Erschienen in: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).
Piscataway, NJ: IEEE, 2023, pp. 21274-21284. ISBN 979-8-3503-0129-8
https://doi.org/10.1109/cvpr52729.2023.02038

3D-POP - An automated annotation approach to facilitate markerless 2D-3D tracking of freely moving birds with marker-based motion capture

Hemal Naik[1234*], Alex Hoi Hang Chan[12*], Junran Yang[2], Mathilde Delacoux[12],
Iain D. Couzin[123], Fumihiro Kano[12†], Máté Nagy[12356†]

[1]Dept. of Collective Behavior and Dept. of Ecology of Animal Societies, Max Planck Institute of Animal Behavior,
[2]Dept. of Biology, University of Konstanz, [3]Centre for the Advanced Study of Collective Behaviour, University of Konstanz,
[4]Computer Aided Medial Procedures, Informatik Department, Technische Universität München,
[5]Dept. of Biological Physics, Eötvös Loránd University, [6]MTA-ELTE 'Lendület' Collective Behaviour Research Group,
Hungarian Academy of Sciences. [*,†]contributed equally. Full affiliation available in supplementary

{hnaik, icouzin}@ab.mpg.de, nagymate@hal.elte.hu,
{hoi-hang.chan, junran.yang, mathilde.delacoux, fumihiro.kano}@uni-konstanz.de

Abstract

Recent advances in machine learning and computer vision are revolutionizing the field of animal behavior by enabling researchers to track the poses and locations of freely moving animals without any marker attachment. However, large datasets of annotated images of animals for markerless pose tracking, especially high-resolution images taken from multiple angles with accurate 3D annotations, are still scant. Here, we propose a method that uses a motion capture (mo-cap) system to obtain a large amount of annotated data on animal movement and posture (2D and 3D) in a semi-automatic manner. Our method is novel in that it extracts the 3D positions of morphological keypoints (e.g eyes, beak, tail) in reference to the positions of markers attached to the animals. Using this method, we obtained, and offer here, a new dataset - 3D-POP with approximately 300k annotated frames (4 million instances) in the form of videos having groups of one to ten freely moving birds from 4 different camera views in a 3.6m x 4.2m area. 3D-POP is the first dataset of flocking birds with accurate keypoint annotations in 2D and 3D along with bounding box and individual identities and will facilitate the development of solutions for problems of 2D to 3D markerless pose, trajectory tracking, and identification in birds.

1. Introduction

Computer vision and machine learning are revolutionizing many facets of conventional research methods. For example, dataset-driven machine learning methods have demonstrated remarkable success in the field of animal behavior, in tasks related to object detection [40, 41], tracking and individual identification [11, 25, 43], species recognition [40], 2D pose estimation [14, 29] and 3D pose estimation [2, 4]. These automatic methods not only reduce the required labor and errors associated with manual coding of behaviours [6, 39] but also facilitate long-term continuous monitoring of animal behavior in both indoor (lab) [32, 36] and outdoor (wild) settings [12, 37]. Engineering and robotics experts use the data on animal locomotion to reverse-engineer the key mechanisms underlying behaviors and movements of animals [18, 20]. The development of new techniques critically depends on the quality of publicly-available datasets with accurate annotations.

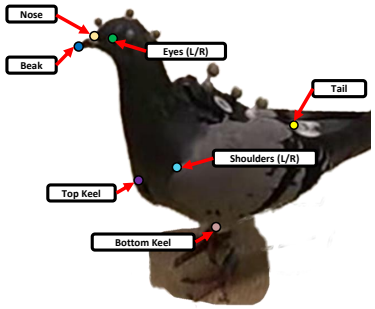

Figure 1. Definition of morphological keypoints offered in the 3D-POP dataset

Creating large datasets with animals is particularly difficult because every species has distinct morphology, and also because it is generally challenging to film freely moving animals in a controlled environment. It is thus important for

datasets to include a wide range of species and behaviors to maximize the practical application of machine learning methods for animal tracking. Although animals have been included in many popular image datasets collected from the internet such as ImageNet [9] and COCO [26], those datasets have not fulfilled more specific needs of animal behavior researchers. Hence, recently several datasets have been created with a focus on animal behavior research, such as species classification [12, 37, 40, 41, 48], behavioral classification [27, 31, 48] and posture tracking [2, 10, 15, 24, 47].

The most common approach for creating datasets of animals is through manual annotations in the image space (2D). As a result, most solutions to single / multiple animal detection, tracking, or pose estimation problems are limited to the 2D space [14, 25], or use 2D image projections to validate the results of 3D predictions without ground truth [2, 4]. For nonhuman animals, a dataset similar to Human 3.6M [16] is necessary to develop solutions for problems on 2D/3D tracking and posture prediction with a range of constraints, such as single or multiviews, single or multi-individual, and tracking using single frame or temporal consistency. More recently, marker-based motion-capture technology has been used to create 3D datasets for rats [10] and dogs [22] with one individual. The application of mo-cap for animal behaviour studies has also increased in popularity, such as studying flight kinematics [23] and gaze behavior in a freely moving group [17, 21]. It is clear that datasets with mo-cap will not only enhance the size of the dataset but also improve the accuracy of annotations, thus providing a large 2D/3D ground truth dataset for the animal position, posture, and identity tracking. However, despite its potential, researchers have only begun using mo-cap for behavior studies, and further work is required in terms of method development and dataset collection.

We propose a new mo-cap-based approach to create large-scale datasets with a bird species (homing pigeons, *Columba livia*), and provide a complete code base for further applications to other species. Along with 2D-3D posture, the dataset also offers annotations for 2D-3D movement trajectories (position) with ground truth on identities for up to 18 individuals. We overcame the unique challenge of needing to attach reflective markers on desired but often inaccessible morphological keypoints on animal bodies and instead determined the relative 3D position of these keypoints to markers attached on accessible parts of the animal (Figure 1).

The method enables a large amount of training data to be generated in a semi-automatic manner with minimal time investment and human labor. Moreover, by tracking freely-moving animals in a relatively large area (3.6m x 4.2m), we were able to track a variety of naturalistic behaviors in a flock consisting of up to 10 individuals under realistic experimental conditions. Finally, we demonstrate through a series of experiments that our method is consistent and the CNN models trained on our dataset are able to predict the postures of birds with no markers attached to their bodies.

2. State of the Art

2.1. 2D posture

Animal Kingdom [31] is by far the largest dataset with 50 hours of video annotations that include 850 species of varied taxa (fish, birds, mammals, etc.), focusing on a generalizable solution for 2D pose estimation and activity recognition for a single individual. Other notable datasets contain images instead of videos and focus on capturing variations in terms of specific taxa *e.g*. mammals [48], birds [41] and monkeys [47], or specific species *e.g*. zebras [14], all of these focus on solving problems for a single individual recorded from a single viewpoint.

Datasets based on single animal-based solutions are sufficient for some cases and rely on detection-based top-down approaches for extending the method for tracking the posture of multiple individuals [42]. There are few datasets that offer posture annotations for multiple individuals [2, 24, 25]. The problem of tracking multiple individuals is often simplified by placing the cameras above the animals, which minimizes occlusions [25, 43]. Tracking multiple individuals from side views may require multiple views, which may be important to resolve occlusions when animals interact in 3D spaces *e.g*. Cowbird dataset [2].

The existing datasets have motivated the development of various methods for posture estimation. However, reliance on manual annotations limits the complexity of datasets in terms of the number of viewpoints or the number of individuals, especially for video sequences.

2.2. 3D posture

Datasets with ground truth on 3D posture are relatively difficult to obtain with a group of animals. One popular method for obtaining 3D ground truth posture is the triangulation of 2D postures using multiple views to record animals. Acinoset [20] (leopard in wild), Fly3D [15] (fly in a lab) and OpenMonkeyStudio [3] (macaque in a lab) use triangulation-based approaches to provide 3D posture of single individuals. The images for these datasets are also annotated manually and, therefore, the accuracy of the computed 3D pose depends on the quality of annotation and calibration.

An alternative approach is to use marker-based mo-cap with a skeleton tracking feature as used with humans [16]. Kearney *et al*. [22] used motion capture to generate 3D ground truth for dogs and combine their approach with depth sensors (RGB-D) with the aim of designing markerless tracking based on RGBD sensors (63 to 82 markers). Dunn *et al*. [10] offered Rat 7M dataset using mo-cap with

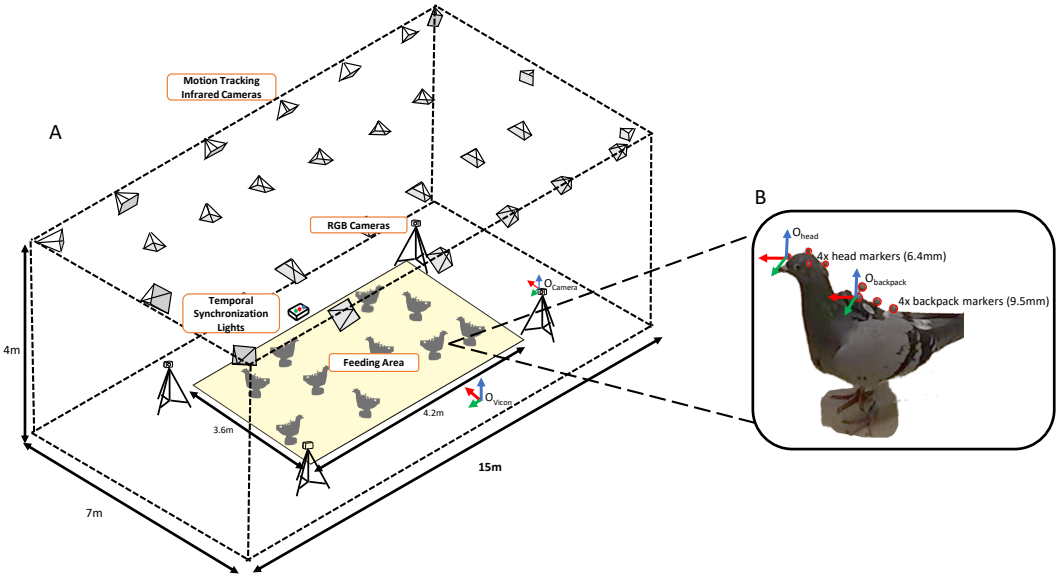

Figure 2. Illustration showing the experimental setup, with different defined coordinate systems, including the Vicon global coordinate system (O_{Vicon}), the camera coordinate system for each RGB camera (O_{Camera}), and the head (O_{head}) and backpack ($O_{backpack}$) coordinate system for each pigeon subject A) Detailed floor plan for data collection. B) Pigeon subject, with corresponding head and backpack markers and coordinate systems

RGB cameras and 20 markers. These datasets are useful for solving posture problems for a single individual from multiple views and offer the option of using temporal consistency.

Recently, Marshall *et al*. published PAIR-R24M [27], the first dataset with 3D ground truth with more than one animal, a pair of rats, using the approach of Dunn *et al*. [10]. Motion capture systems offer the huge advantage of creating millions of annotations in an automatic manner with high accuracy and low noise. The skeleton tracking feature with the mo-cap is primarily designed for tracking human posture and relies on a large number of markers. Additionally, the marker patterns have to be unique (at least partially) for maintaining the identity of each individual. Marker placement is a also limitation for smaller species and wild animals that lack the tolerance for having markers placed on specific locations of their body.

Lack of ground truth in 3D posture had led to innovative work of predicting 3D posture using 2D keypoints and silhouettes [2, 4] or using synthetic datasets [5] or toys [49]. These approaches are promising but lack quantitative evaluation for robust practical applications. Computer vision literature on 3D posture problems mainly focuses on extracting as much of detail as possible. For animal behavior experiments information required from videos is always defined in the context of the experiment. It is worth noting that for birds, tracking the head and body orientations could be sufficient to quantify many key behaviors in ground-foraging contexts, such as feeding (pecking ground), preening, vigilance (head scanning), courtship (head bowing), or walking. Measuring the head direction in 3D also allows gaze reconstruction [17, 21], to be applied in the study of social cognition and collective behavior.

2.3. Multi-object tracking with identity

Identity recognition is a critical problem to solve in the context of biological studies, especially when tracking the behavior of multiple interacting individuals over long periods of time. Tracking and identification of multiple individuals in large groups are especially exciting to quantify group-level behaviors like social networks [44, 46], dominance, or leadership [30].

For indoor experiments, the number of individuals is often controlled and the identification problem is linked with the tracking of animals [25, 43]. The task of tracking and identification is often resolved together using markers [14, 30] or marker-less [7, 11, 43] methods. The existing solutions perform well with specific perspectives (top-down view) and thus often fail to resolve cases of occlusion. Robust evaluation of simultaneous identification and tracking methods is difficult because true ground truth for identities is often not available in datasets with multiple animals or available for only a very short duration [25].

There are many good datasets available to independently solve problems of posture estimation, detection, tracking, and identification. Very few datasets offer the possibility of solving all of these problems simultaneously in realistic experimental scenarios.

We aim to fill this gap with our contribution of a semi-automatic method for producing new datasets with animals. Our dataset, 3D-POP, includes video recordings of 18 unique pigeons in various group sizes (1,2,5,10) from multiple views. We offer ground truth for identity, 2D-3D trajectories, and 2D-3D posture mapping for all individuals across the entire dataset (300K frames). The dataset also consists of annotations for object detection in the form of bounding boxes.

3. Methods

3.1. Experimental Setup

The dataset was collected from pigeons moving on a jute fabric (3.6m x 4.2m) onto which we evenly scattered grains to encourage the birds to feed in that area (Figure 2A). This feeding area was located inside a large enclosure equipped with a mo-cap system (15m x 7m x 4m). This mo-cap system consists of 30 motion capture cameras (12 Vicon Vero 2.2, 18 Vicon Vantage-5 cameras; 100Hz) and can track the 3D positions of reflective markers with sub-millimeter precision. At the corners of the feeding area, we placed 4 high-resolution (4K) Sony action cameras (rx0ii, 30Hz, 3840x2160p) mounted on standard tripods and also an Arduino-based synchronization box which flashes RGB and infrared LED lights every 5 seconds (Figure 2). Details on the synchronization and calibration of RGB cameras are provided in the supplementary text.

3.2. Animal Subjects

Eighteen pigeons (*Columba livia*) were subjected to this study over 6 experimental days. Each day 10 pigeons were randomly selected from the population. Four 6.4mm reflective markers were attached to each subject's head, and four 9.5mm markers were attached to a customized backpack worn by each subject (Figure 2B). Generally, pigeons tolerate markers on the head with minimal effects on their behavior and habituate quickly to backpacks. Backpacks are also widely used for bird studies in behavioral ecology [1, 45]. The four 9.5mm backpack markers had a unique geometric configuration to track the individual identities of each bird throughout each recording. Each day we performed up to 11 trials in the following order: 1 pigeon (4 trials), a pair of pigeons (4 trials), a flock of 5 pigeons (2 trials), and a flock of 10 pigeons (1 trial). It took approximately 1 hour to perform all trials each day. The total frames and duration of samples over the course of the experiment are described in Table 1. An additional session was recorded with birds without attaching any markers to validate the results of models trained on annotated data having birds with markers (see 5.2).

3.3. Data annotation pipeline

3.3.1 Annotation principle

The movement of all features on a rigid body can be tracked simultaneously in a 3D space by computing 6-DOF pose of the rigid object. We use this principle to achieve annotations for keypoint features that are rigidly attached to the head and body of the bird. The four markers attached to the head and body (using a backpack) of each pigeon are used to compute 6-DOF pose of these body parts using the mo-cap system.

By assuming that the head and body are rigid bodies in the case of walking or standing birds, we designed a pipeline to annotate the position of features on the head and body (beak, eyes, shoulder, and tail, etc.) in a few frames to compute their 3D location with respect to marker positions. Once computed, the relationship between markers and features does not change during the sequences and this ensures that 6-DOF pose of head and body for any frame can be used to project 3D positions of keypoint features onto the image space to obtain 2D annotations.

All keypoints defined for the head lie on the skull of the bird (Figure 1). The rigidity assumption is valid for these keypoints as they are rigidly placed on the skull. The keypoints chosen for the body lie actually on the rib cage and shoulders and exhibit a limited range of motion independent of each other. The rigidity assumption for the body is a reasonable assumption for the annotation pipeline if the birds do not move their wings and body(see 5.3).

3.3.2 Manual annotation

6-DOF (Degrees of freedom) tracking of the head and body is used to create a bounding box around the bird and crop the image of the focal individual for annotation. For each individual pigeon, 9 morphological keypoints (Figure 1) are annotated on 5-10 frames from all available view angles. Ideally, four frames (1 per view) is sufficient, but all keypoints are rarely visible within a single instance. Moreover, multiple measurements (3-5 frames per view) improve the robustness of computed 3D keypoint positions. The position of each keypoint is first triangulated using sparse bundle adjustment (in the camera coordinate system), then the relative position of the keypoint is computed with respect to the markers (in the coordinate system of the body part). Finally, all resultant 3D positions of keypoints are averaged and stored as a template file. This process is repeated for each bird on each recording day.

3.3.3 Annotation propagation

In this final step, the ground truth data is generated for each recording using 3D keypoint positions computed in the previous step. The 3D positions of the keypoint features are transferred to the global coordinate system using 6-DOF pose. Next, keypoints are transferred to the coordinate system of each camera and projected to the image space (using calibration parameters). Bounding box annotations for object detection or tracking tasks are derived from keypoint projections. We determined that keypoints with the minimum and maximum x-y pixel values with an offset of 60 pixels are sufficient to define a bounding box. Finally, the 6-DOF tracking with the mo-cap system maintains the identity of each bird and this is also stored with 2D-3D information for the entire sequence.

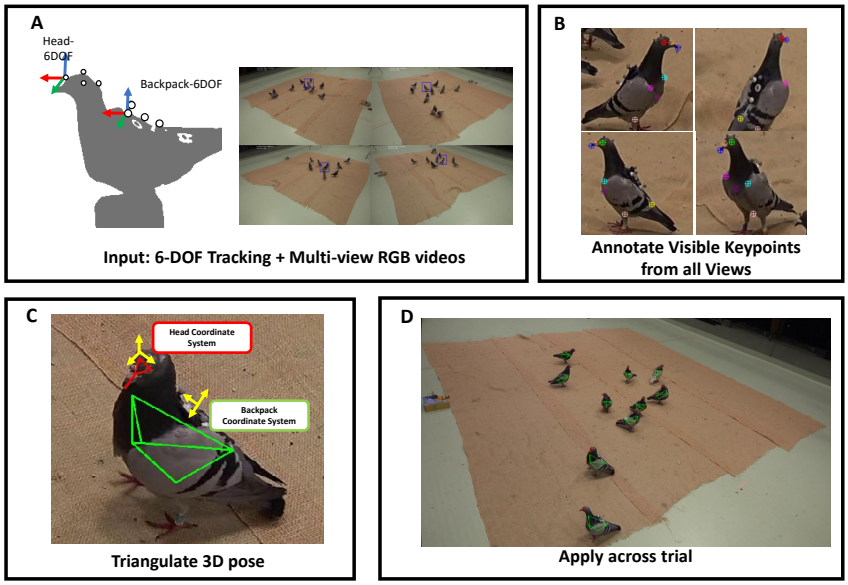

Figure 3. Semi-automated annotation pipeline based on 6DOF tracking and RGB images. A) Input 6-DOF tracking data for head and backpack coordinate systems, and multi-view RGB videos. B) Manually annotate all visible keypoints from all views. C) Triangulate 3D position of all keypoints in the head and backpack coordinate system, assuming that keypoints and tracked markers are a rigid body. D) Apply across trials to get keypoints across all individuals

4. The 3D-POP dataset

4.1. Dataset Description

We present 3D-POP (3D Postures of Pigeons), a dataset that provides accurate ground truth for 3D keypoints, 2D keypoints, bounding boxes, and individual identities. The dataset includes RGB images from four high-resolution cameras (4K) and up to 6 hours of recordings divided into 57 sequences of 1,2,5, and 10 pigeons behaving naturalistically (Table 1). The dataset contains 3D coordinates (unfiltered) and 6-DOF pose obtained from the mo-cap facility along with calibration parameters. We also provide a total of 1 hour of recording (11 sequences) with pigeons (group size:1,2,5,11) without any markers on their body. These videos are provided for users to test the practical effectiveness of markerless solutions without the influence of markers. For realistic assessment, we show that a model trained with our dataset is able to infer keypoints on videos with pigeons without markers (see 5.2). Download the dataset here: https://doi.org/10.17617/3.HPBBC7

No. individuals	Annotated frames	Video length (min)
1	95,513	55
2	135,547	119
5	44,240	85
10	20,321	91

Table 1. Dataset Summary: Total number of labeled frames with ground truth data for different group sizes.

4.2. Customization

We release 3D-POPAP (3D-POP Annotation Pipeline) to manipulate the annotations of the dataset (Download: https://github.com/alexhang212/Dataset-3DPOP). As explained earlier, our use of the 6-DOF tracking decouples the keypoint annotations from the positions of markers used for mo-cap. Due to this design of the annotation approach, we can offer a unique dataset with the ability to easily add new 2D/3D keypoint annotations. The feature of keypoint modification is relevant for future work because defining the posture of birds is a difficult problem and depends on the final application. As of now, there are no datasets available with ground truth on the 3D posture of birds. The lack of ground truth has motivated novel ideas for solving the 3D reconstruction of bird pose using 2D annotations (silhouette and keypoints [2]). Among the available 2D datasets with birds, different numbers of keypoints are selected to define pose *e.g.* CUB-200: 15 [41], Cowbird dataset: 12 [2] and Animal Kingdom: 23 [31].

To the best of our knowledge, the use of posture in behavior studies with birds is still limited and pose definition may rely completely on the nature of the study. Our inspiration for keypoint definition is inspired by gaze studies [17, 21] for which 9 keypoint-based posture sufficiently provides gaze direction with body and head orientation.

4.3. Dataset Validation

The annotations in 3D-POP are obtained automatically, and therefore we designed three different tests to validate the accuracy and consistency of the annotations. The first test compares the accuracy of the 3D features computed with our method and the method presented by Kano *et al.* [21].

The second test measures the consistency of the 3D/2D annotations across the dataset. This test is required to identify errors in annotation introduced by erroneous 3D mo-cap tracking due to occlusion, rapid movement of the birds, or calibration and synchronization errors of the cameras. It is important to perform this test because manually checking millions of annotations is not practical. Finally, the third test checks the variation in the 3D pose captured in all sequences. This test shows that the dataset is not biased to specific types of motion or poses.

4.3.1 Accuracy

Kano *et al*. [21] use a calibration method to measure the 3D position of eyes w.r.t. mo-cap markers. This process involves a custom camera rig, made of 4 separate webcams that capture the head of each pigeon before data collection. We replicated this process to compute the ground truth 3D position of eyes and beak. Further, we compared the ground truth with the 3D position of the same features computed with our approach.

We obtained root mean squared errors (RMSE) for all three features (Beak: 5.0 mm, Left eye: 5.0 mm, Right Eye: 4.9 mm), which is sufficient for pigeons considering that the diameter of the eyes is typically 6-7 mm [8]. This method provides an approximation of the accuracy for a few features only, and a better method is required to test the accuracy of 3D features measured on the body. It should be noted that our method has comparable accuracy and alleviates the need of using dedicated calibrations rigs and thus saves time.

4.3.2 Consistency and outlier detection

It is reasonable to assume that a small portion of the mo-cap sequences contains tracking errors and will produce inaccurate 6-DOF poses for body parts. As a result, the annotation for all keypoints associated with the relevant body parts is likely to be wrong. We know that models trained with large datasets with small noise still generalize to a solution [34]. Yet, it is important to identify and remove these sequences from the dataset. Keeping this in mind, we design a consistency check with the intuition that a well-trained model for keypoint detector will predict 2D features with reasonable accuracies for all frames. Therefore, a comparison between predicted keypoints and propagated keypoints is likely to show very large errors for all keypoints (of the same body part), especially for frames with faulty mo-cap tracking (Figure 5). We use this idea to automatically determine the consistency of the annotations throughout the trial.

We trained a state-of-the-art 2D keypoint detection model (DLC [28]) on 15177 images with a ResNet50 backbone for 30,000 iterations with the adam optimizer. There

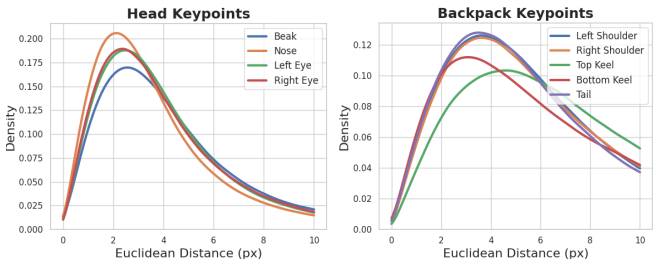

Figure 4. Distribution of Euclidean distances (px) between model predictions of a trained DLC model and annotations, after outlier frames were filtered. Frequency shown in the y-axis and and only points of up to 10px error is shown on the x-axis. A) Head keypoints B) Backpack keypoints

is no reliable method available for tracking the posture of multiple birds simultaneously, therefore we use a top-down approach and train on single individual data using bounding box annotations. The training data excludes highly occluded frames with $>$ 30% overlap with another bounding box to avoid sequences that have multiple individuals in the bounding box due to close proximity. GESD outlier analysis [35] is used for each keypoint independently setting the expected outliers at 20% of the dataset. The frames having more than 1 outlier keypoint are filtered out as we expect a higher number of outliers in case of erroneous annotations (explained above).

Using this method we filtered out 2.9% of the overall dataset, which lowered the average Euclidean distance between annotation and predictions (see Table 2). We used the filtered training data and retrained a model (14,722 images, 30,000 iterations, ResNet50 backbone, adam optimizer), but obtained similar errors compared to the previous model (see Table 2). The consistency check reveals that the annotations are largely consistent with model predictions, with a typical error of 2-3 pixels for head features and 3-4 px for body features (See Figure 4). Figure 5 shows visual examples of outlier frames where mo-cap errors are likely due to behaviors such as flying or occlusions.

The outlier filtering method introduces artificial gaps in the dataset. We computed the number of dropped frames and found that 96.1% of gaps are less than 30 frames (1 second) in length (see supplementary). Researchers in need of continuous temporal data can use gap-free segments or use interpolation to fill small gaps. For sake of completeness, we have included automatically rejected frames in the dataset.

4.3.3 Pose variation

We then compute the number of unique poses that each pigeon exhibit to understand the heterogeneity of pose present in the 3D-POP dataset. It is difficult to compute pose vari-

$RMSE_{Method}$ (px)	Beak	Nose	Left Eye	Right Eye	Left Shoulder	Right Shoulder	Top Keel	Bottom Keel	Tail
$RMSE_{BeforeFiltering}$	10.1	7.9	7.5	7.5	8.4	8.7	9.4	9.9	8.8
$RMSE_{AfterFiltering}$	8.1	6.0	5.9	5.9	7.9	8.2	9.1	9.5	8.2
$RMSE_{AfterRetraining}$	8.4	6.5	6.4	6.3	8.0	8.2	9.1	9.5	8.4

Table 2. Root mean squared 2D Euclidean error (px) of each keypoint with different data subsets and trained DLC 2D keypoint models. BeforeFiltering: Error of model trained on the full dataset with inference on frames before outliers were filtered. AfterFiltering: Errors of the model trained on the full dataset with inference on frames after outliers were filtered. AfterRetraining: Errors of the model trained on the filtered dataset with inference on frames after outliers were filtered

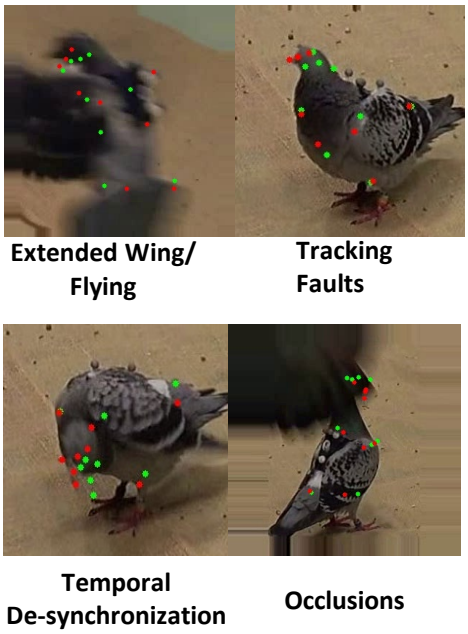

Figure 5. Example frames that are filtered automatically by the outlier analysis, with descriptions of the cause of annotation inaccuracy. Green labels represent annotations, and red labels represent prediction from the trained DLC 2D keypoint detection model.

ation directly using the 6-DOF pose defined by markers because the coordinate system is not defined in a standardized way for each pigeon. To create a standardized comparison, we compute two planes defined by keypoint features to represent the alignment of the head and body in 3D space. The head plane is computed using three points (beak and eyes) and the body plane is computed using three points (shoulders and tail). In this manner, the orientations of all planes representing the pose of all individuals are defined using the same features and can be compared in a unified coordinate system. We use the normal of the planes to compute the angles with the canonical coordinate system (See supplementary). It is assumed that a degree of change in rotational angles of either head or body corresponds to a new pose. We found a total of 74,924 unique orientations of the head and 14,191 unique orientations of the body, and the combined 1.8 million unique poses present in the dataset. A graphical representation of the range of poses is provided in the supplementary material.

5. Experiments

5.1. Marker-based + Markerless Hybrid Approach

The first experiment shows that markerless tracking algorithm trained on 3D-POP is able to solve 3D tracking for cases when mo-cap fails to track markers. The solution is useful as an increasing number of pre-existing experimental setups are designed to use marker-based mocap technologies for biological studies [10, 17, 21, 23, 38]. A hybrid tracking solution, that uses markerless tracking to fill the gaps of the mo-cap system has many potential applications for future behavior studies.

We chose a 5 min sequence with a single individual and artificially removed 25% of mo-cap tracking data. The gaps are randomly introduced for a duration of 30-90 frames (1-3 seconds), to mimic tracking loss. We used the 2D keypoint DLC model (see 4.3) to detect keypoints from all 4 camera views and triangulate the results with sparse bundle adjustment. We compared the result with the ground truth and achieved avg. RMS error of 9.2 mm (details in supplementary). A simple linear interpolation-based approach to fill gaps resulted in avg. RMS error of 52.1 mm. The proposed solution is a viable application because biologists are likely to keep using motion-tracking technology until a robust solution is designed for markerless 3D tracking. However, we acknowledge that better solutions can be designed for a hybrid approach using temporal consistency in the future [20].

5.2. Markerless Bird Tracking

This experiment shows that models trained with our datasetcan be directly used to track birds without any markers attached to their bodies. This experiment works as a "sanity check" to ensure that models trained with 3D-POP dataset are not biased toward the presence of markers. The test also demonstrates the potential contribution of our method toward developing a complete markerless solution

Figure 6. Pictures show that the 2D keypoint detection algorithm trained with the 3D-POP dataset can make predictions on videos with pigeons without any markers attached to the body.

for 3D tracking, posture estimation, and identification.

Using a pre-trained object detection model (YOLOv5s [33]), we extracted the bounding box of a pigeon from a single individual sequence. We then used the 2D keypoint DLC model (see 4.3) to predict keypoints from the sequence. The models generalize well to the images of pigeons without markers (see Figure 6, supplementary video). The result is qualitatively checked, but sufficient to prove our claim. The same solution can be easily extended to multiple pigeon trials by designing a top-down approach (using YOLO) until better solutions are developed using 3D-POP.

5.3. Manual Validation

This experiment demonstrates the validity of our assumption that keypoints on the body (shoulder, keel, etc.) behave like points on a rigid body. We selected 1000 frames randomly and manually annotated keypoints for the body part. We compared the manual annotations with automatic ground truth annotations using PCK05 and PCK10 (percentage correct keypoint within 5% and 10% of bounding box width) metrics. We report an average PCK05 of 66% and PCK10 of 94% across all keypoints on the body (Table 3). We also visually quantified that only 2.8% of the frames are cases where birds are moving their wings, thus the simplified skeletal representation of the body is valid in over 97% of the dataset.

Metric	Left Shoulder	Right Shoulder	Top Keel	Bottom Keel	Tail
PCK05	0.78	0.75	0.58	0.57	0.60
PCK10	0.98	0.98	0.94	0.89	0.92

Table 3. PCK errors per body keypoint between manual annotation and 3DPOP annotation. PCK is defined as the percentage of points that are within 5% and 10% of the bounding box width

6. Limitations and Future work

The annotation method presented in the paper largely relies on the assumption that the head and body mostly behave as rigid bodies. This assumption does not hold for certain body parts such as the neck, tail end, or feet and limits the selection of keypoints at these body parts. For similar reasons, the proposed approach will not support annotation for flying birds or birds that change the shape of body parts while performing certain behaviors *e.g.* courtship [19].

Our approach inherently depends on the tracking accuracy of the mo-cap system. Users must maintain mo-cap systems regularly calibrated for consistent results. Another possible source of error in the annotation pipeline is video camera calibration and its temporal synchronization with the mo-cap system. We do show that our outlier detection method is effective at identifying noisy annotations, however, noise can still be present in the dataset. Finally, since the dataset was curated semi-automatically in an existing motion tracking setup, the data we provide is limited to an indoor environment.

We have improved the existing state of the art for multi-animal tracking by adding complexity in the form of the number of individuals and camera views. In the future, we intend to develop lifting-based approaches [13, 15] to learn the 2D-3D mapping obtained in the 3D-POP dataset to track birds in outdoor environments.

7. Conclusion

In this paper, we introduced a novel method to use a mo-cap system for generating large-scale datasets with multiple animals. We demonstrate that our semi-automated method offers an alternative for generating high-quality datasets with animals without manual effort. We offer 3D-POP, the first dataset with ground truth for 3D posture prediction and identity tracking in birds, which is extremely difficult to achieve even with manual labor. 3D-POP dataset offers an opportunity for the vision community to work on a complex set of vision problems relevant to achieving markerless tracking of birds in indoor and outdoor environments. At the same time, our method will motivate biologists to create new datasets as they have access to and work with different types of animals.

8. Acknowledgements

Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC 2117 (ID: 422037984). The Ethical Committee of Baden-Württemberg approved all the experiments(Regierungspräsidium Freiburg, Referat 35, License Number: 35-9185.81/G-19/107). M.N. acknowledges additional support from the Hungarian Academy of Sciences (grant no. 95152) and Eötvös Loránd University. I.C. also acknowledge Office of Naval Research (grant ONR, N00014-19-1-2556), Horizon Europe Marie Sklodowska-Curie Actions (860949) and the Max Planck Society.

References

[1] Gustavo Alarcón-Nieto, Jacob M Graving, James A Klarevas-Irby, Adriana A Maldonado-Chaparro, Inge Mueller, and Damien R Farine. An automated barcode tracking system for behavioural studies in birds. *Methods in Ecology and Evolution*, 9(6):1536–1547, 2018. 4

[2] Marc Badger, Yufu Wang, Adarsh Modh, Ammon Perkes, Nikos Kolotouros, Bernd G Pfrommer, Marc F Schmidt, and Kostas Daniilidis. 3d bird reconstruction: a dataset, model, and shape recovery from a single view. In *European Conference on Computer Vision*, pages 1–17. Springer, 2020. 1, 2, 3, 5

[3] Praneet C Bala, Benjamin R Eisenreich, Seng Bum Michael Yoo, Benjamin Y Hayden, Hyun Soo Park, and Jan Zimmermann. Automated markerless pose estimation in freely moving macaques with openmonkeystudio. *Nature communications*, 11(1):1–12, 2020. 2

[4] Benjamin Biggs, Thomas Roddick, Andrew Fitzgibbon, and Roberto Cipolla. Creatures Great and SMAL: Recovering the Shape and Motion of Animals from Video. In C.V. Jawahar, Hongdong Li, Greg Mori, and Konrad Schindler, editors, *Computer Vision – ACCV 2018*, Lecture Notes in Computer Science, pages 3–19, Cham, 2019. Springer International Publishing. 1, 2, 3

[5] Luis A Bolaños, Dongsheng Xiao, Nancy L Ford, Jeff M LeDue, Pankaj K Gupta, Carlos Doebeli, Hao Hu, Helge Rhodin, and Timothy H Murphy. A three-dimensional virtual mouse generates synthetic training data for behavioral analysis. *Nature methods*, 18(4):378–381, 2021. 3

[6] Marek L. Borowiec, Rebecca B. Dikow, Paul B. Frandsen, Alexander McKeeken, Gabriele Valentini, and Alexander E. White. Deep learning as a tool for ecology and evolution. *Methods in Ecology and Evolution*, 13(8):1640–1660, 2022. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/2041-210X.13901. 1

[7] Katarzyna Bozek, Laetitia Hebert, Yoann Portugal, Alexander S Mikheyev, and Greg J Stephens. Markerless tracking of an entire honey bee colony. *Nature communications*, 12(1):1–13, 2021. 3

[8] Ray D Chard and Ralph H Gundlach. The structure of the eye of the homing pigeon. *Journal of Comparative Psychology*, 25(2):249, 1938. 6

[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, June 2009. ISSN: 1063-6919. 2

[10] Timothy W Dunn, Jesse D Marshall, Kyle S Severson, Diego E Aldarondo, David GC Hildebrand, Selmaan N Chettih, William L Wang, Amanda J Gellis, David E Carlson, Dmitriy Aronov, et al. Geometric deep learning enables 3d kinematic profiling across species and environments. *Nature methods*, 18(5):564–573, 2021. 2, 3, 7

[11] André C Ferreira, Liliana R Silva, Francesco Renna, Hanja B Brandl, Julien P Renoult, Damien R Farine, Rita Covas, and Claire Doutrelant. Deep learning-based methods for individual recognition in small birds. *Methods in Ecology and Evolution*, 11(9):1072–1085, 2020. 1, 3

[12] Crystal Gagne, Jyoti Kini, Daniel Smith, and Mubarak Shah. Florida wildlife camera trap dataset. *arXiv preprint arXiv:2106.12628*, 2021. 1, 2

[13] Adam Gosztolai, Semih Günel, Victor Lobato-Ríos, Marco Pietro Abrate, Daniel Morales, Helge Rhodin, Pascal Fua, and Pavan Ramdya. LiftPose3D, a deep learning-based approach for transforming two-dimensional to three-dimensional poses in laboratory animals. *Nature Methods*, 18(8):975–981, Aug. 2021. Number: 8 Publisher: Nature Publishing Group. 8

[14] Jacob M Graving, Daniel Chae, Hemal Naik, Liang Li, Benjamin Koger, Blair R Costelloe, and Iain D Couzin. DeepPoseKit, a software toolkit for fast and robust animal pose estimation using deep learning. *eLife*, 8:e47994, Oct. 2019. Publisher: eLife Sciences Publications, Ltd. 1, 2, 3

[15] Semih Günel, Helge Rhodin, Daniel Morales, João Campagnolo, Pavan Ramdya, and Pascal Fua. Deepfly3d, a deep learning-based approach for 3d limb and appendage tracking in tethered, adult drosophila. *Elife*, 8:e48571, 2019. 2, 8

[16] Catalin Ionescu, Dragos Papava, Vlad Olaru, and Cristian Sminchisescu. 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, jul 2014. 2

[17] Akihiro Itahara and Fumihiro Kano. "Corvid Tracking Studio": A custom-built motion capture system to track head movements of corvids. *Japanese Journal of Animal Psychology*, pages 72–1, 2022. 2, 3, 5, 7

[18] Noah T Jafferis, E Farrell Helbling, Michael Karpelson, and Robert J Wood. Untethered flight of an insect-sized flapping-wing microscale aerial vehicle. *Nature*, 570(7762):491–495, 2019. 1

[19] Judith Janisch, Elisa Perinot, Leonida Fusani, and Cliodhna Quigley. Deciphering choreographies of elaborate courtship displays of golden-collared manakins using markerless motion capture. *Ethology*, 127(7):550–562, 2021. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/eth.13161. 8

[20] Daniel Joska, Liam Clark, Naoya Muramatsu, Ricardo Jericevich, Fred Nicolls, Alexander Mathis, Mackenzie W Mathis, and Amir Patel. Acinoset: a 3d pose estimation dataset and baseline models for cheetahs in the wild. In *2021 IEEE International Conference on Robotics and Automation (ICRA)*, pages 13901–13908. IEEE, 2021. 1, 2, 7

[21] Fumihiro Kano, Hemal Naik, Göksel Keskin, Iain D. Couzin, and Máté Nagy. Head-tracking of freely-behaving pigeons in a motion-capture system reveals the selective use of visual field regions. *Scientific Reports*, 12(1):19113, Nov 2022. 2, 3, 5, 6, 7

[22] Sinead Kearney, Wenbin Li, Martin Parsons, Kwang In Kim, and Darren Cosker. Rgbd-dog: Predicting canine pose from rgbd sensors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8336–8345, 2020. 2

[23] Marco KleinHeerenbrink, Lydia A France, Caroline H Brighton, and Graham K Taylor. Optimization of avian perching manoeuvres. *Nature*, 607(7917):91–96, 2022. 2, 7

[24] Rollyn Labuguen, Jumpei Matsumoto, Salvador Blanco Negrete, Hiroshi Nishimaru, Hisao Nishijo, Masahiko Takada, Yasuhiro Go, Ken-ichi Inoue, and Tomohiro Shibata. Macaquepose: A novel "in the wild" macaque monkey pose dataset for markerless motion capture. *Frontiers in behavioral neuroscience*, 14:581154, 2021. 2

[25] Jessy Lauer, Mu Zhou, Shaokai Ye, William Menegas, Steffen Schneider, Tanmay Nath, Mohammed Mostafizur Rahman, Valentina Di Santo, Daniel Soberanes, Guoping Feng, Venkatesh N. Murthy, George Lauder, Catherine Dulac, Mackenzie Weygandt Mathis, and Alexander Mathis. Multi-animal pose estimation, identification and tracking with DeepLabCut. *Nature Methods*, 19(4):496–504, Apr. 2022. Number: 4 Publisher: Nature Publishing Group. 1, 2, 3

[26] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common Objects in Context. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, Lecture Notes in Computer Science, pages 740–755, Cham, 2014. Springer International Publishing. 2

[27] Jesse D. Marshall, Ugne Klibaite, Amanda Gellis, Diego E. Aldarondo, Bence P. Ölveczky, and Timothy W. Dunn. The PAIR-R24M Dataset for Multi-animal 3D Pose Estimation. Technical report, bioRxiv, Nov. 2021. Section: New Results Type: article. 2, 3

[28] Alexander Mathis, Pranav Mamidanna, Kevin M. Cury, Taiga Abe, Venkatesh N. Murthy, Mackenzie Weygandt Mathis, and Matthias Bethge. DeepLabCut: markerless pose estimation of user-defined body parts with deep learning. *Nature Neuroscience*, 21(9):1281–1289, Sept. 2018. Number: 9 Publisher: Nature Publishing Group. 6

[29] Mackenzie Weygandt Mathis and Alexander Mathis. Deep learning tools for the measurement of animal behavior in neuroscience. *Current Opinion in Neurobiology*, 60:1–11, Feb. 2020. 1

[30] Máté Nagy, Gábor Vásárhelyi, Benjamin Pettit, Isabella Roberts-Mariani, Tamás Vicsek, and Dora Biro. Context-dependent hierarchies in pigeons. *Proceedings of the National Academy of Sciences*, 110(32):13049–13054, 2013. 3

[31] Xun Long Ng, Kian Eng Ong, Qichen Zheng, Yun Ni, Si Yong Yeo, and Jun Liu. Animal kingdom: A large and diverse dataset for animal behavior understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19023–19034, 2022. 2, 5

[32] Ali Nourizonoz, Robert Zimmermann, Chun Lum Andy Ho, Sebastien Pellat, Yannick Ormen, Clément Prévost-Solié, Gilles Reymond, Fabien Pifferi, Fabienne Aujard, Anthony Herrel, et al. Etholoop: automated closed-loop neuroethology in naturalistic environments. *Nature methods*, 17(10):1052–1059, 2020. 1

[33] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016. 8

[34] David Rolnick, Andreas Veit, Serge Belongie, and Nir Shavit. Deep learning is robust to massive label noise. *arXiv preprint arXiv:1705.10694*, 2017. 6

[35] Bernard Rosner. Percentage points for a generalized ESD many-outlier procedure. *Technometrics*, 25(2):165–172, 1983. 6

[36] John R Stowers, Maximilian Hofbauer, Renaud Bastien, Johannes Griessner, Peter Higgins, Sarfarazhussain Farooqui, Ruth M Fischer, Karin Nowikovsky, Wulf Haubensak, Iain D Couzin, et al. Virtual reality for freely moving animals. *Nature methods*, 14(10):995–1002, 2017. 1

[37] Alexandra Swanson, Margaret Kosmala, Chris Lintott, Robert Simpson, Arfon Smith, and Craig Packer. Snapshot serengeti, high-frequency annotated camera trap images of 40 mammalian species in an african savanna. *Scientific data*, 2(1):1–14, 2015. 1, 2

[38] Leslie M. Theunissen and Nikolaus F. Troje. Head Stabilization in the Pigeon: Role of Vision to Correct for Translational and Rotational Disturbances. *Frontiers in Neuroscience*, 11, 2017. 7

[39] Devis Tuia, Benjamin Kellenberger, Sara Beery, Blair R. Costelloe, Silvia Zuffi, Benjamin Risse, Alexander Mathis, Mackenzie W. Mathis, Frank van Langevelde, Tilo Burghardt, Roland Kays, Holger Klinck, Martin Wikelski, Iain D. Couzin, Grant van Horn, Margaret C. Crofoot, Charles V. Stewart, and Tanya Berger-Wolf. Perspectives in machine learning for wildlife conservation. *Nature Communications*, 13(1):792, Feb. 2022. Number: 1 Publisher: Nature Publishing Group. 1

[40] Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. The inaturalist species classification and detection dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8769–8778, 2018. 1, 2

[41] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The Caltech-UCSD Birds-200-2011 Dataset, July 2011. Issue: 2010-001 Num Pages: 8 Number: 2010-001 Place: Pasadena, CA Publisher: California Institute of Technology. 1, 2, 5

[42] Urs Waldmann, Hemal Naik, Nagy Máté, Fumihiro Kano, Iain D Couzin, Oliver Deussen, and Bastian Goldlücke. I-muppet: Interactive multi-pigeon pose estimation and tracking. In *DAGM German Conference on Pattern Recognition*, pages 513–528. Springer, 2022. 2

[43] Tristan Walter and Iain D Couzin. Trex, a fast multi-animal tracking system with markerless identification, and 2d estimation of posture and visual fields. *Elife*, 10:e64000, 2021. 1, 2, 3

[44] Hal Whitehead. Analysing animal social structure. *Animal behaviour*, 53(5):1053–1067, 1997. 3

[45] Jessie L Williamson and Christopher C Witt. A lightweight backpack harness for tracking hummingbirds. *Journal of Avian Biology*, 52(9), 2021. 4

[46] Shiting Xiao, Yufu Wang, Ammon Perkes, Bernd Pfrommer, Marc Schmidt, Kostas Daniilidis, and Marc Badger. Multi-view tracking, re-id, and social network analysis of a flock of visually similar birds in an outdoor aviary. *arXiv preprint arXiv:2212.00266*, 2022. 3

[47] Yuan Yao, Praneet Bala, Abhiraj Mohan, Eliza Bliss-Moreau, Kristine Coleman, Sienna M. Freeman, Christopher J. Machado, Jessica Raper, Jan Zimmermann, Benjamin Y. Hayden, and Hyun Soo Park. OpenMonkeyChallenge: Dataset and Benchmark Challenges for Pose Estimation of Non-human Primates. *International Journal of Computer Vision*, Oct. 2022. 2

[48] Hang Yu, Yufei Xu, Jing Zhang, Wei Zhao, Ziyu Guan, and Dacheng Tao. Ap-10k: A benchmark for animal pose estimation in the wild, 2021. 2

[49] Silvia Zuffi, Angjoo Kanazawa, David W Jacobs, and Michael J Black. 3d menagerie: Modeling the 3d shape and pose of animals. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6365–6373, 2017. 3