Multi-view Tracking, Re-ID, and Social Network Analysis of a Flock of Visually Similar Birds in an Outdoor Aviary

Shiting Xiao[1] · Yufu Wang[1] · Ammon Perkes[2] · Bernd Pfrommer[1] · Marc Schmidt[2] · Kostas Daniilidis[1] ·
Marc Badger[1]

Abstract

The ability to capture detailed interactions among individuals in a social group is foundational to our study of animal behavior and neuroscience. Recent advances in deep learning and computer vision are driving rapid progress in methods that can record the actions and interactions of multiple individuals simultaneously. Many social species, such as birds, however, live deeply embedded in a three-dimensional world. This world introduces additional perceptual challenges such as occlusions, orientation-dependent appearance, large variation in apparent size, and poor sensor coverage for 3D reconstruction, that are not encountered by applications studying animals that move and interact only on 2D planes. Here we introduce a system for studying the behavioral dynamics of a group of songbirds as they move throughout a 3D aviary. We study the complexities that arise when tracking a group of closely interacting animals in three dimensions and introduce a novel dataset for evaluating multi-view trackers. Finally, we analyze captured ethogram data and demonstrate that social context affects the distribution of sequential interactions between birds in the aviary.

✉ Marc Badger
mbadger@seas.upenn.edu

Shiting Xiao
gxiao@seas.upenn.edu

Yufu Wang
yufu@seas.upenn.edu

Ammon Perkes
aperkes@sas.upenn.edu

Bernd Pfrommer
pfrommer@seas.upenn.edu

Marc Schmidt
marcschm@sas.upenn.edu

Kostas Daniilidis
kostas@cis.upenn.edu

[1] Department of Computer and Information Science, University of Pennsylvania, Philadelphia, PA, USA

[2] Department of Biology, University of Pennsylvania, Philadelphia, PA, USA

1 Introduction

In social animals, moment-to-moment interactions among individuals drive the formation of long-term social networks. In turn, both an animal's position in the social network and its immediate social context change how it behaves and interacts with others (e.g. Anderson et al.,2021; White, 2010). The dynamics of a group's social network drives how individuals access food, shelter, and mates, and ultimately determines the group's reproductive success (Kohn et al., 2013). As we work toward a quantitative understanding of social behavior, it is essential that we develop animal and engineering systems for studying the interplay between the behavior of individuals and group dynamics.

Capturing the dynamics of social networks is not an easy task. Individuals must be accurately tracked and re-identified over long time periods and interactions between individuals must be detected and characterized to create an ethogram, or record of salient behaviors and their timestamps, for all individuals. Manual focal sampling by behavioral experts is one way of creating ethograms, but such efforts only capture a small slice of important behaviors for a few individuals at a time. Many recent works have developed automated systems

supporting the creation of behavioral ethograms, including those focusing on 2D tracking and re-ID (Pérez-Escudero et al., 2014; Romero-Ferrero et al., 2019; Walter & Couzin, 2021), pose estimation in 2D (Mathis et al., 2018; Lauer et al., 2022; Segalin et al., 2021; Pereira et al., 2019, 2022; Graving et al., 2019; Chen et al., 2020), and 3D (Gosztolai et al., 2021; Bala et al., 2020; Joska et al., 2021; Dunn et al., 2021; Günel et al., 2019; Badger et al., 2020; Wang et al., 2021; Zuffi et al., 2019), behavioral mapping (Berman et al., 2014), and analysis of collective behavior (Heras et al., 2019; Katz et al., 2011; Evangelista et al., 2017).

Of foundational importance to all multi-animal pipelines is the ability to track and re-identify individuals. With a few exceptions (Graving et al., 2019; Joska et al., 2021; Badger et al., 2020), current systems have only been deployed and tested in 2D settings with consistent lighting and static backgrounds, which make the problems of detection and tracking significantly easier. Interesting social dynamics, however, usually do not occur in isolation. Instead, they are embedded in the surrounding 3D environment, which introduces many challenges for automated perception. Groups of interacting animals spread over regions orders of magnitude larger than their body size, requiring many cameras to capture details for every individual. Individuals may be visually similar, yet their appearance may change dramatically as they move in 3D, puff their fur, or fluff their feathers. Variable lighting further alters the appearance of individuals. Backgrounds are visually complex and dynamic, and animals are frequently occluded by each other and structures in the environment. Many animals also have multimodal motion distributions making tracking extremely difficult. The extent to which automated systems can overcome these difficulties and capture groups of animals interacting within large and complex 3D environments is not well understood.

In this work, we aim to study behavioral dynamics in a socially gregarious species of songbirds (White et al., 2012; Maguire et al., 2013). We present 1) approaches for tracking a flock of birds and capturing their social interactions in a dynamic, multi-view setting, and 2) a new challenging dataset for evaluating the real-world performance of multi-view multi-object trackers.

Tracking in 3D is a complex problem. Some methods perform 3D reconstruction followed by tracking (Reconstruction-then-Tracking, or RT) and other methods first form tracks in 2D and then associate the tracks across views (Tracking-then-Reconstruction, or TR) (Wu et al., 2009). The advantage of performing reconstruction first is that tracking ambiguities are much less common in 3D than in 2D, so associating detections across time is far easier in 3D. On the other hand, matching sequences of points from 2D tracks improves cross-view association by reducing the potential for false matches,

which create ghost trajectories. When used for tracking bats, these two approaches show a tradeoff between the number of track fragments and false positive tracks (Wu et al., 2009) and the best-performing approach will depend on both camera geometry and the performance of the 2D tracker. We implement two RT approaches because the camera views frequently contain many occlusions and the baseline 2D trackers such as SORT (Bewley et al., 2016) did not perform well under these situations.

Our first approach uses foreground masks to construct a 3D pointcloud, which is then clustered to form points for tracking in 3D. Our second approach performs stereo matching of detections across views to reconstruct 3D points. In both approaches, 3D points are subsequently linked over time to form tracks using a motion prior. We test the performance of both trackers on an evaluation dataset containing long trajectories (∼36000 frames) with sparse 3D annotations and ground truth identities.

Our evaluation dataset includes a challenge task along with code for loading and viewing examples and evaluating performance on the task. In the task, which we call Where'd It LanD (WILD), the 3D locations of a single bird's head and tail are provided along with a sequence of frames. The tracker must then return the 3D location of the same bird's head at the end of the sequence as the target bird hops or flies with other birds in the aviary. Predictions are marked as correct if the returned 3D location is within a given threshold distance of the ground truth 3D location. Tracking performance is evaluated by the fraction of correctly predicted sequences across a range of distance thresholds. Finally, we use our dataset to perform a behavioral analysis of birds interacting in the aviary and show that social context influences the distribution of actions used by birds during courtship.

2 Contributions

1. A system for automatically extracting behavioral ethograms from a flock of birds interacting in an outdoor aviary. Components include synchronized camera and microphone array recording for months-long durations, and pipelines for detection, reconstruction, tracking, and re-identification.
2. An exploration of reconstruction-then-tracking approaches to multi-view multi-object tracking.
3. A unique dataset and codebase with tracking challenges for evaluating multi-view multi-object tracking algorithms.
4. An analysis of the social network of a flock of cowbirds showing how social context affects behavioral choices made by male and female birds during courtship.
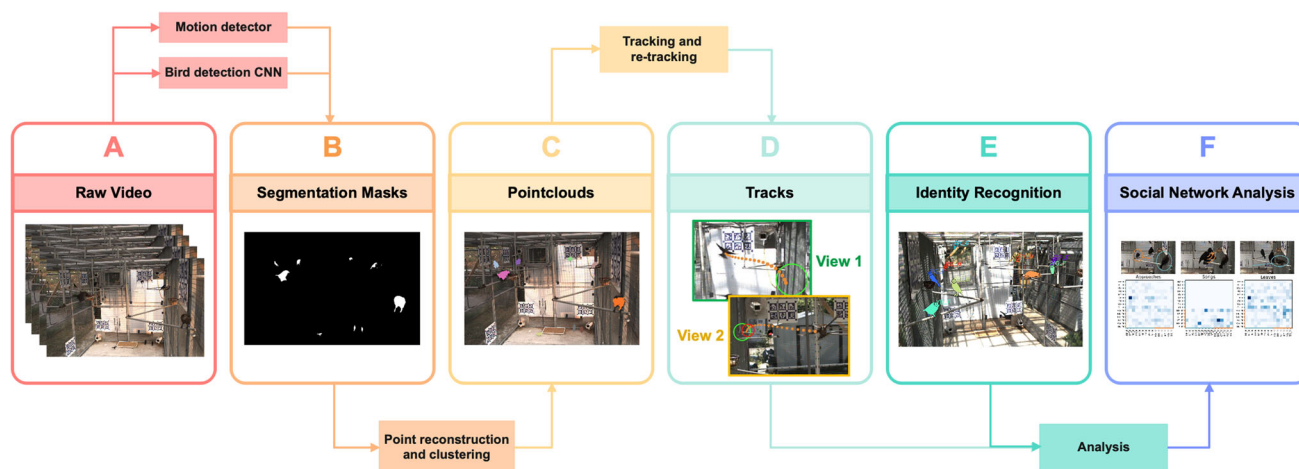
Fig. 1 Full pipeline for cowbird tracking and recognition. **A** A synchronized set of raw videos from multiple views are processed in a frame-by-frame manner. **B** Segmentation masks of bird instances are obtained using a Mask R-CNN network and background subtraction. **C** Pointclouds are reconstructed by multi-view matching, triangulation, and clustering. **D** Tracking, which is implemented using a Lagrangian Particle Tracking (LPT) algorithm, links pointclouds in time to form tracklets. Re-tracking associate 3D tracklets to generate longer 3D tracks. **E** Individual identity recognition using the FastReID framework. **F** Output from the pipeline can then be used for social network analysis

3 Related Work

3.1 Multi-object Tracking

3.1.1 Detection

Most state-of-the-art tracking methods follow the tracking-by-detection paradigm (Bergmann et al., 2019; Bewley et al., 2016; Karunasekera et al., 2019; Wojke et al., 2017; Wu et al., 2009; Cavagna et al., 2021; Sinhuber et al., 2019; Ling et al., 2018), in which the quality of detection is critical to the tracking performance. Convolutional Neural Network (CNN) based detectors (Girshick et al., 2014; Girshick, 2015; He et al., 2017; Ren et al., 2015; Liu et al., 2016; Redmon et al., 2016; Wang et al., 2020; Lin et al., 2017) have outperformed previous methods for object detection and instance segmentation tasks. In particular, the R-CNN family (Girshick et al., 2014; Girshick, 2015; He et al., 2017; Ren et al., 2015) find category-agnostic bounding box candidates, and then perform classification and refinement on them based on feature maps. A latest work Context R-CNN (Beery et al., 2020) keeps a "memory bank" based on contextual frames and uses attention to improve detection. SSD (Liu et al., 2016), the YOLO family (Redmon et al., 2016; Wang et al., 2020), and RetinaNet (Lin et al., 2017) directly regress to category-specific bounding box candidates. Detection can fail though, if an object's appearance changes dramatically between sightings. Unless enough examples are available in the training data, networks may not be robust to such changes. In the aviary, for example, motion blur caused by birds in flight is rare in training data and hence difficult to

detect. Background subtraction is a widely used technique to detect dynamically moving objects from static cameras. Zivkovic (2004) and Zivkovic and van der Heijden (2006) use a Gaussian mixture model that captures gradual changes in the background such as illumination changes, which is an important factor when running outdoor experiments where the sun is the light source. By using both a CNN based detector and a background subtraction based motion detector, we can reliably detect birds despite variations in their postures and movements.

3.1.2 Trajectory Generation

The ability to track an individual animal as it moves throughout its 3D environment is fundamental for addressing a broad range of questions in behavioural ecology and the study of animal social networks. Some interesting methods obtain 3D detections using point cloud observations from LiDAR data (Weng et al., 2020; Chiu et al., 2020; Yin et al., 2021), but obtaining such data is unrealistic in long-term wildlife monitoring. Recently, video data has become ubiquitous and indispensable in the study of collective behavior (Caravaggi et al., 2017; Schofield et al., 2019; Ling et al., 2018; Sinhuber et al., 2019). When individuals interacting in a 3D environment pass behind each other or objects in the environment, 2D occlusions occur. Because single camera views do not provide depth information, such occlusions create ambiguities and often result in lost tracks, identity swaps, or other tracking errors (Ciaparrone et al., 2020). Occlusions occur more frequently in crowded environments and identity swaps that occur during such occlusions can be difficult to recover from

if animals have similar appearances. An intuitive solution is to use multiple calibrated cameras and fuse information from different viewpoints to resolve ambiguities.

To track multiple objects in multiple camera views, data association must be performed not only across time (Tracking), but also spatially across views (Reconstruction). Doing reconstruction and tracking at the same time is computationally infeasible (Atanasov et al., 2014), so current methods typically adopt either a Tracking-then-Reconstruction (TR) route or a Reconstruction-then-Tracking (RT) route (Wu et al., 2009; Cavagna et al., 2021). TR methods first form 2D tracks in each camera views and then match them to reconstruct 3D tracks. Many state-of-the-art 2D tracking algorithms (Bergmann et al., 2019; Bewley et al., 2016; Karunasekera et al., 2019; Wojke et al., 2017) can be readily extended to track in 3D using cross-view data association techniques (Wang et al., 2014; Wu & Betke, 2016), but the complexity of most data association methods grows quickly with the number of simultaneously processed frames. Working in the 2D space, TR methods also have to handle both 2D and 3D occlusions in the reconstruction procedure (Cavagna et al., 2021; Wu & Betke, 2016).

Conversely, RT methods first reconstruct 3D representations using cross-view matching techniques, and then link them in time to form 3D trajectories. 2D occlusions are solved during the reconstruction procedure, which is typically performed independently for each frame, so the complexity of RT methods is substantially lower than the TR methods. Sinhuber et al. (2019); Ling et al. (2018) associate detections from multiple camera views using the stereo matching technique and use predictive Lagrangian Particle Tracking (LPT) (Ouellette et al., 2006) to form short 3D trajectories, or tracklets. A re-tracking strategy (Xu, 2008) is then used to solve 3D occlusions and link these short tracklets to form longer trajectories. A recent RT work by Cavagna et al. (2021) reconstructs each target as a point cloud in 3D and resolves 3D occlusions by solving a partitioning problem through a semi-definite optimization technique. While this method has proven to be effective for tracking birds moving at non-zero velocities in a dense flock, it performs poorly and cannot separate birds that perch close together for minutes (several hundred frames) because the complexity of the partitioning problem becomes too high to be solved reliably. Beyond using simple 2D locations to reconstruct 3D representations of targets, other methods also encompass orientation (Cheng et al., 2015), keypoints (Dong et al., 2021), and deep appearance features (Dong et al., 2021; Zhou et al., 2015) to perform association across views. In this work, we only use 2D locations and masks to reconstruct the targets in 3D for simplicity and efficiency.

3.1.3 Datasets

State-of-the-art multi-object tracking (MOT) datasets predominantly target people and vehicles, motivated by surveillance and self-driving applications (Sun et al., 2020; Gan et al., 2021; Han et al., 2021). Datasets for animal tracking and related tasks are presented by a comparatively small amount of previous literature. Recent work AP-10K dataset (Yu et al., 2021) is the first large-scale benchmark for mammal animal pose estimation which consists of 10,015 images from 23 animal families and 54 species. The OVIS dataset (Qi et al., 2021) for video instance segmentation consists of 20 animal species in hundreds of occluded scenes. Recently, a larger scale dataset for Tracking Any Object (TAO) (Dave et al., 2020) has been compiled containing 2,907 videos. We contribute our multi-view 3D tracking dataset of cowbirds for evaluating generalist trackers.

In the biology context, most behavioral studies acquire the dataset with carefully designed lab conditions: ideal illumination, arenas with a plain background, and well-quantified or no environmental stimuli (Sinhuber et al., 2019; Pérez-Escudero et al., 2014; Romero-Ferrero et al., 2019). While well-defined lab environments make it easier for tracking the objects, they restrict the complexity of the objects' movements that can be measured. Birds, in particular, exhibits rich postures and movements. Current datasets for the tracking of birds, however, contain only scenarios of bird flocks in migration Ling et al. (2018); Wu et al. (2014). In contrast, our multi-view tracking dataset contains large variation in bird pose, orientation, appearance, and social interaction across different lighting conditions that characterize "wild" footage.

3.2 Animal Re-Identification

In spite of the vast literature on multi-object tracking, handling occlusions remains the biggest challenge, especially in crowded scenes. Visual appearance features can aid frame-to-frame association (Wojke et al., 2017a; Romero-Ferrero et al., 2019; Pereira et al., 2022), and the ability to re-identify (re-ID) an individual animal upon re-encounter is extremely helpful in preserving the correct identities after occlusions. However, few ecological studies have taken advantages of the deep learning re-ID methods despite their success in human re-ID (Schneider et al., 2018). More recently, Schofield et al. (2019) used a variant of the VGG-M architecture (Chatfield et al., 2014) for both identity and sex classification of wild chimpanzees. When pre-trained on the ImageNet dataset, the VGG19 CNN architecture (Simonyan & Zisserman, 2014) can recognize individuals within small groups of birds (Ferreira et al., 2020) and giant pandas (Hou et al.,

2020). While classification approaches have demonstrated good overall performance (Luo et al., 2019) and can generalize across age-related changes in individual appearance (Schofield et al., 2019), the extent of their generalizability to unseen individuals in a small dataset (small in the number of individuals and training examples) is an important question that remains unexplored. Deep metric learning approaches, on the other hand, have shown good generalization across difference individuals and datasets (Yi et al., 2014; Zou et al., 2021). Here we collect a dataset for bird re-identification and train an identity embedding network using both metric-learning-based and classification-based losses (Luo et al., 2019).

4 Data Collection

4.1 Aviary

Many songbird species exhibit complex social structures, including the highly gregarious brown-headed cowbirds (*Molothrus ater*). Cowbirds present an excellent study system because exhibit complex patterns of behavioral interactions and the dynamics and structure of a group's social network predicts overall reproductive success (Kohn et al., 2013). Interactions between birds occur on timescales ranging from seconds to months. In just a few seconds a male could sing aggressively towards another male and then fly toward and land near a female, who then might make a chatter vocalization, lunge at the male, or fly away. Through hundreds of these interactions pair bonds between males and females emerge and a stable social network forms over the course of the three month breeding season. Several interesting questions remain unanswered, including what interactions influence the formation of pair bonds between males and females, how these interactions change over time, and how female feedback and multi-way interactions influence the development of the social network throughout the breeding season. Furthermore, these dynamics and the possible quantification of the social network will allow for eventual neurobiological studies that probe the influence of social context on brain dynamics in a naturalistic environment. To address these questions, we studied a flock of 15 cowbirds housed in a large outdoor aviary.

The UPenn Aviary is a covered outdoor arena (length × width × height: 6 × 2.4 × 2.4 meters) enclosed by rigid wire mesh. Inside are 12 central perches (located 40 cm below the ceiling) and 8 additional perches on the long sides (50 cm below the ceiling) of the aviary (see Fig. 4b,c for a diagram). Each corner has one camera (BLFY-PGE-23S6C with a Kowa 12.5 mm C-Mount lens) pointing inwards. The height × width field of view of the cameras is approximately 31 × 48 degrees and they are angled so that all points in the aviary

Fig. 2 Variation of captured images. Lighting and background appearance varies widely across viewpoint, time of day, and season throughout the birds' breeding period

volume can be observed by at least two cameras. Ten of the twelve central perches can be seen by all four top cameras. The bottom four cameras capture birds when they descend the ground to feed or bathe. Cameras are synchronized by a hardware trigger and capture 1920 × 1200 pixel frames at 40 Hz, which are sent over Gigabit Ethernet to a central server. Cameras are calibrated using a standard checkerboard (intrinsics) and an array of 96 AprilTags (Krogius et al., 2019) printed on 16 aluminum boards attached to the walls of the aviary (extrinsics). The aviary also captures audio signals using an array of 24 microphones (Behringer ECM8000), which are organized in eight triplets (with ∼ 10 cm between microphones within a triplet) around the exterior of the aviary and sampled at 48 kHz. The server writes all camera and microphone messages and their timestamps to one ROS bag (Quigley et al., 2009) for each day of recording.

Using the recording system described above, we recorded a flock of 15 interacting cowbirds (*Molothrus ater*) for approximately 16 hours per day for 104 days (March 16, 2019 to June 28, 2019). Captured images varied significantly in appearance across views and with the time of day, weather, and season (Fig. 2). There were six males and nine females in the flock. Males have black bodies with dark brown plumage on their heads and are larger than females, which are brown colored with lighter gray-brown breasts (see Fig. 4a for examples). We banded the left and right legs of each bird with a unique color combination drawn from blue, teal, green, pink, red, and yellow colors. Leg bands were approximately 1 cm in diameter and birds could be manually identified from nearby cameras whenever there was sufficient lighting and their bands were not occluded. Birds usually perched on the perches when not flying around the aviary, but they occasionally perched on the walls or walked along the floor between food and water trays. Perching periods varied dramatically, lasting from a fraction of a second to over 15 minutes. During long periods of perching, shadows shifted more rapidly than the birds themselves. In flight, however, birds crossed the 6 meter aviary in about 1 second (40 frames) and moved more than a body length between consecutive frames.

4.2 Multi-view Multi-bird Dataset and Challenge Tasks

Our dataset for multi-view multi-object tracking originates from four 15 minute segments drawn from one day in early April and two days in mid May. We chose these months because we expected to see rapid change in the social network across this period. The social network, including pair bonds, is not yet formed in April but solidifies by mid-May. Because cowbirds' behavior in the aviary makes it relatively easy to annotate periods of perching, we chose to annotate the beginning and end of these stationary periods for every bird in the aviary.

Each annotation effort began by selecting a bird and viewing a synchronized multi-view recording from the aviary in the VIA Video Annotator (Dutta & Zisserman, 2019). Once a bird stopped flying or walking (e.g. by landing on a perch), the center of the bird's head and the tip of its tail were clicked in at least two views. Very small motions during stationary periods (< 10 cm), such as steps along the same perch, were annotated with midpoints. Just before the bird started its next flight, its head and tail were annotated and labeled as an end point of the stationary sequence. The bird was then followed visually in flight until it landed again and a new stationary sequence was started. A behavioral annotation was also created whenever a target male sang. We ignore female chatter vocalizations because the visual chattering cue is subtle and annotators had a hard time assigning chatter when the female was not close to the camera. We plan to incorporate sound detection and localization to reliably assign chatters in future work. We confirmed the identity of each bird whenever both its leg bands were visible. All 15 birds in all four segments were positively identified and no two birds in the same segment were given the same identity. After all birds were annotated for a given segment, annotations were triangulated to obtain a sparse sequence of 3D locations and body axis orientations for each bird. For stationary segments, the positions of the head and tail were interpolated between the start and endpoints (using any available midpoints). Annotations were inspected for tracking errors (ID swaps or merges) by plotting pairwise distances between all birds. Whenever the distance between any two birds became less than 15 cm, the annotations were manually checked to ensure that trajectories had not merged (i.e. that no identity merge had occurred during manual annotation). From the annotations, we extracted 1098 stationary sequences of widely varying length. Averaged across birds, the 10th, 50th, and 90th percentiles of stationary sequence length were 3.7, 17.6, and 165 seconds respectively. These stationary sequences were used to form a training dataset for re-ID described below.

Untracked periods between stationary sequences were collected to obtain 986 motion sequences and formed our "Where'd It LanD" or WILD challenge. Each motion
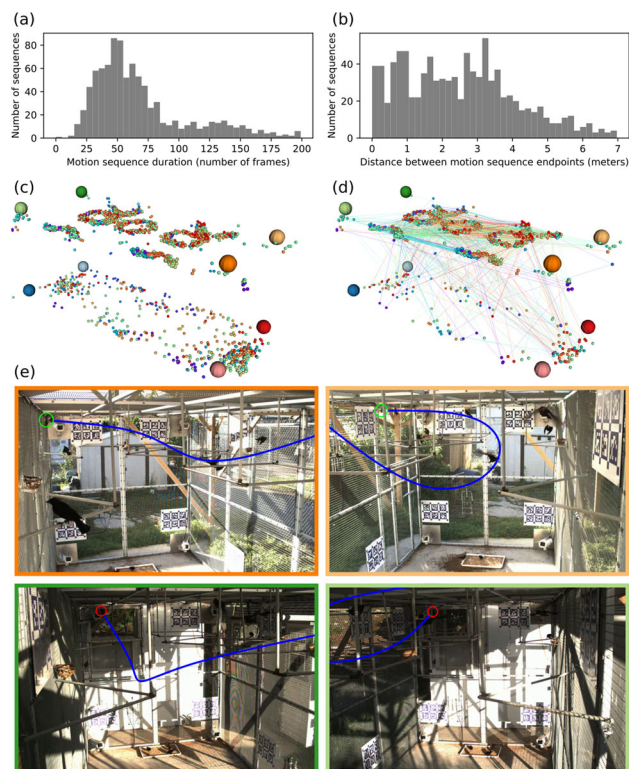

Fig. 3 The WILD dataset. Motion sequences are usually between 15 and 200 frames (**a**) between endpoints separated by 0-6 meters (**b**). In a reconstruction of all stationary sequence start and end points (**c**), areas of high point density reveal the perch geometry and ground plane. Lines between motion sequence start and end points (**d**) reveal flights from perch to perch, and from perches and the ground. Lines connect start and end points belonging to the same sequence; they do not indicate the actual trajectories. Points in (**c**) and (**d**) are colored by bird ID. Large spheres show the locations of the camera centers. An example from the dataset (**e**) shows the target bird's start location (green), approximate flight trajectory (blue), and ending location (red). Image borders denote the camera and correspond to large sphere colors in (**c, d**)

sequence is annotated with 3D start and end points (Fig. 3e; the endpoint of a stationary sequence serves as the start point of the following motion sequence). Averaged across birds, the 10th, 50th, and 90th percentiles of motion sequence length were 0.88, 1.6 and 4.5 seconds (35, 63, and 180 frames) respectively (Fig. 3a). The average number of motion sequences per bird was 66 (minimum: 8, maximum: 269) or an average total duration of 157 seconds per bird (minimum: 15.5 s, maximum: 552 s). The mean distance between motion sequence endpoints was 1.9 m (Fig. 3b, d).

Motion sequences in WILD vary dramatically in difficulty. In "easy" examples, a bird might hop between two perches and the entire sequence can be seen from the same set of cameras (e.g. Fig. 4b). In more challenging examples, birds change direction multiple times, fly behind other birds or through dark areas, or land in areas that are not visible by the original set of cameras (e.g. Fig. 4c). In the most difficult cases, birds might be fully visible by only one camera and be

partially or fully occluded from view by a second camera, and might then fly and land in an opposite corner of the aviary, where they are not visible by the original set of cameras (e.g Fig. 3e).

As part of the WILD challenge, we provide a data loader that takes in an example index and returns metadata, 3D start and end points of the target bird and an iterator containing the sequence of synchronized multi-view frames. We also provide an example visualization script that creates a video showing the start and end points of a sequence reprojected onto all visible views. Finally, we provide an evaluation script that takes in a list of indices and predicted 3D endpoint locations and returns the fraction of correctly predicted sequences using several distance thresholds.

5 Multi-view Multi-bird Tracking

5.1 Approach

We present an automated pipeline that can detect and track multiple cowbirds from raw video footage and demonstrate its use on the WILD challenge. The pipeline consists of the following components: (A) detection of bird instances, (B) 3D position recovery based on point cloud reconstruction and clustering, (C) 3D tracks generation using a predictive Lagrangian Particle Tracking (LPT) algorithm, and (D) occlusion handling in a re-tracking procedure.

5.2 Detection

We use a Mask R-CNN network pretrained on COCO instance segmentation to localize bird instances. Similar to our previous work (Badger et al., 2020), we removed weights for non-bird classes (leaving bird and background) and then fine-tuned all layers on on the Aviary Dataset (Badger et al., 2020). While Mask R-CNN would be robust to variations of bird postures given enough training examples, it is not reliable when detecting birds in certain postures which are rarely seen in training data, such as birds in flight with motion blur. To account for this issue, we add a background subtraction module (Zivkovic, 2004) to detect flying birds. For each frame in a raw video, we first convert it to a grey scale image, and then remove stationary features from the scene, eg. the aviary settings and gradual changes in illumination, adaptively learned from 500 temporally consecutive frames using Gaussian mixture probability density. We then segment the foreground image into distinct blobs of pixels corresponding to bird instances. However, shadows often move faster than perched birds, so pure background subtraction is not reliable when capturing birds that remain stationary during a substantial part of the video footage. We therefore exploit advantages of both Mask R-CNN detector and motion-based

detector, keeping a union of their detections without duplicates as inputs to the next stage of the pipeline. By combining the two methods, we are able to reliably detect birds both in stationary and in motion.

5.3 Reconstruction

The reconstruction step aims to recover the 3D positions of the detected instances. We first reconstruct dense 3D point clouds, and subsequently perform clustering to obtain cluster centers as the 3D positions of the birds. This point cloud reconstruction step is essential in the tracking pipeline because we found in experimentation that reconstructing the 2D center points alone is not able to sufficiently represent the 3D location of the bird. The single point representation is extremely sensitive to the quality of detections. As the shape of the bird changes dramatically during flight, shape of the bounding boxes and segmentation masks vary between frames (see Fig. 5a), and accordingly, the center/weighted centers of the bounding boxes/segmentation masks shift a lot. This adds additional unstableness to tracking. In our experiments, representing the bird by taking the center of the dense cloud of 3D points is more smooth and stable.

We use a similar method to (Cavagna et al., 2021) to reconstruct 3D point clouds. At each instant of time, given a union of segmentation masks from each camera view, we find matched pairs of active pixels from 2 distinct camera views based on epipolar distance. In the aviary, a region can be seen in another 2-4 camera views. We consider a pair to be a good match if it satisfies the trifocal constraint (Hartley & Zisserman, 2003) with another active pixel from one of those views. The matched pairs of pixels are then triangulated using a standard DLT method (Hartley & Zisserman, 2003). A potential challenge may occur if a bird were to enter the camera view at an extremely near distance, which results in a big mask with a large number of pixels that could blow up the memory. To solve this, one could sub-sample a mask if the number of pixels in it exceeds certain number. After reconstructing all 3D points, ghost points due to bad triangulation or false detection are filtered temporally if their nearest neighbor cannot be found in the neighboring frames.

We then cluster the 3D point clouds using the DBSCAN clustering algorithm. Centers of the clusters are the inputs to the tracking algorithm described in the next subsection.

5.4 Tracking

Once the 3D positions of the detected bird instances are reconstructed at each instant of time, we link them in time through an LPT (Ouellette et al., 2006) procedure. This tracking method has been successfully applied to study dynamic behaviour in aggregations of animals, including swarms of

midges (Sinhuber et al., 2019) and flocks of birds (Ling et al., 2018).

At a generic time t, let $\mathbf{x}_i^t$ denote the ith 3D point. The objective of the tracking problem is to find an $\mathbf{x}_j^{t+1}$ for every $\mathbf{x}_i^t$ such that $\mathbf{x}_j^{t+1}$ corresponds to the 3D location of the point at time $t+1$ that was at position $\mathbf{x}_i^t$ at time t. We define ϕ_{ij}^n to be the cost of associating each pair of $\mathbf{x}_i^t$ and $\mathbf{x}_j^{t+1}$. As this multidimensional assignment problem and is known to be NP-hard (Ouellette et al., 2006), minimizing the overall cost spanning hundreds of frames is computationally expensive. Therefore, we limit the temporal association to only a few frames at a time.

We generate 3D trajectories for each individual in the following two stages:

1. *Tracking*: Associate 3D points in time to form short tracklets in a frame-by-frame manner. At first instant of time, $t = 1$, we perform Hungarian matching based only on the distance between points as there's no dynamic information from the past. For each matched pair of points, we add a velocity vector to points at $t = 2$ defined as follows:

$$\mathbf{v}_j^2 = \frac{1}{\Delta t}(\mathbf{x}_j^2 - \mathbf{x}_i^1) \tag{1}$$

Starting from $t = 2$, we estimate the expected position of each particle in the future frame as

$$\mathbf{p}_i^t = \mathbf{x}_i^t + \mathbf{v}_i^t \Delta t \tag{2}$$

We define the cost of association ϕ_{ij}^n to be the distance between particles $\mathbf{x}_j^{t+1}$ and the estimated position $\mathbf{p}_i^t$. A particle can be linked to the tracklet if the cost of linking is below a set threshold. The velocity corresponding to point $\mathbf{x}_j^{t+1}$ can be calculated as

$$\mathbf{v}_j^{t+1} = \frac{1}{\Delta t}(\mathbf{x}_j^{t+1} - \mathbf{x}_i^t) \tag{3}$$

If multiple particles can be linked to the same tracklet, we stop the tracklet and start new ones. We set the threshold conservatively to minimize false linking. This results in shorter tracklets, which will be further connected in the re-tracking procedure described next. At last, the position and velocity of each point in a tracklet will be smoothed by a one dimensional Gaussian Filter (Mordant et al., 2004).

2. *Re-tracking*: Associate 3D tracklets to generate longer 3D tracks. All tracklets generated from the last stage are projected forward and backward in time using the positions and velocities at the endpoints (Xu, 2008). If distance between a forward projection of one tracklet is

close to the backward projection of another tracklet, the two tracklets are joined. When there are multiple possible matches, closeness of the velocity vectors is used to determine the best match. In addition, we handle the transient disappearance and appearance of a particle from the field of view due to miss detection by extrapolation based on its previous motion history. At last, trajectories shorter than 10 frames are removed from the final set to avoid ghost trajectories.

Generated tracks could be used to calculate motion priors of birds in the aviary, both of the collective as a whole as well as of the individuals.

Comparison with 2D tracking systems: the majority of the tracking literature focuses on 2D tracking. Many recent methods are learning-based and require track-level training data that is expensive to annotate. A general framework that does not require track-level training is a tracking-by-detection approach, where 2D detections are temporarily linked to form 2D tracklets (Bewley et al., 2016). In the context of multi-animal tracking, popular systems such as DeepLabCut and SLEAP are based on tracking by detection (Lauer et al., 2022; Pereira et al., 2022).

An extension to the 3D scenario is to match 2D tracklets across views stereoscopically to form 3D tracklets. But there are many complications in our setting: 2D tracking in a single view is less robust in complex scenes (Fig. 3); matching tracks across view is an NP-hard multidimensional assignment problem that can produce duplicate 3D tracks; merging or deleting duplicates relays on heuristics. The complexity of the system quickly increased to cover many corner cases.

In contrast, tracking directly with a 3D representation reduces ambiguities, and the system is simple in principle. Recent 3D tracking papers have made a similar observation (Rajasegaran et al., 2021; Cavagna et al., 2021). A more comprehensive comparison is left for future research.

5.5 Re-ID with the Bird15 Dataset

To form a dataset for bird re-identification, we exported images from stationary sequences. Images were passed through the bird detector and the sequence annotations (ground truth locations and identities of perched birds) were used to assign an identity with each detection. We exported tight crops from all available views, except when two or more birds occluded each other, in which case only the crop for the bird closest to the camera was exported for that view. To improve the spatial and pose diversity of exported crops, we partitioned the aviary into 3D bins (10 cm side length) and tracked the number of crops exported for each bird in each bin. For each bird, we exported crops every 10 frames until the bin for that bird and location had 10 images. Once the bin was filled, we continued to export crops, but only every

40 frames. We use this method to bias collection towards a diversity of locations generated by brief periods of perching as birds move throughout the aviary. All crops were resized to 256x256 pixels. Image filenames contain bird ID, camera view, sequence number, and frame number information following the Market1501 format (Zheng et al., 2015).

We split the dataset into training and test sets, composed of crops obtained from the first half and second half of each 15 minute segment, respectively. The training and test sets each contain 18,000 images. Birds were fairly evenly represented in both sets (mean ± std. training images per bird: 1225±531, test images per bird: 1229±339), with the exception of one female with Red+Yellow leg bands, which only had four examples in the training set and 620 in the test set. The number of examples from each of the top cameras was similar between training and test sets, and was consistently higher than the number of examples from the bottom cameras (as expected based on the lack of visibility of the perches). We randomly selected 7,500 training images to serve as a validation set.

We then trained an embedding network for bird identification on the Bird15 dataset. The network consists of a ResNet50 (He et al., 2016) pre-trained on ImageNet, which takes in a 256×256 image and outputs a 2048 vector of re-ID features f, followed by a BNNeck (Luo et al., 2019) and a classification head, which outputs identity logits p. The network was supervised using both triplet (Weinberger & Saul, 2009) and cross-entropy identity losses and we used Adam and the FastReID codebase (Luo et al., 2019) to optimize the model. We use the default FastReID baseline "bag of tricks", except that we do not use horizontal flipping augmentation because bird identities depend on the ordering of the left/right leg band colors, which would be swapped upon reflection. During inference, we apply a softmax function to the logits p to obtain a distribution over bird IDs for each image.

6 Results and Experiments

6.1 Short-term Tracking of Individual Birds in Cluttered Scenes using WILD

Experiment. We tested our tracker on the WILD dataset. Among the 952 motion segments we evaluated against, 741 segments have short sequences of ≤ 100 frames, 186 segments have $100 \sim 300$ frames, and 25 have rather long sequences of ≥ 300 frames. For each motion segment, we provide the start and end locations of the target bird's head and tail points in 2D and 3D, as well as an iterator containing the sequence of synchronized multi-view frames. The task is to track the target bird and predict its 2D/3D position at the end of the sequence. The experiment was conducted as fol-

lows. We ran our multi-object tracker on the provided frame sequence and output a set of track hypotheses for all birds in the scene. At the start frame, we established correspondence between the target and the closest hypothesis based on 3D Euclidean distance, and at the end frame, we measured the 3D distance between the target's end location and the same hypothesis. All remaining hypothesis that were not associated with ground truth were ignored.

We compared our Pointcloud Reconstruction based tracker with the Stereo Matching method introduced by Ling et al. (2018). This method has been demonstrated to successfully resolve multi-view optical occlusions and improve tracking performance. The evaluation process for these two methods differs only in the point reconstruction stage, with the rest - detection and tracking - remaining the same (see Fig. 1ABCD). One major difference of these two methods is the way they represent each target in 3D. Taking only the center of the detection mask/bounding box as input, the Stereo Matching method reconstructs the target as only one single point in space. The Pointcloud Reconstruction method, on the other hand, reconstructs the target as a dense cloud of points.

Evaluation metric. The end position of the track hypothesis retrieved by our tracking pipeline, see Fig. 4, is compared with the ground truth end position. "AC0.X", the fraction of reconstructed hypotheses landing within 0.X meters of the ground truth, is reported in Table 1; its ideal value is equal to 100 percent. We chose this evaluation metric because distance based metrics were very sensitive to outliers. For example, samples that were not tracked successfully can land far away from the ground truth and end up dominating the average and inflating the standard deviation. We also provide an alternative evaluation metric in Table 3. In stead of using metric meter as error measure, we use the average height of the cowbirds (15 cm). For example, "AC2×" is the fraction of tracks landing within 30 cm of the ground truth. We do not evaluate the result using the standard CLEAR MOT evaluation method of Bernardin and Stiefelhagen (2008), because the MOT statistics are based on frame-by-frame annotations and the production of frame-by-frame 3D ground truth trajectories is currently severely limited by the amount of human effort and expertise required for manual annotation.

Result Analysis. We present qualitative results of the our tracker in Fig. 4. The quantitative results of both the Pointcloud Reconstruction method and the Stereo Matching method on the WILD dataset are reported in Table 1. The table shows that Pointcloud Reconstruction method outperforms the Stereo Matching method in every category. Video visualization shows that points reconstructed by Stereo Matching are more unstable than pointclouds, as the single-point representation is more sensitive to the quality of detections. A slight change of the detection (box size and

Fig. 4 Qualitative tracking results. **a** Examples of detected bird instances with variations in pose, shape, lighting, scale, occlusion, and motion blur. **b** Example of a successful short track (56 frames) followed by its 2D projections in 3 different views. The green cube/circle is the start 3D/2D position and the red cube/circle is the end position. Dots in the 2D images are smaller/larger as the bird gets further away/closer. **c** Example of a successful long track (375 frames). During flight, the individual hops on the wall and briefly pauses for 1–2 seconds. Examples in (**b**) and (**c**) are from video segments drawn from different days, demonstrating variable time of day and lighting

Table 1 Quality of the trajectories retrieved by Stereo Matching method and Pointcloud Reconstruction method. AC0.X denotes percent tracks land within 0.X meter, of the ground truth end position

Method	Length (# frames)	Segment Counts	AC0.1	AC0.3	AC0.5	AC1.0
Stereo	≤ 100	741	0.17	0.34	0.41	0.52
	100 ∼ 300	186	0.10	0.20	0.27	0.47
	> 300	25	0.04	0.08	0.20	0.28
Pointcloud	≤ 100	741	0.44	0.60	0.67	0.75
	100 ∼ 300	186	0.30	0.41	0.49	0.61
	> 300	25	0.16	0.28	0.32	0.44

**Table 2** Quality of the trajectories by our tracker assuming "oracle" matching through ambiguities

Length (# frames)	Segment Counts	AC0.1	AC0.3	AC0.5	AC1.0
≤ 100	741	0.50	0.73	0.78	0.87
100 ∼ 300	186	0.45	0.62	0.67	0.76
> 300	25	0.36	0.44	0.44	0.60

shape) in the next frame will result in very different 2D location of the center and resulting reconstructed 3D points.

As the tracking performance of the Stereo Matching method is significantly limited by the single-point representation, we restrict the following discussion to the Pointcloud Reconstruction method only. As Table 1 shows, most tracks are either successful with low error (44% of the short tracks land within 0.1m to the ground truth) or are not at all close (33% of the short tracks land more than 0.5m from the ground truth). Increasing the threshold does not increase the overall accuracy very much. Table 1 also shows that our tracker performs better on short segments than on the longer ones. To understand the influence of failures originating from ambi-

guities, we collected statistics of percent accuracy assuming "oracle" matching through ambiguities. That is, we kept all possible matches during the re-tracking stage, and linked them to the tracking hypothesis to form a tree structure. We counted a hypothesis as a success as long as one of it's leaf nodes landed within the threshold of the ground truth. Statistics are reported in Table 2. As the table shows, accuracy of the longer tracks has increased notably, indicating ambiguities are an important source of failure. This problem could be aided by re-ID or visual features as discussed in the next section.

Assuming failures are solely due to accumulating errors ambiguities or missed detections are encountered, if 44%
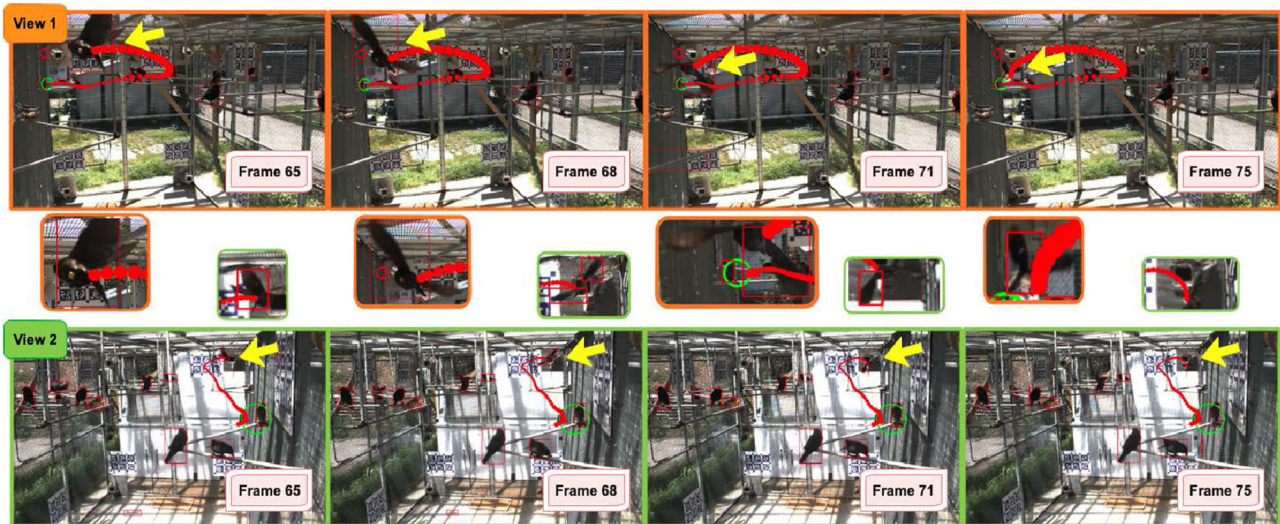
(a) Challenging tracking cases due to dramatic shape changes and motion blurs during flight, which may result in incomplete detections (view 1&2, frame 71), split detections (view 2, frame 68), and missed detections (view 2, frame 75). Shape changes also cause the reconstructed pointclouds to merge (small wingspread) and split (large wingspread), animated in Figure 6b.
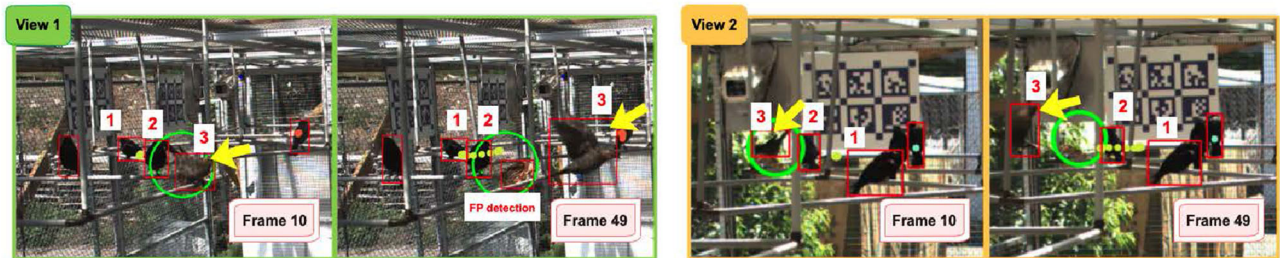
(b) Challenging tracking cases due to 3D occlusions. At the start of the motion segment, birds No.1, 2, and 3 are perching motionlessly for around 10 frames. After being reconstructed as dense point clouds in 3D, the clustering algorithm was not able to separate them as they are too close in 3D (see animation in Figure 6a). Our tracking algorithm failed to track the target bird No.3 because of the 3D occlusion and false positive detections (view 1, frame 49).
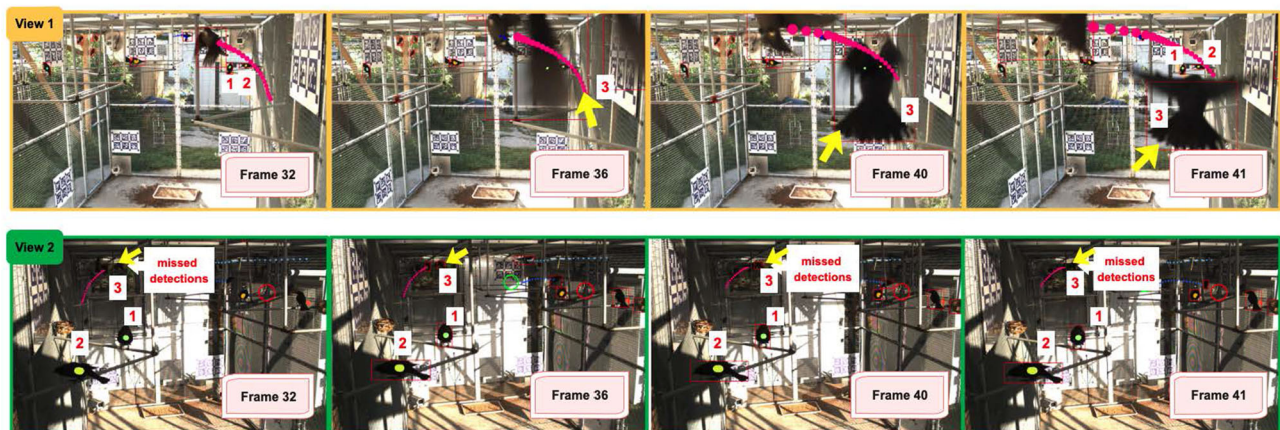
(c) Challenging tracking cases due to large variation in apparent sizes and extreme lighting conditions. The target bird No.3 flied right to the camera view 1, resulting in big apparent sizes in view 1 and small sizes in view 2. In some extreme poses only partial detections (view 1, frame 36) can be obtained. Under heavy shadow occlusions, our detection network and motion segmentation module both failed to detect target bird No.3 (view 2, frame 32, 40, and 41). Too many missed detections resulted in a tracking failure. However, our tracking pipeline was able to re-track birds that disappeared shortly. For example, birds No.1 and 2 were occluded by bird No.3 for a few frames (view 1, frame 36&40). Based on the velocity and position record of the bird, we linked the tracklets before and after occlusion and produced a continues 3D track (view 1&2).

Fig. 5 Challenges of multi-view multi-bird tracking. Yellow arrow indicates the target bird to track. Image borders denote the camera view and correspond to large sphere colors in Fig. 3(c,d)

Table 3 Alternative evaluation metric for the proposed tracking method, with the threshold for correct track being 1×, 2× and 3× of the average height of cowbirds. We used 15cm as the average height

Length (# frames)	Segment Counts	AC1×	AC2×	AC3×
≤ 100	741	0.52	0.60	0.66
100 ∼ 300	186	0.37	0.41	0.47
> 300	25	0.20	0.28	0.32
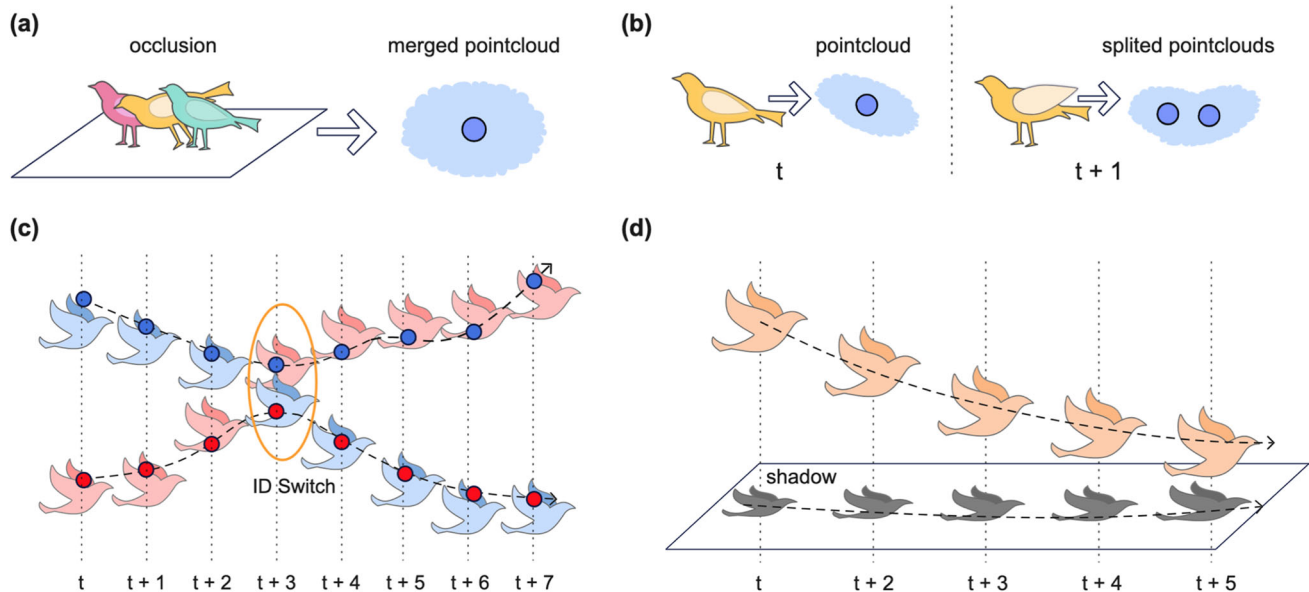
Fig. 6 Failure cases. **a** Inseparable pointcloud due to occlusions. **b** Merged/split clusters due to shape change of an individual at different instants of time, which could result in ghost trajectories. **c** Identity Switch. At first, the blue hypothesis is correctly tracking the ground truth blue bird. After a few frames, though, the blue bird and the red bird cross paths and blue hypothesis follows the wrong target. **d** Ghost trajectory resulting from false positive detections, eg. shadows of a bird

of tracks are successful for 100 frames, then we can expect only 19% of tracks to survive to 200 frames and 9% to survive to 300 frames. Because the performance is better than this expectation, it is possible that the tracker is struggling elsewhere. For example, during initialization, there might be no track available to assign to the target start, or the wrong track could be assigned to the target start. A discussion of the failure cases is provided in the next paragraph.

Failure cases catalogue. Our tracker produced many plausible results but also many failure cases, shown in Figs. 5 and 6. To better understand the nature of the complexity of the WILD dataset, we manually examined 20 failure cases by looking into the outputs (detections, pointclouds, tracklets) produced in each stage of the pipeline frame by frame. We found that the tracker struggles in the following cases:

1. Missed detections: extreme poses, occlusions from poles and other individuals, and extreme lighting conditions in the aviary occasionally cause the detector to fail (Fig. 5ac).

2. False positive detections: shadows of birds, for example, create ghost pointclouds and ghost trajectories (Fig. 6d). Nests are often falsely detected as birds too (5b).

3. An inseparable pointcloud due to occlusions (Fig. 6a): multiple targets in close 3D proximity can occlude each other in all camera views. They then become reconstructed as one pointcloud as a whole and share one track.

4. Merged and split pointclouds: when individuals change shape or size (Fig. 6b), pointclouds can split into two or more clusters. During flight, the appearance of a bird changes dramatically in a very short period of time (Fig. 4a), which results in differently shaped clouds of points. In many cases, points representing one bird are grouped into multiple clusters (Fig. 6b), which introduces unstable and unpredictable ghost pointclouds. Such instability increases the difficulty of tracking.

5. Identity switches: true identities of different hypotheses can become switched, particularly if two individuals remain directly next to each other for several seconds (Fig. 6c).
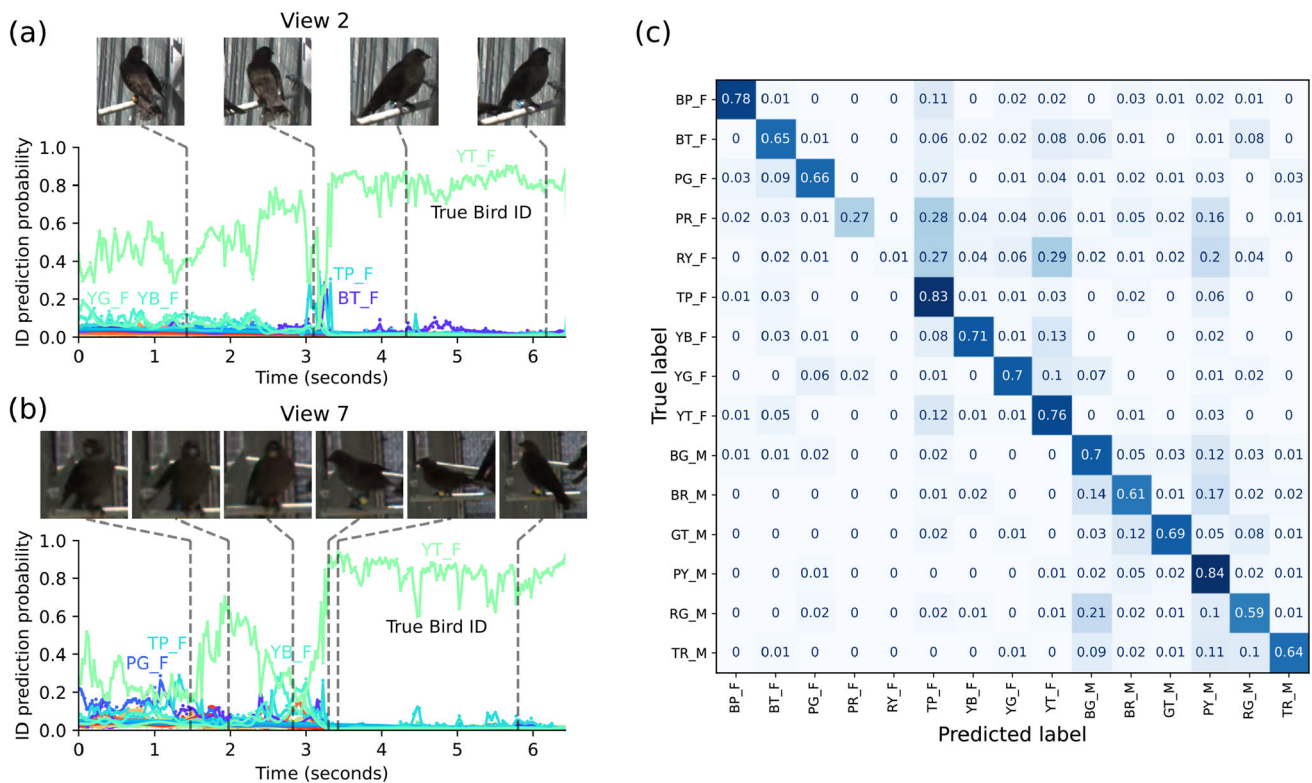
Fig. 7 Bird re-identification. We use a ResNet50 network supervised with triplet and ID losses to predict the identity of perched birds. In an example from the Bird15 test, a female with Yellow+Teal leg bands is visible from views 2 (**a**) and 7 (**b**). From view 2 (**a**) only its left leg band is initially visible, but the network has learned other features (such as tail shape, or background features if the bird is in a repeatedly used location) that allow it to correctly predict the identity. When no bands are visible (second image from the left in (**a**), the confidence decreases. Once both bands are visible (third and forth images) confidence increases again.

From another view (**b**), both bands are visible, but are in a shadow and some initial color distortion causes the network to incorrectly predict Pink+Green, Teal+Pink, and Yellow+Blue, albeit with low confidence. As the bird reorients to face the other direction, both bands become visible with better lighting and confidence increases. A normalized confusion matrix (**c**) shows most birds are correctly identified 60–80% of the time in the test set. Increasing the detection confidence threshold from 0 to 0.8 improves accuracy from 0.68 to 0.97 while still correctly identifying 52% of the examples in the Bird15 test set

6.2 Bird Re-identification

We evaluate the performance of the re-ID network using the Bird15 test set, which we constructed using the ground truth locations of perched birds. Overall, the network correctly identified 68% of examples in the test set and most individuals are identified correctly 60–80% of the time (Fig. 7c). Instead of returning whichever bird corresponds to the highest probability (even if it is very low), setting a detection confidence threshold to 0.8 increases the accuracy to 0.97 while correctly predicting 52% of samples in the test set. Most confusion appears to be within females and within males separately, with relatively low confusion between males and females. Unless lighting is very poor, males can usually be distinguished from females by their darker color.

When deployed on crop sequences from tracked birds (Fig. 7a,b), probability trajectories over time reveal interesting patterns of the re-ID network. From camera view 2 (Fig. 7a), the network predicts the correct identity despite only

being able to see one band (three other female birds have yellow bands). When both bands are hidden, however, the network becomes less confident. Interestingly, these observations suggest that the network has learned to rely on the bands, but that it has also learned to rely on additional features such as slight variations in bird color or patterning, or perhaps features of the background behind the favorite perch locations for each bird. This hypothesis could be tested by training on a masked dataset, where the network receives only pixels corresponding to the bird and no pixels from the background. Improving the diversity of perch positions by collecting additional annotations throughout the breeding season may also help improve the robustness of the bird re-ID pipeline.

6.3 Social Network Analysis

Using our dataset we analyzed the birds' social network and investigated how birds' behavior depends on social context.

In addition to human labeled song annotations, we also added "approach", "stay", "leave", and "sing to" interactions using the start and endpoints of the stationary sequences. Whenever a bird flew to a location within an interaction distance (0.5 meters) of another, we added a "b1 approached b2" annotation. Whenever a bird was within the interaction distance of another and flew away we added a "b1 left b2" annotation. Whenever a male sang, we added "b1 sang to b2" annotations for all birds within the interaction distance. Finally whenever a bird was approached, if it did not leave within one second, we add a "b1 stayed with b2" annotation (Anderson et al., 2021). After collecting the interactions between all pairs of birds, we grouped interactions depending on social context factors, such as those belonging to male-male interactions, or those between a pair-bonded or non-pair-bonded male and female. We defined a pair bond between a male and a female whenever the female received more than 50% of her total song interactions from that specific male (Anderson et al., 2021). From the sets of interactions, we constructed transition ethograms and inspected how the probabilities of interaction transitions changed with social context. We focus our analyses on two 15 minute segments with song annotations from mid May.

From the patterns of approaches and leaves, we observed differences in the overall activity levels of individuals (Fig. 8). Two females, F1 (TP_F) and F2 (YT_F), repeatedly flew back and forth among perches within the interaction distance of another female F3 (BT_F). The approach and leave interactions among males revealed that M1 (PY_M) frequently approaches other males (BG_M, BR_M, GT_M), shown as the darker PY_M row in the approaches matrix. At the same time, these three males frequently fly away from M1, shown as the darker PY_M column in the leaves matrix. These patterns clearly indicate that M1 is dominant over these other males.

From the song interaction data, we observed six pair bonds between males and females. Male M1 was bonded with two females (BP_F and YT_F). Similarly, male BG_M was bonded with two females (PG_F and YB_F). Finally, RG_M was bonded with RY_F, and TR_M was bonded with PR_F. Based on these pair bonds, we split the set of interaction transitions into pair bond and non-pair bond groups (Fig. 9). Inspecting the differences in transition probabilities of pair-bonded birds relative to non-pair-bonded birds (Fig. 9c) reveals that females are more likely to leave when approached by non pair bond males than when approached by their pair bond male. When a female stays with its pair bond male, the male is more likely to sing to her and less likely to leave than when a female stays near a non pair bond male. When a female leaves her pair bond male, the male is more likely to follow and approach her again, than when a female leaves a non-pair bond male.
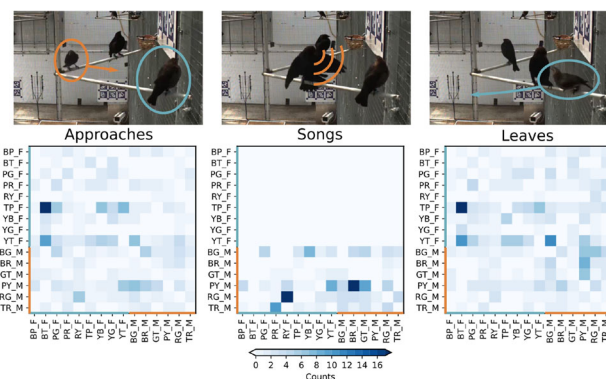

Fig. 8 Pairwise interactions. Approaches, songs, and leave interactions occur frequently between individuals in the aviary. Each matrix shows the frequency of interactions for each pair of individuals. The bird performing the action is shown on the left axis (the approaching, singing, or leaving bird) and the target or recipient of the action is shown along the bottom axis (the approached, receiving, or remaining bird). Orange indicates males and blue indicates females. Approaches and leaves show relative movement between individuals and reveal differences in activity levels and dominance (see Sect. 6.3). We also observed six pair bonds between males and females, which are defined whenever a female receives more than 50% of songs from a single male (Anderson et al., 2021)

It will be interesting to analyze how patterns of interaction vary throughout time of day and over the breeding season. For example, in one of the annotated 15 minute segments in April, males were actively singing for nearly the entire period, but we recorded very few flight sequences, leaves, and approaches because most birds remained on their perches. Without many more periods of observation, it will remain unclear whether such differences in interaction patterns are a normal part of social network formation, or whether they can be explained by other environmental variables such as time of day, temperature, and weather.

Finally, we anticipate that estimating the pose and shape of individuals in the aviary (Badger et al., 2020) will allow us to incorporate more fine-grained behaviors and interactions, such as the head-up display shown in Fig. 10.

7 Conclusion

In this work we develop a system for capturing the behavioral interactions of a group of 15 songbirds. Although we found that our pointcloud reconstruction method performed better than a stereo matching method, there is still much room for performance improvements on our difficult multi-view multi-animal Where'd It LanD (WILD) dataset. We introduce several complexities that arise when studying animals that maneuver and interact in three dimensions. Tracking many individuals across multiple sensors is a challenging task with points of failure. The relative lack of flying birds
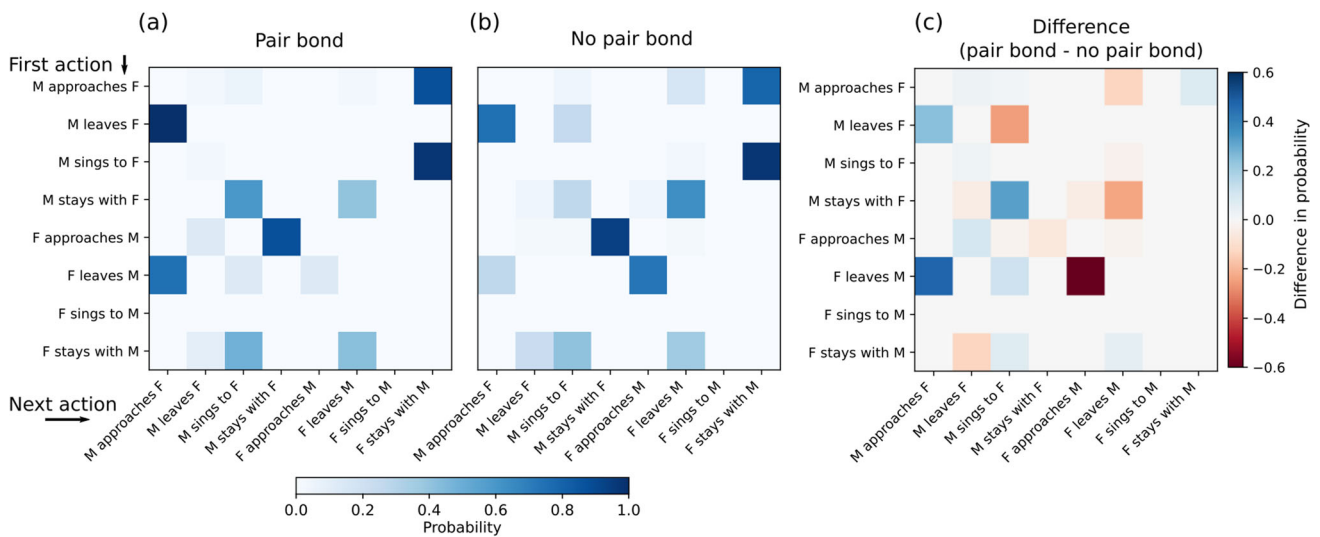
Fig. 9 Interaction sequences. Interaction transition probabilities differ between pair-bonded (**a**, n = 163 transitions) and non-pair-bonded (**b**, n = 187 transitions) males and females. For a given row, filled-in cells show interactions that occurred next based on their frequency in the dataset. Counts are normalized within rows and darker blue shows greater probability. (**c**) The difference in transition probabilities for bonded pairs relative to non-bonded pairs. Darker blue indicates a transi-

tion is more likely for a bonded pair than for a non-bonded pair; darker red indicates a transition is more likely for a non-bonded pair than a bonded pair. Transition probabilities reveal that pair-bonded females are generally more receptive to approaches by their pair bond male than by other males and that pair-bonded males are more likely to follow females with which they have formed a pair bond
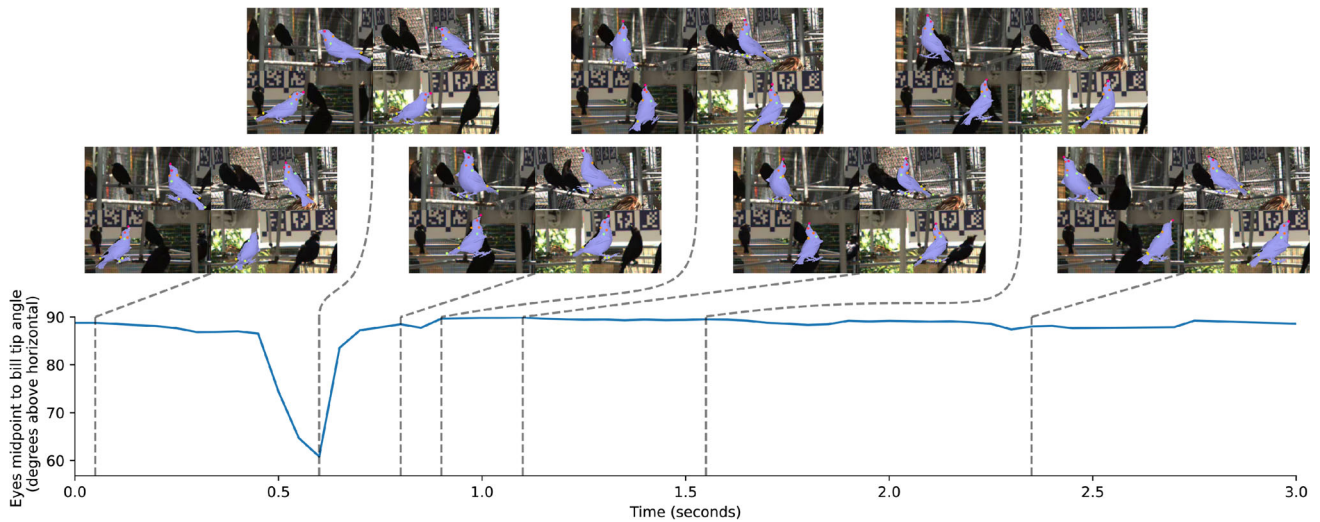

Fig. 10 Pose trajectories. Behaviors extracted from pose trajectories can reveal fine-grained interactions such as head-up aggressive displays by males. In every other frame, a three dimensional parameterized mesh (Badger et al., 2020) is fit to multi-view anatomical keypoints. In this

example, the angle between horizontal and the vector from the midpoint between the eyes to the bill tip (visualized in the plot) captures this behavior well

in our detection dataset (birds spend most of their time sitting perched) hindered our object detection pipeline and lead us to add the additional complexity of a motion detector. Replacing this motion detector with a neural network designed specifically for detecting objects in motion could significantly improve our pipeline by reducing the number of false positive detections (and ensuing ghost trajectories

and tracking failures) generated by background motion. We also found that birds occluded each other much more than expected because the perches were positioned only slightly below plane of the top cameras. We plan to improve the layout of the aviary in order to reduce such occlusions. We also highlight the need for additional work that integrates detection, tracking, re-ID, and pose estimation pipelines

without relying on extensively annotated tracking datasets, which become prohibitively expensive to create in multi-view multi-animal settings. Using our system and dataset of ground-truth identities, we developed a re-ID pipeline, extracted detailed ethograms for all birds in the aviary, and demonstrated that the presence of a pair bond changes the interaction dynamics between males and females.

Data and code availability Data and code will be made publicly available via Google Drive and GitHub.

Declarations

Conflict of interest The authors declare no competing or conflicts of interest.

Ethical approval The aviary and cowbird data collection were approved by the University of Pennsylvania Institutional Animal Care and Use Committee.

References

Anderson, H. L., Perkes, A., Gottfried, J. S., Davies, H. B., White, D. J., & Schmidt, M. F. (2021). Female signal jamming in a socially monogamous brood parasite. *Animal Behaviour, 172*, 155–169. https://doi.org/10.1016/j.anbehav.2020.10.011

Atanasov, N., Zhu, M., Daniilidis, K., & Pappas, G. J. (2014). Semantic localization via the matrix permanent. *Robotics: Science and Systems, 2*, 1–10.

Badger, M., Wang, Y., Modh, A., Perkes, A., Kolotouros, N., Pfrommer, B., & Daniilidis, K. (2020). 3d bird reconstruction: A dataset, model, and shape recovery from a single view. *Eccv*.

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 Communications, 11*(1), 4560. https://doi.org/10.1038/s41467-020-18441-5

Beery, S., Wu, G., Rathod, V., Votel, R., & Huang, J. (2020). Context r-cnn: Long term temporal context for per-camera object detection. *Cvpr*.

Bergmann, P., Meinhardt, T., & Leal-Taixé, L. (2019). Tracking without bells and whistles. *Iccv*. https://doi.org/10.1109/ICCV.2019.00103

Berman, G. J., Choi, D. M., Bialek, W., & Shaevitz, J. W. (2014). Mapping the stereotyped behaviour of freely moving fruit flies. *Journal of The Royal Society Interface, 11*(99), 20140672. https://doi.org/10.1098/rsif.2014.0672

Bernardin, K., & Stiefelhagen, R. (2008). Evaluating multiple object tracking performance: The clear mot metrics. *EURASIP Journal on Image and Video Processing, 2008*(1), 246309. https://doi.org/10.1155/2008/246309

Bewley, A., Ge, Z., Ott, L., Ramos, F., & Upcroft, B. (2016). Simple online and realtime tracking. In *2016 ieee international conference on image processing (icip)* (pp. 3464–3468). https://doi.org/10.1109/ICIP.2016.7533003

Bewley, A., Ge, Z., Ott, L., Ramos, F., & Upcroft, B. (2016). Simple online and realtime tracking. *Icip*

Caravaggi, A., Banks, P. B., Burton, A. C., Finlay, C. M., Haswell, P. M., Hayward, M. W., & Wood, M. D. (2017). A review of camera trapping for conservation behaviour research. *Remote Sensing in Ecology and Conservation, 3*(3), 109–122.

Cavagna, A., Melillo, S., Parisi, L., & Ricci-Tersenghi, F. (2021). Sparta tracking across occlusions via partitioning of 3d clouds of points. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 43*, 1394–1403.

Chatfield, K., Simonyan, K., Vedaldi, A., & Zisserman, A. (2014). Return of the devil in the details: Delving deep into convolutional nets. In *Proceedings of the British Machine Vision Conference (BMVC)*.

Chen, Z., Zhang, R., Eva Zhang, Y., Zhou, H., Fang, H.-S., Rock, R. R., & Lu, C. (2020). Alphatracker: A multi-animal tracking and behavioral analysis tool. bioRxiv. https://doi.org/10.1101/2020.12.04.405159

Cheng, X., Qian, Z.-M., Wang, S. H., Jiang, N., Guo, A., & Chen, Y. (2015). A novel method for tracking individuals of fruit fly swarms flying in a laboratory flight arena. *PloS One, 10*, e0129657. https://doi.org/10.1371/journal.pone.0129657

Chiu, H. -k., Prioletti, A., Li, J., & Bohg, J. (2020). Probabilistic 3d multi-object tracking for autonomous driving. arXiv preprint arXiv:2001.05673.

Ciaparrone, G., Luque Sánchez, F., Tabik, S., Troiano, L., Tagliaferri, R., & Herrera, F. (2020). Deep learning in video multi-object tracking: A survey. *Neurocomputing, 381*, 61–88. https://doi.org/10.1016/j.neucom.2019.11.023

Dave, A., Khurana, T., Tokmakov, P., Schmid, C., & Ramanan, D. (2020). Tao: A large-scale benchmark for tracking any object. *Eccv*.

Dong, J., Fang, Q., Jiang, W., Yang, Y., Huang, Q., Bao, H., & Zhou, X. (2021). Fast and robust multi-person 3d pose estimation and tracking from multiple views. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 44*(10), 6981–92.

Dunn, T. W., Marshall, J. D., Severson, K. S., Aldarondo, D. E., Hildebrand, D. G. C., Chettih, S. N., & Ölveczky, B. P. (2021). Geometric deep learning enables 3d kinematic profiling across species and environments. *Nature Methods, 18*(5), 564–573. https://doi.org/10.1038/s41592-021-01106-6

Dutta, A., & Zisserman, A. (2019). The VIA annotation software for images, audio and video. In *Proceedings of the 27th acm international conference on multimedia*. New York, NY, USAACM. https://doi.org/10.1145/3343031.3350535

Evangelista, D. J., Ray, D. D., Raja, S. K., & Hedrick, T. L. (2017). Three-dimensional trajectories and network analyses of group behaviour within chimney swift flocks during approaches to the roost. *Proceedings of the Royal Society B: Biological Sciences, 284*(1849), 20162602. https://doi.org/10.1098/rspb.2016.2602

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.

Gan, Y., Han, R., Yin, L., Feng, W., & Wang, S. (2021). Self-supervised multi-view multi-human association and tracking. *Acm mm*.

Girshick, R. (2015). Fast r-cnn. Iccv.

Girshick, R., Donahue, J., Darrell, T., & Malik, J. (2014). Rich feature hierarchies for accurate object detection and semantic segmentation. *Cvpr*.

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*(8), 975–981. https://doi.org/10.1038/s41592-021-01226-z

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. https://doi.org/10.7554/eLife.48571

Han, X., You, Q., Wang, C., Zhang, Z., Chu, P., Hu, H., & Liu, Z. (2021). *MMPTRACK: Large-scale Densely Annotated Multi-camera Multiple People Tracking Benchmark*. Mmptrack: Large-scale densely annotated multi-camera multiple people tracking benchmark.

Hartley, R., & Zisserman, A. (2003). *Multiple view geometry in computer vision*. Cambridge: Cambridge University Press.

He, K., Gkioxari, G., Dollár, P., & Girshick, R. (2017). Mask r-cnn. Iccv.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (cvpr)* (pp. 770–778). https://doi.org/10.1109/CVPR.2016.90

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. https://doi.org/10.1371/journal.pcbi.1007354

Hou, J., He, Y., Yang, H., Connor, T., Gao, J., Wang, Y., et al. (2020). Identification of animal individuals using deep learning: A case study of giant panda. *Biological Conservation, 242*, 108414.

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). https://doi.org/10.1109/ICRA48506.2021.9561338

Karunasekera, H., Wang, H., & Zhang, H. (2019). Multiple object tracking with attention to appearance, structure, motion and size. *IEEE Access, 7*, 104423–104434.

Katz, Y., Tunstrøm, K., Ioannou, C. C., Huepe, C., & Couzin, I. D. (2011). Inferring the structure and dynamics of interactions in schooling fish. *Proceedings of the National Academy of Sciences, 108*(46), 18720–18725. https://doi.org/10.1073/pnas.1107583108

Kohn, G. M., King, A. P., Dohme, R., Meredith, G. R., & West, M. J. (2013). In the company of cowbirds, molothrus ater ater: Robust patterns of sociability predict reproductive performance. *Journal of Comparative Psychology, 127*, 40–48. https://doi.org/10.1037/a0029681

Krogius, M., Haggenmiller, A., Olson, E. (2019 October). Flexible layouts for fiducial tags. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*.

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*(4), 496–504.

Lauer, J., Zhou, M., Ye, S., Menegas, W., Schneider, S., Nath, T., et al. (2022). Multi-animal pose estimation, identification and tracking with deeplabcut. *Nature Methods, 19*(4), 496–504.

Lin, T.-Y., Goyal, P., Girshick, R., He, K., & Dollár, P. (2017). Focal loss for dense object detection. Iccv.

Ling, H., Mclvor, G. E., Nagy, G., MohaimenianPour, S., Vaughan, R. T., Thornton, A., & Ouellette, N. T. (2018). Simultaneous measurements of three-dimensional trajectories and wingbeat frequencies of birds in the field. *Journal of The Royal Society Interface, 15*(147), 20180653.

Liu, W., Anguelov, D., Erhan, D., Szegedy, C., Reed, S., Fu, C.-Y., & Berg, A.C. (2016). Ssd: Single shot multibox detector. Eccv.

Luo, H., Gu, Y., Liao, X., Lai, S., & Jiang, W. (2019). *June*. Cvpr workshops: Bag of tricks and a strong baseline for deep person re-identification.

Luo, H., Jiang, W., Gu, Y., Liu, F., Liao, X., Lai, S., & Gu, J. (2019). A strong baseline and batch normalization neck for deep person re-identification. *IEEE Transactions on Multimedia*. https://doi.org/10.1109/TMM.2019.2958756

Maguire, S. E., Schmidt, M. F., & White, D. J. (2013). Social brains in context: Lesions targeted to the song control system in female cowbirds affect their social network. *PLOS ONE, 8*(5), 1–8. https://doi.org/10.1371/journal.pone.0063239

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*(9), 1281–1289.

Mordant, N., Crawford, A., & Bodenschatz, E. (2004). Experimental lagrangian acceleration probability density function measurement. *Physica D: Nonlinear Phenomena, 193*(1), 245–251. https://doi.org/10.1016/j.physd.2004.01.041

Ouellette, N. T., Xu, H., & Bodenschatz, E. (2006). A quantitative study of three-dimensional lagrangian particle tracking algorithms. *Experiments in Fluids, 40*(2), 301–313.

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*(1), 117–125. https://doi.org/10.1038/s41592-018-0234-5

Pereira, T. D., Tabris, N., Matsliah, A., Turner, D. M., Li, J., Ravindranath, S., et al. (2022). Sleap: A deep learning system for multi-animal pose tracking. *Nature Methods, 19*(4), 486–495.

Pérez-Escudero, A., Vicente-Page, J., Hinz, R. C., Arganda, S., & de Polavieja, G. G. (2014). idtracker: Tracking individuals in a group by automatic identification of unmarked animals. *Nature Methods, 11*(7), 743–748. https://doi.org/10.1038/nmeth.2994

Qi, J., Gao, Y., Hu, Y., Wang, X., Liu, X., Bai, X. & Bai, S. (2021). Occluded video instance segmentation: A benchmark. arXiv preprint arXiv:2102.01558.

Quigley, M., Gerkey, B., Conley, K., Faust, J., Foote, T., Leibs, J. & Ng, A. (2009). Ros: an open-source robot operating system. In *Proc. of the IEEE Intl. Conf. on Robotics and Automation (icra) Workshop on Open Source Robotics*.

Rajasegaran, J., Pavlakos, G., Kanazawa, A., & Malik, J. (2021). Tracking people with 3d representations.

Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. *Cvpr*. https://doi.org/10.1109/CVPR.2016.91

Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in Neural Information Processing Systems, 28*.

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*(2), 179–182. https://doi.org/10.1038/s41592-018-0295-5

Schneider, S., Taylor, G., Linquist, S., & Kremer, S. (2018). Past, present, and future approaches using computer vision for animal re-identification from camera trap data. *Methods in Ecology and Evolution*. https://doi.org/10.1111/2041-210X.13133

Schofield, D., Nagrani, A., Zisserman, A., Hayashi, M., Matsuzawa, T., Biro, D., & Carvalho, S. (2019). Chimpanzee face recognition

from videos in the wild using deep learning. *Science Advances, 5*(9), eaaw0736.

Segalin, C., Williams, J., Karigo, T., Hui, M., Zelikowsky, M., Sun, J. J., & Kennedy, A. (2021). The mouse action recognition system (mars) software pipeline for automated analysis of social behaviors in mice. *eLife, 10*, e63720. https://doi.org/10.7554/eLife.63720

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *Sci: Comput. Therm*

Sinhuber, M., Van Der Vaart, K., Ni, R., Puckett, J. G., Kelley, D. H., & Ouellette, N. T. (2019). Three-dimensional time-resolved trajectories from laboratory insect swarms. *Scientific Data, 6*(1), 1–8.

Sun, P., Kretzschmar, H., Dotiwalla, X., Chouard, A., Patnaik, V., & Tsui, P. (2020). others. Waymo open dataset. Cvpr: Scalability in perception for autonomous driving.

Walter, T., & Couzin, I. (2021). Trex, a fast multi-animal tracking system with markerless identification, 2d posture estimation and visual field reconstruction. *eLife, 10*, e64000. https://doi.org/10.7554/eLife.64000

Wang, B., Wang, G., Luk Chan, K., & Wang, L. (2014). Tracklet association with online target-specific metric learning. *Cvpr*.

Wang, X., Kong, T., Shen, C., Jiang, Y., & Li, L. (2020). SOLO: Segmenting objects by locations. *Eccv*.

Wang, Y., Kolotouros, N., Daniilidis, K., & Badger, M. (2021). Birds of a feather: Capturing avian shape models from images. *Computer Vision and Pattern Recognition (cvpr)*.

Weinberger, K. Q., & Saul, L. K. (2009). Distance metric learning for large margin nearest neighbor classification. *Journal of Machine Learning Research, 10*(9), 207–244.

Weng, X., Wang, J., Held, D., & Kitani, K. (2020). 3D Multi-object tracking: A baseline and new evaluation metrics. *Iros*.

White, D. J. (2010). A social ethological perspective applied to care of and research on songbirds. *ILAR Journal, 51*(4), 387–393. https://doi.org/10.1093/ilar.51.4.387

White, D. J., Gersick, A. S., & Snyder-Mackler, N. (2012). Social networks and the development of social skills in cowbirds. *Philosophical Transactions of the Royal Society of London. Series B, Biological sciences, 367*(1597), 1892–1900. https://doi.org/10.1098/rstb.2011.0223

Wojke, N., Bewley, A., & Paulus, D. (2017). Simple online and realtime tracking with a deep association metric. *Icip*.

Wojke, N., Bewley, A., & Paulus, D. (2017). Simple online and realtime tracking with a deep association metric. In *2017 IEEE International Conference on Image Processing (icip)* (pp. 3645–3649). https://doi.org/10.1109/ICIP.2017.8296962

Wu, Z., & Betke, M. (2016). Global optimization for coupled detection and data association in multiple object tracking. *Computer Vision and Image Understanding, 143*, 25–37.

Wu, Z., Fuller, N., Theriault, D., & Betke, M. (2014). A thermal infrared video benchmark for visual analysis. *Cvpr workshops*.

Wu, Z., Hristov, N. I., Kunz, T. H., & Betke, M. (2009). Tracking-reconstruction or reconstruction-tracking? comparison of two multiple hypothesis tracking approaches to interpret 3d object motion from several camera views. *Workshop on Motion and Video Computing (wmvc)*. https://doi.org/10.1109/WMVC.2009.5399245

Xu, H. (2008). Tracking lagrangian trajectories in position-velocity space. *Measurement Science and Technology, 19*, 075105. https://doi.org/10.1088/0957-0233/19/7/075105

Yi, D., Lei, Z., Liao, S., & Li, S. Z. (2014). Deep metric learning for person re-identification. *Icpr*.

Yin, T., Zhou, X., & Krahenbuhl, P. (2021). Center-based 3d object detection and tracking. *Cvpr*.

Yu, H., Xu, Y., Zhang, J., Zhao, W., Guan, Z., & Tao, D. (2021). Ap-10k: A benchmark for animal pose estimation in the wild. arXiv preprint arXiv:2108.12617.

Zheng, L., Shen, L., Tian, L., Wang, S., Wang, J., & Tian, Q. (2015). Scalable person re-identification: A benchmark. In *2015 IEEE International Conference on Computer Vision (iccv)* (pp. 1116–1124). https://doi.org/10.1109/ICCV.2015.133

Zhou, X., Zhu, M., & Daniilidis, K. (2015). Multi-image matching via fast alternating minimization. *Iccv*.

Zivkovic, Z. (2004). Improved adaptive gaussian mixture model for background subtraction. *Icpr*. https://doi.org/10.1109/ICPR.2004.1333992

Zivkovic, Z., & van der Heijden, F. (2006). Efficient adaptive density estimation per image pixel for the task of background subtraction. *Pattern Recognition Letters, 27*(7), 773–780. https://doi.org/10.1016/j.patrec.2005.11.005

Zou, G., Fu, G., Peng, X., Liu, Y., Gao, M., & Liu, Z. (2021). Person re-identification based on metric learning: a survey. *Multimedia Tools and Applications, 80*(17), 26855–26888.

Zuffi, S., Kanazawa, A., Berger-Wolf, T., & Black, M.J. (2019). Three-d safari: Learning to estimate zebra pose, shape, and texture from images" in the wild". In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 5359–5368).