu et al., 2022), with a custom learning rate of 0.00005.

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), and for ViTPose, we use the highest mean average precision (mAP) score.

This procedure can lead to a different number of training epochs in each experiment. Nevertheless experiments are comparable in the sense that each model is trained to perform best without over-fitting to the training data.

4 Evaluation

We evaluate each module of 3D-MuPPET on test sequences of the 3D-POP dataset. We separate our evaluation into three parts, to provide an idea of how each component of the framework performs. First, we evaluate keypoint estimation accuracy in Sect. 4.2. Second, we evaluate identity tracking accuracy and third, we evaluate inference speed. The latter two evaluations are both in Sect. 4.3. We first briefly discuss the evaluation metrics we use in Sect. 4.1, then report quantitative results on each of the components above. Finally, we also show qualitative results on all tasks.

Since the current framework extends the work of I-MuPPET (Waldmann et al., 2022), which relies on triangulating 2D posture estimates into 3D, in Sect. 4.2 we evaluate both 2D performance and 3D performance for all tasks to provide insights into how errors propagate.

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 et al. (2009) to measure track quality, and the trajectory-based metrics proposed in Ristani et al. (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 code provided by Luiten and Hoffhues (2020), Dendorfer (2020).

Inference Speed We also benchmark the inference speed of our framework in 2D and 3D with all 1000 frames in the test set from 3D-POP (Naik et al., 2023), cf. Sect. 3.1.1. For this evaluation, we use a workstation with a 16GB Nvidia Geforce RTX 3070 GPU, 11th Gen Intel(R) Core(TM) i9-11900 H @ 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). We consider the framework as interactive if the inference speed is $\geq$ 1 fps.

4.2 Pose Estimation

We report quantitative and qualitative results of 2D and 3D poses on the indoor pigeon data (cf. Sect. 3.1.1) and compare 3D-MuPPET to a 3D baseline based on 3D CNNs (Iskakov et al., 2019). Furthermore, to illustrate the applicability to other species, we also compare the KeypointRCNN (cf. Sect. 3.2) to DLC (Mathis et al., 2018) on their 2D odor trail tracking data and to 3D Bird Reconstruction (Badger et al., 2020) on their 2D cowbird keypoint dataset, both available in the supplementary materials.

3D Baseline For a 3D comparison, we train the “Learnable Triangulation of Human Pose” framework (LToHP, Iskakov et al. (2019)), on the same training dataset specified in Sect. 3.1.1. We perform this comparison because the framework is state of the art for human 3D posture estimation, and uses a 3D CNN architecture, which is shown to be more accurate than simple triangulation (Iskakov et al., 2019). With this comparison we can evaluate how well the triangulation based 3D-MuPPET performs, since models like LToHP rely on 3D ground truth datasets, which are rare in animal posture tracking.

The framework predicts a 2D heatmap from each view that is projected into a 3D voxel grid, then learns to predict 3D keypoints using a 3D CNN architecture. Since the model requires a 3D root point as input, we train both an algebraic and volumetric triangulation model by providing cropped images of pigeon individuals based on ground truth bounding boxes. During inference, we follow the same workflow as in Iskakov et al. (2019) by first obtaining a root point estimate (top keel) using the algebraic model, then run the volumetric model to obtain 3D keypoint estimates. We refer to Iskakov et al. (2019) for more details.

Table 1 Quantitative evaluation of 2D pigeon poses

Metric/Method	KP-RCNN	DLC*	ViTPose*
RMSE $(px)\downarrow$	**28.1**	39.0	38.9
Median $(px)\downarrow$	5.7	4.7	**4.4**
PCK05 (%) ↑	82.4	89.1	**91.1**
PCK10 (%) ↑	95.4	**96.8**	**96.8**
Mean Speed (fps) ↑	**7.5**	3.0	2.1

We report the RMSE and its median (px), PCK05 (%) and PCK10 (%) for estimated 2D poses on the 3D-POP test sequences. Comparison between KeypointRCNN (KP-RCNN, cf. Sect. 3.2), modified DeepLabCut (DLC*) and modified ViTPose (ViTPose*). *: We combine YOLOv8 (Jocher et al., 2023) for instance detection with single-object DLC (Mathis et al., 2018) and ViTPose (Xu et al., 2022). We also report the mean 2D inference speed for the complete pipelines in fps. For details on the inference speed we refer to Sect. 4.3. Upwards and downwards arrows represent whether a higher or lower value is better, respectively. Best results per row in bold

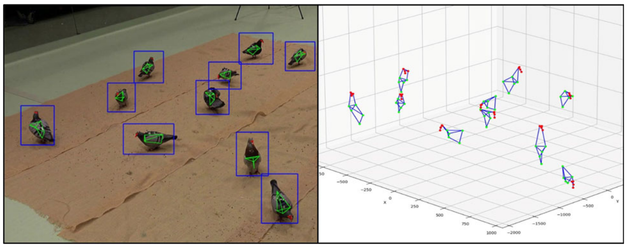

Fig. 3 Qualitative 3D results. Example frame from 3D-POP, cf. Sect. 3.1.1. 2D (left side) and 3D (right side) pose estimates using 3D-MuPPET

For model training, we train the algebraic model for 292 epochs and the volumetric model for 782 epochs with default augmentation parameters, both models having lowest validation loss.

Finally, note that since LToHP is a single subject framework, we make use of ground truth bounding boxes to crop the image inputs during training and inference, with the goal of providing a baseline for 3D posture estimation accuracy, but not as a complete pipeline. Implementing a complete pipeline for multi-animal 3D CNN based posture estimation is outside the scope of this study, and can be a further application of 3D-MuPPET, where it can replace the algebraic model together with the ground truth bounding boxes by providing root point estimate, bounding boxes and identities to the volumetric model of LToHP.

Results We train the different posture estimation modules of 3D-MuPPET on multi-pigeon data from Naik et al. (2023), cf. Sect. 3.1.1, and choose the best weights with the lowest validation loss. We train the KeypointRCNN (cf. Sect. 3.2) for 44 epochs. In the case of DLC* and ViTPose* (cf. Sect. 3.2), we train YOLOv8 (Jocher et al., 2023) for 27 epochs, ViTPose (Xu et al., 2022) for 175 epochs and DLC (Mathis et al., 2018) for 86000 iterations.

Quantitative results for 2D pose estimation are in Table 1. We find that ViTPose* performs best across most metrics like median error (4.4 px) and PCK (PCK05 91.1%, PCK10 96.8%). When a more generous threshold is considered in PCK10, both DLC* and ViTPose* are equally accurate (PCK10 96.8%). KP-RCNN has the lowest RMSE, likely due to reduced outliers since the RMSE metric is quite sensitive to large outliers which is also reflected in a relatively small median error compared to RMSE (RMSE 28.1 px, median 5.7 px). This difference is likely due to bounding box detection errors in the YOLOv8 model within DLC* and ViTPose*.

For 3D, when comparing between models in the posture estimation module of 3D-MuPPET, 3D-ViTPose* performs the best across all evaluation metrics with a RMSE of 24.0 mm, its median of 7.0 mm, PCK05 of 71.0% and PCK10 of 92.5%, cf. Table 2. This is not surprising since ViTPose* already performs the best in 2D, cf. Table 1, and shows that 2D accuracy propagates into 3D.

We conclude that in applications where high accuracy is needed, researchers should prefer 3D-ViTPose* for the pose estimation module of 3D-MuPPET.

Comparing 3D-MuPPET with the 3D baseline in LToHP (Iskakov et al., 2019), we find that LToHP has the best performance across all metrics with a RMSE of 14.8 mm, its median of 5.8 mm, PCK05 of 76.7% and PCK10 of 94.3%, cf. Table 2. One of the reasons is that the bounding boxes of the subjects are provided from the ground truth for LToHP, removing the reliance on 2D and 3D multi-animal identity tracking. In addition, the model can also learn the general 3D structure of a pigeon instead of relying on 2D detection and triangulation.

Nevertheless, we show that 3D-MuPPET produces comparable estimates compared to LToHP (cf. Figs. 3 and 4), given a median difference of only 1.2 mm between the best model in 3D-MuPPET (3D-ViTPose*) and LToHP, cf. Table 2. This difference in error is very small in the context of keypoints on a pigeon, and will likely not affect any downstream behavioural experiments. For example, the diameter of the eye of a pigeon is on average around $10-13$ mm (Chard & Gundlach, 1938), which is much larger than the difference between the model estimates.

4.3 Tracking Performance

Figures 4 and 5 show results of the 3D pose estimation and tracking task for multiple pigeons. Further qualitative results can be found in our supplementary video at 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. Sect. 3.1.1. 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

Table 2 Quantitative evaluation of 3D pigeon poses

Metric/Method	3D-KP-RCNN	3D-DLC*	3D-ViTPose*	LToHP (Iskakov et al., 2019)
RMSE (*mm*) ↓	25.0	25.0	24.0	**14.8**
Median (*mm*) ↓	9.4	7.5	7.0	**5.8**
PCK05 (%) ↑	53.2	66.1	71.0	**76.7**
PCK10 (%) ↑	85.4	90.9	92.5	**94.3**
Mean Speed (fps) ↑	**1.76**	0.72	0.51	0.38

We report the filtered (cf. Sect. 3.2) RMSE and its median (*mm*), PCK05 (%) and PCK10 (%) for the 3D poses on the 3D-POP test sequences. Comparison between LToHP (Iskakov et al., 2019) and 3D-MuPPET (highlighted in gray). *: We combine YOLOv8 (Jocher et al., 2023) for instance detection with single-object DLC (Mathis et al., 2018) and ViTPose (Xu et al., 2022). We also report the mean 3D inference speed for the complete pipeline in fps. For details on the inference speed we refer to Sect. 4.3. Upwards and downwards arrows represent whether a higher or lower value is better, respectively. Best results per row in bold. See text for a discussion of the results

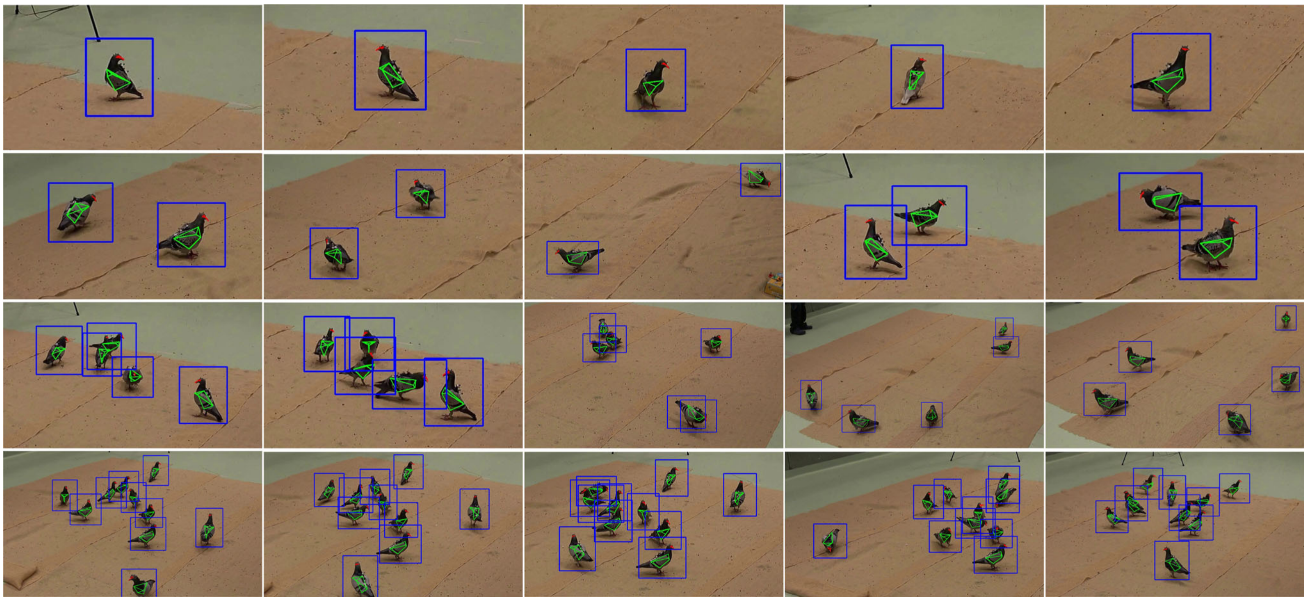

Fig. 4 Qualitative results. Example frames from 3D-POP (Naik et al., 2023) for multi-pigeon pose estimation and tracking in 3D, reprojected to 2D view. Green lines connect the body, red lines the head keypoints. Some frames are cropped for a better view

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 ViTPose* (cf. Sect. 3.2; the most accurate model from Sect. 4.2) and the metrics specified in Sect. 4.1. Note that for sequence 59, we remove the first 3 s (90 frames) since 2 pigeons are initially outside the tracking volume which causes the first frame identity matching (see Sect. 3.2) to fail.

Detailed 2D results for a detection confidence threshold of 0.5 are shown in Table 3. Overall, we achieve good results with our framework on the 2D video sequences including a HOTA of 86%, 98% multi-object tracking accuracy (MOTA), 90% multi-object tracking precision (MOTP), a recall of 98%, 99% precision, 99% mostly tracked (MT), and 0% mostly lost (ML) trajectories, 0.08 false positives per frame (FPF), and a IDF1 of 94% (metrics specified in Sect. 4.1 and our supplemental material).

In Table 4 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 (2020) to 30 mm. We choose 30 mm as this threshold is well within the body size of a pigeon, while taking into account the possible distance an individual can move within one frame. Overall, we achieve good 3D results with 3D-MuPPET including 85% multi-object tracking accuracy (MOTA), 90% mostly tracked (MT), and 0% mostly lost (ML) trajectories (metrics specified in Sect. 4.1 and our supplemental material).

Inference Speed Finally, we benchmark the inference speed of the pipeline, and we show that 3D-MuPPET can estimate 2D and 3D postures at interactive speeds (defined by ≥ 1 fps). Tables 5 and 7 provide detailed inference speed estimates for different numbers of individuals for 2D and 3D respectively, and we see that inference speed decreases

Table 3 Quantitative tracking evaluation in 2D

Test seq	HOTA↑	MOTA↑	MOTP↑	Rcll↑	Prcn↑	MT↑	ML↓	FPF↓	IDS↓	Frag↓	IDF1↑
11, view 1	0.82	0.92	0.90	0.96	0.96	0.90	0	0.39	2	14	0.92
11, view 2	0.84	0.92	0.88	0.96	0.96	0.90	0	0.41	0	7	0.96
11, view 3	0.84	0.92	0.89	0.96	0.96	0.90	0	0.41	0	11	0.96
11, view 4	0.85	0.94	0.90	0.97	0.97	1	0	0.26	3	29	0.95
19, view 1	0.90	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	0.92	1	0.91	1	1	1	0	0	0	4	1
19, view 4	0.89	0.99	0.92	0.99	1	1	0	0	4	11	0.94
30, view 1	0.83	0.96	0.92	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	8	15	0.96
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	0.87	1	0.89	1	1	1	0	0	1	14	0.96
48, view 2	0.90	1	0.90	1	1	1	0	0.02	0	6	1
48, view 3	0.91	1	0.91	1	1	1	0	0	0	4	1
48, view 4	0.91	1	0.90	1	1	1	0	0	0	3	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
Combined	0.86	0.98	0.90	0.98	0.99	0.99	0	0.08	73	318	0.94

We test 20 video sequences quantitatively with the metrics specified in Sect. 4.1 and our supplementary materials. Upwards and downwards arrows represent whether a higher or lower value is better, respectively. The threshold for the confidence score of ViTPose* (cf. Sect. 3.2) is set to 0.5

Table 4 Quantitative tracking evaluation in 3D

Seq	MOTA↑	MT↑	ML↓	IDS↓	Frag↓
11	0.92	1	0	0	173
19	0.89	0.90	0	0	214
30	0.92	1	0	0	225
48	0.93	1	0	0	245
59	0.57	0.60	0	8	334
Combined	0.85	0.90	0	8	1191

We test five sequences quantitatively with the metrics specified in Sect. 4.1. For detailed explanations on abbreviations and metrics, please refer to our supplemental material. Upwards and downwards arrows represent whether a higher or lower value is better, respectively. See text for a discussion of the results

Table 5 2D inference speed

Method/Num. of Ind	1	2	5	10
KP-RCNN	**7.50**	**8.07**	**8.02**	**6.22**
DLC*	3.85	3.42	2.75	2.03
ViTPose*	3.22	2.61	1.57	0.99

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

Table 6 2D inference speed

Batch size	Frame rate [fps]			
	1 pigeon	2 pigeons	5 pigeons	10 pigeons
1	8.24	8.13	8.03	6.77
2	8.70	8.54	8.27	6.91
4	8.90	8.81	8.43	7.09
8	9.10	8.96	8.61	7.17
16	**9.45**	**9.29**	**8.88**	**7.29**

Benchmark for our in-memory pipeline using the KeypointRCNN, cf. Sect. 3.2. We benchmark our pipeline with our video sequences preloaded in memory and report values for different batch sizes
Best results per column in bold

Table 7 3D inference speed

Method/Num. of Ind	1	2	5	10
3D-KP-RCNN	***1.89***	***1.84***	***1.73***	***1.59***
*3D-DLC**	*0.93*	*0.84*	*0.65*	*0.47*
*3D-ViTPose**	*0.79*	*0.64*	*0.38*	*0.24*
LToHP (cf. Sect. 4.2)	0.83	0.44	0.17	0.08

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 italic, 3D-MuPPET versions highlighted in italic. See text for a discussion of the results with increasing number of individuals across all models (at most by 2.23 fps for ViTPose* in 2D, cf. Table 5, and 0.75 fps for LToHP in 3D, cf. Table 7). Overall, we see that the mean inference speed is the fastest for the KeypointRCNN, reaching 7.5 fps in 2D and 1.76 fps in 3D, cf. Tables 1 and 2 respectively.

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 Table 6 for results. We report values for batch sizes up to 16, restricted by the hardware that we use, cf. Sect. 4.1. The maximum speed is at a batch size of 16 with an interactive speed of about $7-9$ fps depending on the number of pigeons present in the video sequence.

We conclude that researchers prioritizing inference speed for multi-animal posture estimation and tracking may consider the KeypointRCNN for the pose estimation module in 3D-MuPPET.

The speed evaluation shows that our pipeline can potentially be applied to closed-loop experiments (see Naik (2021)), based on the requirements of the researcher. For example, if an experiment requires general position and orientation of pigeons in closed-loop, inference speeds of 1.76 fps (cf. Table 2; can be pushed even further by preloading the data in memory and processing batches, cf. Table 6) might be sufficient. However, we do note that the current inference speed estimates do not include video acquisition time, so researchers considering such applications will need to develop a multi-view video acquisition framework independently.

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, cf. Sect. 4.1. A rough comparison yields that SLEAP (Pereira et al., 2022) is about an order of magnitude faster than the KeypointRCNN (SLEAP up to $\sim$ 800 fps; numbers read off from Pereira et al. (2022), Figs. 2b, 3e and Extended Data Fig. 6c). Considering the fact that the image resolution provided in 3D-POP is higher than the one of the flies and mice (3840 × 2160 px vs. 1280 × 1024 px) 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, we hope 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. Sect. 3.2 and Fig. 2) to achieve frame rates comparable to SLEAP.

Table 8 Quantitative results for our single to multi-aninal domain shift

Metric/Num. of Ind	1	2	5	10
2D				
RMSE (px) ↓	8.6	20.1	57.2	272.5
Median (px) ↓	4.3	6.0	7.7	17.9
PCK05 (%) ↑	90.5	76.9	66.7	42.9
PCK10 (%) ↑	98.7	93.4	83.6	53.9
3D				
RMSE (mm) ↓	11.1	26.9	93.2	434.3
Median (mm) ↓	6.9	6.0	15.4	246.7
PCK05 (%) ↑	70.4	54.1	30.2	11.4
PCK10 (%) ↑	94.9	82.4	60.3	19.7

We report RMSE and its median (px and mm in 2D and 3D respectively), PCK05 (%) and PCK10 (%) for estimated 2D and 3D posture from the 3D-POP dataset using the KeypointRCNN trained with single pigeon data. Upwards and downwards arrows represent whether a higher or lower value is better, respectively. We report results for sequences containing different number of individuals (1, 2, 5, and 10), cf. Sect. 3.1.2

5 Applications

We showcase the flexibility of 3D-MuPPET by presenting two domain shifts. First we show that 3D-MuPPET can be trained on annotated data that contains only single individuals and applied to multi-animal data which can reduce the annotation effort needed for new species or experimental setups (also see our supplemental material for 2D single mouse and cowbird pose estimation). Secondly, we show that 3D-MuPPET is robust to an indoor to outdoor environment domain shift by applying a model trained on indoor data to data from outdoors without further fine-tuning.

5.1 Single to Multi-animal Domain Shift

We train the KeypointRCNN (cf. Sect. 3.3) for 30 epochs on the single-pigeon dataset specified in Sect. 3.1.2. Results can be found in Table 8, showing difference in error across different number of individuals.

Overall, the single pigeon model performs well in 2D, but not as well in 3D, with the model not being able to generalize for 3D tracking of 10 pigeons. For sequences with 1 and 2 individuals, the performance is similar to using a multi-animal dataset for both 2D and 3D (cf. Tables 1, 2 and 8). For example, when comparing results of 2 individuals using the single pigeon model (Table 8) with the KeypointRCNN trained with multi-pigeon data (averaged over 1, 2, 5, 10 individuals, Table 2), we achieve a RMSE for the single-

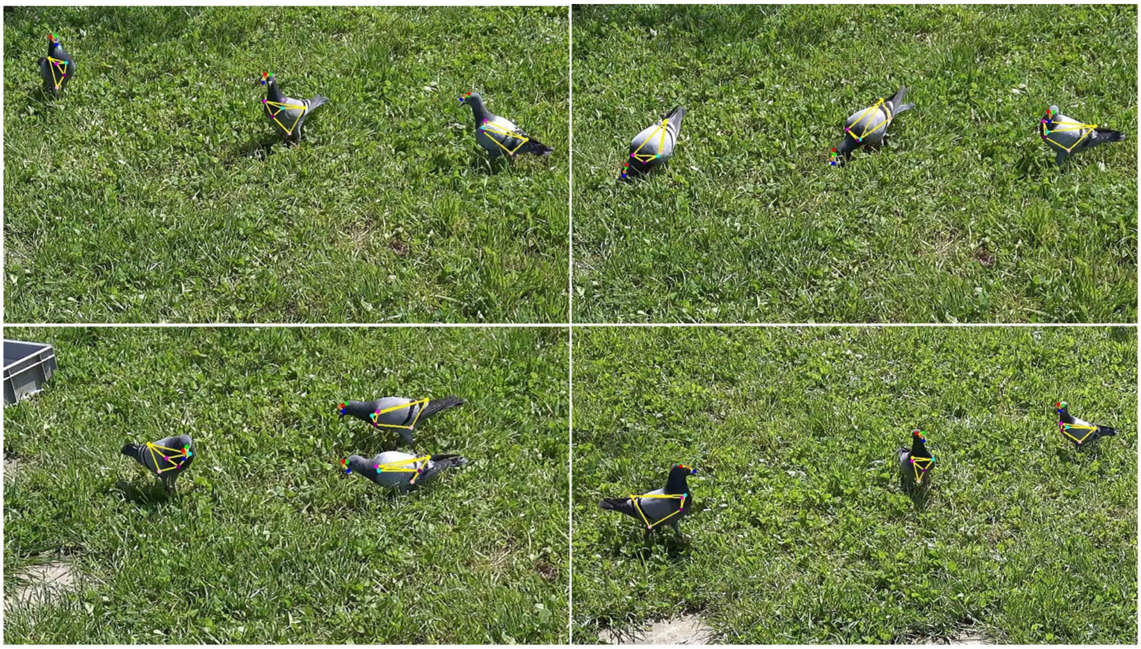

Fig. 5 Qualitative results of pigeons in the wild. Example frames for 3D multi-pigeon pose estimation and tracking in the wild, reprojected to 2D view. Notably, we did not fine-tune 3D-MuPPET (here with Wild-DLC) on our novel Wild-MuPPET data recorded in the wild, cf. Sect. 3.3

pigeon model of 26.9 mm vs. multi-pigeon of 25.0 mm with a median for single-pigeon of 6.0 mm vs. multi-pigeon of 9.4 mm.

For sequences with 5 and 10 individuals, performance differs. In 2D, we observe outliers as evident from the large RMSE values (5 individuals: 57.2 px, 10 individuals: 272.5 px), but from the median and PCK values from the multi-pigeon model (median of 5.7 px, PCK10 of 95.4%), the single-pigeon model show comparable accuracy for 5 individuals (median of 7.7 px, PCK10 of 83.6%), and good accuracy for 10 individuals (median of 17.9 px, PCK10 of 53.9%).

For 3D posture estimation, we expect accuracy to propagate from 2D estimation errors, as shown in the multi-animal model evaluation (cf. Tables 1 and 2), but we show that while 3D error is still low at 15.4 mm (median error) for 5 individuals, the model fails to generalize in 3D for 10 pigeons (median of 246.7 mm).

We think there are two main reasons that the model fails to generalize to 10 pigeons. Firstly, the detection of the bird individuals is less robust with the single pigeon model, where 10 pigeons are not always detected from all frames, and can affect the first frame identity matching and subsequent 2D tracking in the 3D-MuPPET pipeline. So, an incorrect 2D tracklet in one view can already increase the 3D error while additional ID switches in further camera views further deteriorate the 3D accuracy. This is reflected in Table 8 where the median error is $\sim 16\times$ higher for 10 compared to 5 individuals in 3D while it is "only" $\sim 2\times$ higher in 2D; the 2D errors from different views potentiate in 3D. Another reason is occlusions, where the model struggles to predict keypoints when the objects are too occluded, which is often the case in the 10 pigeon sequences. This shortcoming is also expected since the model was only trained on single pigeon data.

Nevertheless, we highlight that training a model with only single pigeon data can allow 2D and 3D posture estimation of up to 5 pigeons, which can simplify the domain shift to new species or systems, because annotating single animal data is less labour intensive than multi-animal annotations.

While less reliable in 3D, we show that the single-pigeon can predict keypoints in 2D reliably, so if researchers wish to annotate multi-individual data, the single-individual model can also be used as a pre-labelling tool. This can further reduce annotation time by first predicting keypoints from the 2D frame and manually correcting faulty detections, similar to methodologies provided in Pereira et al. (2022), Mathis et al. (2018), Graving et al. (2019).

5.2 Pigeons in the Wild

We train the Wild-ViTPose model for 124 epochs and Wild-DLC for 93000 iterations.

In Table 9 we report quantitative results on the test set of our novel Wild-MuPPET dataset. We first show that Wild-ViTPose (ViTPose* is the most accurate model in Sect. 4.2) does not generalize well for pigeons in the wild, compared to Wild-DLC, likely due to differences in augmentation parameters (median error of 146.0 mm and 15.0 mm respectively).

Table 9 Quantitative evaluation of 3D pigeon poses in our novel Wild-MuPPET dataset

Metric/Method	Wild-ViTPose	Wild-DLC	DLC-Fine-tuned	DLC-Scratch
RMSE (*mm*) ↓	166.0	53.4	58.2	**45.0**
Median (*mm*) ↓	146.0	15.0	**11.4**	12.7
PCK05 (%) ↑	0	25.1	**44.7**	40.1
PCK10 (%) ↑	0.2	74.4	**81.3**	77.4

We report RMSE and its median (*mm*), PCK05 (%) and PCK10 (%) for the 3D poses of pigeons in the wild, on the 100 test frames in the Wild-MuPPET dataset cf. Sect. 3.1.2. Wild-ViTPose and Wild-DLC are models trained on masked images from 3D-POP using ViTPose (Xu et al., 2022) and DLC (Mathis et al., 2018) respectively, without additional annotations from the wild. DLC-Fine-tuned and DLC-Scratch are trained on sampled images from Wild-MuPPET training set (cf. Sect. 3.1.2), with DLC-Fine-tuned using Wild-DLC as initial weights. See text for a discussion of the results

Best results per row in bold

However, for Wild-DLC, we show that the model performs well on Wild-MuPPET, with a median accuracy of 15.0 mm, only with training data of pigeons indoors, cf. Sect. 3.1.2. Additionally, we also use Wild-DLC for inference in a 3 pigeon sequence in the wild, which reflects our promising quantitative results, cf. Fig. 5 and supplementary video.

To further explore how a model trained on pigeons indoors can aid the domain shift to the wild, we also fine-tune the Wild-DLC model (named DLC-Fine-tuned) using sampled 2D frames from the training set of Wild-MuPPET (see cf. Sect. 3.1.2). To compare whether initializing model weights using data of pigeons indoors can lead to better accuracy in the wild, we also trained a DLC model from scratch, without fine-tuning (named DLC-Scratch), using the same outdoor image dataset, cf. Sect. 3.1.2. Fine-tuning takes 61000 iterations, and training from scratch takes 99000 iterations to reach lowest validation loss.

We show that both fine-tuning and training from scratch improves the performance of Wild-DLC (cf. Table 9), and both methods yield comparable accuracy. However, the fine-tuned model performs slightly better than the model trained from scratch (median of 11.4 mm vs. 12.7 mm respectively). Finally, we note that while keypoint estimation accuracy in the latter two cases is comparable, fine-tuning requires less iterations for the model to converge, allowing reduced training time for domain shifts across datasets.

All together, our two applications show that 3D-MuPPET is flexible and robust, promising to open up new ways for biologists to study animal collective behaviour in a fine-scaled way with multi-animal 3D posture tracking.

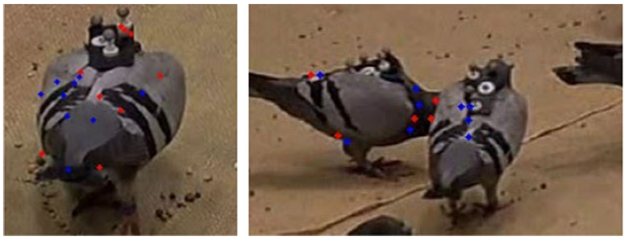

Fig. 6 Limitations. Cropped frames of failure cases from 3D-POP (Naik et al., 2023) data for 2D pose estimation using the KeypointRCNN (cf. Sect. 3.2), due to occlusions. Blue denotes the ground truth, red denotes the prediction

6 Limitations and Future Work

Keypoint detection can fail e.g. due to self-occlusions or occlusions from other individuals (cf. Fig. 6), which can affect the triangulation procedure. This may have caused outliers present in 2D and 3D keypoint evaluation, as indicated by the high RMSE values in contrast to their median errors. While we use a Kalman filter to smooth 3D posture estimates, the method can fail when there are multiple consecutive frames of error. Other filtering and smoothing methods that consider temporal consistency in an offline fashion can alleviate this problem if online processing is not required (e.g. Lauer et al. (2022), Joska et al. (2021)).

For pigeons in the wild, we limit the pigeon segmentation to Kirillov et al. (2023) and He et al. (2017) and the tracking to Bewley et al. (2016)), other methods available like Bekuzarov et al. (2023) and Yang et al. (2023) for segmentation and Karaev et al. (2023) for tracking might boost the performance. Using another tracker might also boost our single to multi-animal domain shift when dealing with 10 individuals.

Finally, our current tracking approach relies on all subjects being present in the first frame for first frame re-identification, 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; Waldmann et al., 2023).

7 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. Tables 1 and 2) at interactive speeds with up to 9.45 fps in 2D and 1.89 fps in 3D. While our results are comparable to a state of the art 3D pose estimator in terms of median error and Percentage of Correct Keypoints, cf. Table 2, 3D-MuPPET achieves a faster inference speed, cf. Tables 5 and 7, and only relies on training a 2D posture estimation model. Additionally, we perform the first quantitative tracking evaluation on 3D-POP and obtain good results, cf. Tables 3 and 4.

In applications where a higher accuracy is needed, researchers should prefer 3D-ViTPose* for the pose estimation module of 3D-MuPPET, cf. Fig. 2. Researchers that prioritize inference speed for multi-animal posture estimation and tracking or are interested in the single to multi-animal domain shift may consider the KeypointRCNN for the pose estimation module in 3D-MuPPET.

Finally, we demonstrate that training a pose estimation module on single pigeon training data yields comparable results compared to a model trained on multi-pigeon data for up to 5 pigeons (cf. Sect. 5.1), as well as showing that a model trained with indoor data can be generalized to data in the wild, cf. Sect. 5.2. This highlights the potential of a domain shift to new species and environments without the need for laborious manual annotation.

3D-MuPPET is the first 3D pose estimation framework for more than four animals that also works with data recorded in the wild, cf. Sect. 3.2. While previous work (Bala et al., 2020; Han et al., 2023; An et al., 2023) has demonstrated 3D pose estimation for up to four animals, 3D-MuPPET shows that it is possible to track the 3D poses of up to 10 pigeons if a 2D posture estimation model and a multi-camera setup is available. Our work offers a promising and flexible framework opening up new ways for biologists to study animal collective behaviour and we hope that this leads to further systematic progress in the field.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11263-024-02074-y.

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

Funding Open Access funding enabled and organized by Projekt DEAL.Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC 2117 – 422037984, and the Federal Ministry of Education and Research (BMBF) within the research program KI4KMU under grant number 01IS23046B (ARGUS). I.C. acknowledges the Office of Naval Research Grant N00014-19-1-2556, Horizon Europe Marie Sklodowska-Curie Actions (860949) and the PathFinder European Innovation Council Work Programme 101098722. M.N. acknowledges the Hungarian Academy of Sciences (grant no. 95152), The Hungarian National Research, Development and Innovation Office (grant no. K 128780), and the Isaac Newton Institute for Mathematical Sciences (EPSRC Grant Number EP/R014604/1).

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

Declarations

Conflict of interest 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.

References

Altmann, J. (1974). Observational study of behavior: Sampling methods. *Behaviour, 49*(3–4), 227–266.

An, L., Ren, J., Yu, T., Hai, T., Jia, Y., & Liu, Y. (2023). Three-dimensional surface motion capture of multiple freely moving pigs using mammal. *Nature Communications, 14*(1), 7727.

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. In *European conference on computer vision* (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. *Nature Communication, 11*, 4560.

Bekuzarov, M. , Bermudez, A. , Lee, J.- Y. , & Li, H. (2023 October). Xmem++: Production-level video segmentation from few annotated frames. In *Proceedings of the IEEE/CVF international conference on computer vision (iccv)* (pp. 635–644).

Berman, G. J. (2018). Measuring behavior across scales. *BMC Biology* **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.

Bernshtein, 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. In *IEEE international conference on image processing*. (pp. 3464–3468).

Biggs, B. , Roddick, T. , Fitzgibbon, A. , & Cipolla, R. (2019). Creatures great and smal: Recovering the shape and motion of animals from video. In *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., Doebeli, C., & Murphy, T. H. (2021). A three-dimensional virtual mouse generates synthetic training data for behavioral analysis. *Nature Methods, 18*, 378–381.

Bridgeman, L. , Volino, M. , Guillemaut, J.- Y. , & Hilton, A. (2019). Multi-person 3d pose estimation and tracking in sports. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*.

Chard, R. D., & Gundlach, R. H. (1938). The structure of the eye of the homing pigeon. *Journal of Comparative Psychology, 25*(2), 249.

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. In *Proceedings of the twenty-ninth international conference on international joint conferences on artificial intelligence* (pp. 5300–5302).

Couzin, I.D. , & Heins, C. (2023). 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. (2020). 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. *International Journal of Computer Vision, 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. In *2009 IEEE conference on computer vision and pattern recognition* (pp. 248-255).

Dunn, T. W., Marshall, J. D., Severson, K. S., Aldarondo, D. E., Hildebrand, D. G., Chettih, S. N., et al. (2021). Geometric deep learning enables 3d kinematic profiling across species and environments. *Nature 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., Renoult, 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). Neural puppeteer: Keypoint-based neural rendering of dynamic shapes. In *Proceedings of the Asian conference on computer vision (ACCV)* (pp. 2830–2847).

Gomez-Marin, A., Paton, J., Kampff, A. R., Costa, R. M., & Mainen, Z. F. (2014). Big behavioral data: Psychology, ethology and the foundations of neuroscience. *Nature Neuroscience, 17*, 1455–1462.

Gosztolai, A., Günel, S., Lobato-Ríos, V., Pietro Abrate, M., Morales, D., Rhodin, H., & Ramdya, P. (2021). Liftpose3d, a deep learning-based approach for transforming two-dimensional to three-dimensional poses in laboratory animals. *Nature Methods, 18*, 975–981.

Graving, J.M. , Chae, D. , Naik, H. , Li, L. , Koger, B. , Costelloe, B.R. , & Couzin, I.D. (2019). Deepposekit, a software toolkit for fast and robust animal pose estimation using deep learning. *eLife* **8**, e47994 https://doi.org/10.7554/eLife.47994

Günel, S. , Rhodin, H. , Morales, D. , Campagnolo, J. , Ramdya, 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. , & et al. (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. In *Proceedings of the IEEE international conference on computer vision*

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

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. In *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 Transactions on Pattern Analysis and Machine Intelligence, 36*(7), 1325–1339.

Iskakov, K. , Burkov, E. , Lempitsky, V. , & Malkov, Y. (2019). Learnable triangulation of human pose. In *Proceedings of the IEEE/CVF international conference on computer vision*

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*https://doi.org/10.1101/2023.08.10.552747

Jocher, G. , Chaurasia, A. , & Qiu, J. (2023). Yolo by ultralytics. https://github.com/ultralytics/ultralytics

Joska, D. , Clark, L. , Muramatsu, N. , Jericevich, R. , Nicolls, F. , Mathis, A. , & Patel, A. (2021). Acinoset: A 3d pose estimation dataset and baseline models for cheetahs in the wild. In *2021 ieee international conference on robotics and automation (icra)* (pp. 13901–13908).

Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. *Journal of Basic Engineering, 82*(1), 35–45.

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.

Karaev, N. , Rocco, I. , Graham, B. , Neverova, N. , Vedaldi, A. , & Rupprecht, C. (2023). Cotracker: It is better to track together. *arXiv preprint* arXiv:2307.07635

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*

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. *Nature Methods, 19*, 496–504.

Li, Y. , Huang, C. , & Nevatia, R. (2009). Learning to associate: Hybrid-boosted multi-target tracker for crowded scene. In *2009 IEEE conference on computer vision and pattern recognition* (p. 2953-2960).

Lin, T.- Y. , Dollar, P. , Girshick, R. , He, K. , Hariharan, B. , & Belongie, S. (2017). Feature pyramid networks for object detection. In *2009 IEEE conference on computer vision and pattern recognition*.

Luiten, J. , & Hoffhues, A. (2020). Trackeval. https://github.com/JonathonLuiten/TrackEval.

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. *International Journal of Computer Vision, 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. *Nature Neuroscience, 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., Ákos, Z., Biro, D., & Vicsek, T. (2010). Hierarchical group dynamics in pigeon flocks. *Nature, 464*(7290), 890–893.

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. *Science Advances* (in press)

Nagy, M., Vásárhelyi, G., Pettit, B., Roberts-Mariani, I., Vicsek, T., & Biro, D. (2013). Context-dependent hierarchies in pigeons. *Proceedings of the National Academy of Sciences, 110*(32), 13049–13054.

Naik, H. (2021). *Xr for all: Closed-loop visual stimulation techniques for human and non-human animals (Dissertation)*. München: Technische Universität 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. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 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. *Nature Protocol, 14*, 2152–2176.

Newell, A. , Yang, K. , & Deng, J. (2016). Stacked hourglass networks for human pose estimation. In *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. *Nature Methods, 17*, 1052–1059.

Papadopoulou, M., Hildenbrandt, H., Sankey, D. W., Portugal, S. J., & Hemelrijk, C. K. (2022). Self-organization of collective escape in pigeon flocks. *PLoS Computational Biology, 18*(1), e1009772.

Paszke, A. , Gross, S. , Massa, F. , Lerer, A. , Bradbury, J. , Chanan, G. , & Chintala, S. (2019). Pytorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems*.

Pedersen, M. , Haurum, J.B. , Bengtson, S.H. , & Moeslund, T.B. (2020). 3d-zef: A 3d zebrafish tracking benchmark dataset. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*.

Pereira, T. D., Aldarondo, D. E., Willmore, L., Kislin, M., Wang, S.S.-H., Murthy, M., & Shaevitz, J. W. (2019). Fast animal pose estimation using deep neural networks. *Nature 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. *Nature 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. In *European conference on computer vision* (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. *Nature Methods, 16*, 179–182.

Sanakoyeu, A. , Khalidov, V. , McCarthy, M.S. , Vedaldi, A. , & Neverova, N. (2020 June). Transferring dense pose to proximal animal classes. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition (cvpr)*.

Sasaki, T., & Biro, D. (2017). Cumulative culture can emerge from collective intelligence in animal groups. *Nature Communications, 8*(1), 15049.

Sun, J.J. , Karashchuk, L. , Dravid, A. , Ryou, S. , Fereidooni, S. , Tuthill, J.C. , & others (2023). Bkind-3d: Self-supervised 3d keypoint discovery from multi-view videos. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 9001–9010).

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. In *Proceedings of the IEEE conference on computer vision and pattern recognition*.

Waldmann, U. , Bamberger, J. , Johannsen, O. , Deussen, O. , & Goldlücke, B. (2022). Improving unsupervised label propagation for pose tracking and video object segmentation. In *Dagm German conference on pattern recognition* (pp. 230–245).

Waldmann, U. , Johannsen, O. , & Goldluecke, B. (2023). Neural texture puppeteer: A framework for neural geometry and texture rendering of articulated shapes, enabling re-identification at interactive speed. arXiv preprint arXiv:2311.17109

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. In *Dagm German conference on pattern recognition* (pp. 513–528).

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. In *Proceedings of the IEEE conference on computer vision and pattern recognition*.

Wang, P. , Shen, X. , Lin, Z. , Cohen, S. , Price, B. , & Yuille, A.L. (2015). Joint object and part segmentation using deep learned potentials.

In *Proceedings of the IEEE international conference on computer vision*

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. In *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. In *Proceedings of the European conference on computer vision (ECCV)*.

Xu, Y. , Zhang, J. , Zhang, Q. , & Tao, D. (2022). ViTPose: Simple vision transformer baselines for human pose estimation. *Advances in Neural Information Processing Systems*.

Yang, J. , Gao, M. , Li, Z. , Gao, S. , Wang, F. , & Zheng, F. (2023). Track anything: Segment anything meets videos. arXiv preprint arXiv:2304.11968

Yang, Y., & Ramanan, D. (2013). Articulated human detection with flexible mixtures of parts. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 35*(12), 2878–2890.

Yomosa, M., Mizuguchi, T., Vásárhelyi, G., & Nagy, M. (2015). Coordinated behaviour in pigeon flocks. *Plos One, 10*(10), e0140558.

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/

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.