

## 8 Supplementary Materials

### 8.1 Full Affiliations

<sup>1</sup>Centre for the Advanced Study of Collective Behaviour, University of Konstanz, Konstanz, Germany,

<sup>2</sup>Department of Collective Behavior, Max Planck Institute of Animal Behavior, Konstanz, Germany,

<sup>3</sup>Department of Biology, University of Konstanz, Konstanz, Germany

<sup>4</sup>Department of Computer and Information Science, University of Konstanz, Germany,

<sup>5</sup>Luondua Boreal Field Station, Arvidsjaur, Sweden,

<sup>6</sup> School of Biological and Medical Sciences, Oxford Brookes University, OX3 0BP, Headington, UK,

<sup>7</sup>Department of Zoology, Stockholm University, Stockholm, Sweden

<sup>8</sup>Department of Ecology of Animal Societies, Max Planck Institute of Animal Behavior, Konstanz, Germany

### 8.2 Supplementary Methods

#### 8.2.1 Fieldwork and data collection

All data included in the CHIRP dataset was collected between 2019-2022 in a long-term field study of Siberian jays in northern Sweden, ongoing since 1989. The field site is split into two sites, a “managed” area which is a patch of managed forest with routine timber extraction, and “Reivo” area, within the Reivo nature reserve. Routine monitoring allows for acquiring detailed information on an individual’s attributes (e.g., age, sex), life history events (e.g., dispersal, group membership) and fitness (e.g., reproductive output and survival, as well as group information such as group size, composition, and social dynamics [18]). Due to the uniqueness of the species biology as a group living bird that cooperates in the contexts of skill acquisition and anti-predator strategies, but not reproduction (they are not cooperative breeders), the study system provides a unique opportunity to study the evolution and maintenance of group-living and cooperation [29], amidst the backdrop of a rapidly changing environment [35].

Each year, researchers visit every territory within the field site to observe, ring and sample every Siberian jay, including conducting 1-2 standardized behavioural observations by placing a two-level wooden perch, with two pieces of pig fat tied in the center of each level. The design of the feeding perch allows for colour rings to be easily identified by researchers with binoculars, and on video when birds are perched. Once birds arrive, researchers start a behavioural recording using a HD resolution (1920 x 1080) handy camcorder on a tripod between 5-10m away, by ensuring the feeding stick was centred and covered the entire screen, and that nothing blocks the view between the camera and the birds. During the observation, the researcher narrates ev-

ery time an individual is doing certain behaviours, including feeding, displacement, chasing and submissive behaviours. The recording was stopped once 30 minutes were up, or 15 min if the group only has 2 individuals. All experiments and observations were approved by Umea ethics board, A23-20. Bird ringing was done under the licence of the Swedish Museum of Natural History.

Once all data were collected, a large number of videos (around 430) were coded manually using BORIS software. For each video, presence, eating, displacement and submissive behaviours were coded along with the identities of the birds performing each behaviours (Table S.2). These coded data are used for further analysis of social behaviour and dynamics in the Siberian jay (e.g. [25, 27, 41]), but act as an important foundation for the CHIRP dataset.

#### 8.2.2 CHIRP Dataset preparation

Firstly, we prepared the bounding box, segmentation and ring segmentation datasets. We first manually selected 17 videos with diverse backgrounds, weather conditions and lighting. From these videos, we equally sampled 70 frames from each video, which acted as the base dataset for bounding box and segmentation mask annotations. We then annotated bounding boxes for the fat and all bird instances. In addition, we annotated segmentation masks for the wooden perch and birds, since segmentation masks gives finer scale shape information about the objects compared to bounding boxes. To speed up annotation, we used the labelled bounding boxes as prompt to SAM2 [51], to automatically generate masks for birds. We validated the approach by annotating 688 bird instances with segmentation masks, and computed the mean IOU. We confirmed that the SAM 2 approach creates accurate masks, at 0.84 mean IOU (Figure S.2).

For a subset of the randomly sampled frames, we labelled segmentation masks of every visible colour ring on cropped images of birds. We ensured that every ring colour present in the population is represented, with chrome and lime colour having the least sample because of relatively low frequency in the population (Figure 2B). For another subset of the frames, we also labelled 13 unique keypoints (Figure S.1), by labelling keypoints only when they are visible from the image. A portion of 2D keypoint annotations were also done in another subset of images, taken from a 4-camera setup of foraging jays on the ground.

Next, using the bounding box annotations described above, we trained a YOLOv8 model to detect Siberian jays from videos. Using the trained model, we ran inference on all videos to obtain bounding boxes from the YOLO model and trajectories using BoTSORT. These bounding boxes were then used to automatically select 3s segments with behaviours of interest, including eating and submissive behaviours, by matching video timestamps to BORIS

annotations. We then manually reviewed all extracted clips and scored the behaviour that was present in each clip, and labelling “other” when the bird was neither eating nor doing submissive calls.

To prepare the video re-id dataset, we also used bounding box and trajectory information by matching tracks to BORIS annotations, and identifying video segments where only 1 bird is present in the feeder. These segments were extracted and cropped as 1 second clips with the associated ID from BORIS annotations. These 1 second clips were then manually reviewed to confirm the ID assignment was correct, and rejecting tracks if the tracks contain objects that were not fays, when the ID was incorrect, or when ID cannot be visibly confirmed from the rings (e.g if all rings were occluded). These short clips then contributed to the video re-id dataset. We note here that for a small subset of videos, there are “unringed” individuals present, and those were removed from the final dataset.

Finally, we chose 4 independent videos that were not part of the core CHIRP dataset as the application specific benchmark. Similar to the above, we first ran a trained YOLOv8 model with tracking using BoTSORT, to output unique tracks from each video. Each track was then manually assigned to a colour bird ID, with unique track numbers every time a bird leaves the frame and returns. Since each ID track (colour ID + index) can include multiple tracks from BoTSORT, these were all combined, with missing frames filled with linear interpolation. Next, the same videos were coded using BORIS, where every time a bird individual touches the food (fat) with its beak, a state event was marked, together with the ID of the bird. This was then combined with the bounding box data, by adding a 3-frame window before and after the manually annotated frame, as an arbitrary time window for each peck.

### 8.2.3 CORVID: COlour-based Video re-IDentification

Placing colour rings on birds is a common way for individual marking of large bird populations, as it provides an effective way for individual identification either in the field through binoculars or from video data, when birds are standing upright. However, while work on re-identification of birds have been done [20, 64], to the best of our knowledge, no work have directly detected and used colour rings for individual identification. Unlike other animals like zebras or giraffes that has visually distinguishable features between individuals, birds generally have very subtle individual differences in visual appearance, making a re-id approach based on colour rings much more flexible and generalizable to other bird systems.

Here, we first extensively explored the ring segmentation dataset in CHIRP using classic colour extraction and unsupervised techniques. We explored different hyperparameters and image features that was most representa-

tive of the colour classes, including colour space, type of features and dimension reduction methods. To do that, we first created cropped rings using the provided segmentation masks, then resized and transformed the images to 20x20 pixels, with simple perspective transform to ensure the ring covers as much of the square as possible. After that, we tested different colour space transforms, including RGB, LAB, AB (LAB space without lightness value L) and HSV. We also tested different type of feature engineering methods, including taking the minimum, maximum, mean and standard deviation from each colour dimension (feature length of 12), and colour histograms, by tallying value of 10 equally spaced bins for each colour dimension (feature length of 30). After extracting features, we tested both UMAP, t-SNE, and no dimension reduction, and finally grouped clusters using kmeans clustering with differing number of target clusters. We refer to Table S.3 for all parameters explored. After obtaining clusters, we assigned all rings within each cluster with a colour class, based on the most frequent class in the cluster (from the ground truth), and computed colour assignment accuracy in the test set. After testing all combinations, we found that hsv colour space, with colour histograms, and dimension reduction to 3 dimensions using UMAP yielded the highest accuracy of 62% in the test set (Figure S.3A).

After identifying the feature engineering method that most represents the colour ring dataset, we used the same colour transforms (hsv space, colour histograms) to preprocess the ring images, and trained a random forest model, with random parameter search. Instead of training a classifier, we trained a regressor to predict confidences for each class, by converting ground truth classes into one-hot encoding. The final random forest model has a top 1 accuracy of 65%, and top 3 accuracy of 87% (Figure S.3B). This random forest model was then incorporated into the COBRA pipeline (Figure 4).

The proposed CORVID pipeline is split into a few distinct steps (Figure 4). Firstly, we trained a mask2former instance segmentation model [11] to detect rings from images of birds. To train the instance segmentation model, we removed class information and combined all colour classes into a single ring class. During inference time, the detected rings were first assigned to ring pairs if the centroids were within 30px Euclidean distance, defined arbitrarily by visualizing the histogram of distances (Figure S.5). After obtaining ring pairs, we used the random forest model above to predict a vector of confidences for each colour class (length 12) for each ring segmentation, which was then summed as a 12x12 matrix, representing the probability of colour pairs. Since each instance in the re-id dataset had 25 frames, we did the same for each frame and summed all the matrices, divided by the number of images to standardize values. Finally, we took the metadata of possible birds for the data

sample, and computed a score for each bird simply as the sum of the ring pair scores for the given ring combinations it has (e.g. oa or will have a score computed as the sum of the probability that the detected ring pairs were oa and or). The bird with the highest score was the predicted bird using the framework.

#### 8.2.4 Benchmarking

Finally, we benchmarked the three main tasks in CHIRP with state of the art methods and baselines. In this section, we describe detailed training and evaluation procedures.

For 2D keypoint estimation, we trained ResNet52, ResNet101, ResNet152, HRNet, ViTPose\_small, ViTPose\_large models using the MMPose library. All models were trained with default augmentation parameters, for 100 epochs. PCK of the test set was evaluated per epoch, and the epoch with the best test set PCK was chosen for the final results, to avoid over-fitting. All models were evaluated using standardized metrics, including Root mean squared error (RMSE), mean and median of the Euclidean distance between ground truth and detected points, as well as percentage correct keypoints (PCK05, PCK10), defined as proportion of keypoints that lies 5% or 10% of ground truth keypoints, scaled by largest bounding box dimensions.

For behavioural classification, we trained slowfast, I3D and C3D models using the MMAction2 library. All models were trained with default augmentation parameters and standard cross entropy loss function, for 100 epochs, with best model chosen as the epoch with highest test accuracy. For evaluation, we reported weighted average precision, recall, f1-score and accuracy across all 3 classes for each model.

For video re-id, we compared our proposed CORVID pipeline with Mega Descriptor [69], a recently proposed foundation model for animal re-id. For both closed set and disjointed set, we trained MegaDescriptor-L-384 model using ArcFace loss. Since consecutive frames within a video are very similar, we only used the first frame from each video as training. During training, we computed the loss on the test set to avoid overfitting, and selected the best model based on lowest test loss. For inference, we extracted image features of all frames within all videos in the re-id dataset, and computed top-5 nearest neighbours based on cosine similarity of the feature space between each image pair in the database. To leverage the extra information from having 25 frames per instance, we extracted top 5 nearest neighbours for all 25 frames per instance and tallied the matched birds.

For evaluation, we compared 3 approaches, first is our proposed CORVID pipeline, second is a pre-trained MegaDescriptor-L-384 model pre-trained on a wide range of animal re-id datasets [69], and thirdly the fine-tuned version of the same MegaDescriptor model. During evaluation,

we further compared two types of performance, first with the constraint of only matching to a database with birds that are possible in the given video (named “within territory”), and second with all birds in the population included as possible birds (named “all”). We report top-1 and top-3 accuracies across all data splits and categories.

For tracking, we used 4 distinct trackers, 1) BoTSORT, 2) OC-SORT, 3) StrongSORT and 4) simple centroid-based tracker. The first three trackers were implemented using the boxmot library [6], with default hyper parameters and osnet\_x1.0 model pretrained on the Market1501 dataset [68] for re-id. The simple centroid-based tracker was implemented by computing bounding box centroid for each frame, then assigning bounding boxes as the same track based on a simple Hungarian algorithm using Euclidian distances between the bounding box centroids. All MOT benchmark results can be found in Table S.4.

#### 8.2.5 Application specific benchmark pipeline

To provide a baseline for the application specific benchmark, we designed a simple 4 step pipeline that combines methods presented in individual task benchmarks, to obtain individual level feeding rates. The pipeline has a modular design, so that each individual component has standardized input and output formats, allowing new methods to be slotted into any part of the pipeline. The pipeline is as follows: 1) object detection, where bounding box of each bird was detected. 2) tracking, where bounding box detections were tracked across spatial and temporal scales to create tracklets. 3) Individual recognition: assigning an identity to each individual tracklet from step 2, and finally 4) behavioural recognition, where eating behaviours are detected within each track.

For the baseline provided we experimented with the individual recognition component of the pipeline, and keeping the other components consistent. We first used a YOLOv8 model for object detection, then BoTSORT for tracking. For individual identification, we used our proposed CORVID pipeline, fine-tuned MegaDescriptor model in the disjointed set, and random assignment to assign IDs to tracks. We chose the MegaDescriptor model fine-tuned on the disjointed set because the model retrieved the highest accuracy between all MegaDescriptor evaluations (Table 1). Since the application specific benchmark is an independent test set, we used the whole video re-id dataset as the gallery for MegaDescriptor. For all methods, we only provided the birds present in the video as possible birds, enforcing the “within-territory” constraint. The rest of the pipeline is identical to the benchmarking procedure for video re-id, other than more frames being pooled for each individual track (instead of standardized 25 frames per track provided in the main dataset). Finally, for behavioural recognition, we used the best performing C3D model from our

action recognition evaluation (Table 1), and predicted behaviours by splitting each track into 1 second time windows (25 frames), and predicting behaviour for each window.

To evaluate the proposed pipelines in the application specific benchmark, we proposed a few novel metrics. We provide a standardized script to compute all metrics, for future work to reproduce the benchmarks. Firstly, we calculate the proportion of ground truth frames with correctly identified individuals. This involves matching ground truth and predicted bounding boxes based on proportion overlap, and computing proportion correct frames from the number of frames that has correctly matched identities. Next, we calculated precision, recall and F1-score for individual level feeding rates. We split the whole video into 1s time windows, and considered a prediction as a true positive if both a ground truth eating event and a detected instance occur within the same window for a given individual.

Next, we computed two biological relevant measures. The first is individual-level feeding rate, defined as the total amount of times an individual pecked at the food per minute. We computed this measure for each individual, by simply tallying the total amount of detected eating instances and dividing it by each video's length. The second metric is a co-occurrence rate, defined as the total amount of frames two individuals were detected together, divided by video duration, as a simple measure for social affiliation. For both biological measures, we computed the absolute measurement error and presented the mean, median and standard deviation of the absolute error, as well as a Pearson's correlation value between ground truth and predicted values.

Finally, to set a human benchmark for automated algorithms to reach, an independent human annotator manually annotated every eating instance and the time period each individual was present in the frame for the first 5-minute of each test video. We then extracted both biological measures, to act as a baseline for all automated methods to be compared with.

## 9 Supplementary Video

Attached supplementary video shows sample inference results that showcase all available annotations in the CHIRP dataset. Link to video <https://youtu.be/s1g0aXG1TM4>

Table S.1. Existing animal datasets that combines multiple computer vision tasks

Dataset	Taxa: Species	Re- Identification	Action Recognition	Posture Estimation	Object Detection	Instance Segmentation	Multi-Object Tracking (MOT)	Captive or Wild
Animal Kingdom[49]	Mammals, birds, fishes, amphibians, reptiles, insects	–	✓	✓	✓	–	–	Wild
3D-POP[47]	Bird: Pigeon	✓	–	✓	✓	–	✓	Captive
ChimpACT[38]	Mammal: Chim-panzee	✓	✓	✓	✓	–	✓	Semi-Captive
Baboonland[17]	Mammal: Baboon	–	✓	–	✓	–	✓	Wild
BuckTales[48]	Mammal: Black-buck	✓	–	–	✓	–	✓	Wild
LoTE-animal[37]	Mammals	–	✓	✓	✓	✓	–	Wild
WILD Dataset[64]	Bird: Cowbird	✓	–	–	✓	✓	✓	Captive
<b>CHIRP</b>	<b>Bird: Siberian jay</b>	✓	✓	✓	✓	✓	✓	<b>Wild</b>

Table S.2. Ethogram for BORIS annotations on 15-30min behavioural videos.. We note that this ethogram is different from the behaviours offered in the dataset.

Behavior	Description	Modifiers / Notes
Presence	Individual present in the area.	Individual generally present (seen) in observation area around observation time. If the observer notes the bird has not been seen since a certain time, stop presence after the last event involving that bird.
Feeding	Individual is feeding on the feeder.	On the feeding stick or feeding from below. Press key when bird is in frame and wings are still (or already) open.
Within 1m	Bird is visible within 1m of feeder, but not feeding.	Individual visible in the camera frame (even partially), but not on feeder. Note this in spreadsheet.
SubmissiveSE	Continuous submissive wing fluttering or squeaking.	“A”: Submissive bird. “R”: Other subject/unknown. Must be in frame. Continuous behavior (no breaks).
SubmissivePE	Short submissive wing fluttering or squeaking, separated by pauses.	“A”: Submissive bird. “R”: Other subject/unknown. Must be in frame. Punctuated events rather than continuous.
Displacement	Actor displaces another bird from a branch or feeder perch.	Actor pushes/pecks/lands to force receiver to move. Must happen in frame, on feeder-side branches.
Chasing	Actor actively pursues another bird.	Actor leaves position to chase receiver. Energy-costly. Not always visible in frame; rely on observer comments.
Eating	Individual is actually eating.	Confirm visible ingestion.

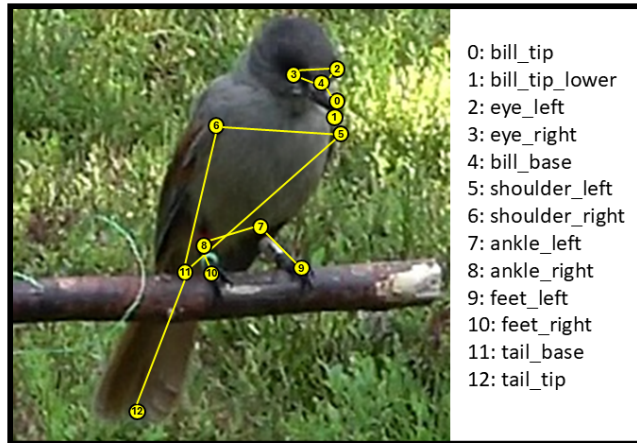
Figure S.1. **Keypoint definitions.** Annotated keypoints provided in the keypoint annotation dataset, ordered by the index as provided in the dataset.



Table S.3. Features and hyper parameter explored on the ring segmentation dataset

Category	Type	Hyper-parameters
Colour space	HSV	
	LAB	
	RGB	
	AB (LAB space without L)	
Feature type	Min, max, mean, sd	
	Colour histogram (10 bins per dimension)	
Dimension reduction method	UMAP	Number of dimensions: 2, 3
	t-SNE	
	None	
Clustering method	k-means	Number of clusters: 10, 15, 20, 25, 30

Table S.4. **MOT Metrics on tracking algorithms.** HOTA: Higher order tracking accuracy, MOTA: Multi-object tracking accuracy. MT/PT/ML: Mostly/Partially/Mostly Lost tracks percentage. Bold denotes best performing model for each metric.

Algorithm	HOTA	MOTA	MT (%)	PT (%)	ML (%)	Frag.	ID Switches
BoTSORT	<b>61.26</b>	<b>91.64</b>	89.50	7.87	2.62	2644	<b>1348</b>
OC-SORT	57.31	90.46	88.63	8.45	2.92	3201	1890
StrongSORT	54.49	90.48	88.92	8.16	2.92	2801	3932
Simple	37.58	90.94	<b>90.38</b>	<b>7.29</b>	<b>2.33</b>	<b>2326</b>	3765

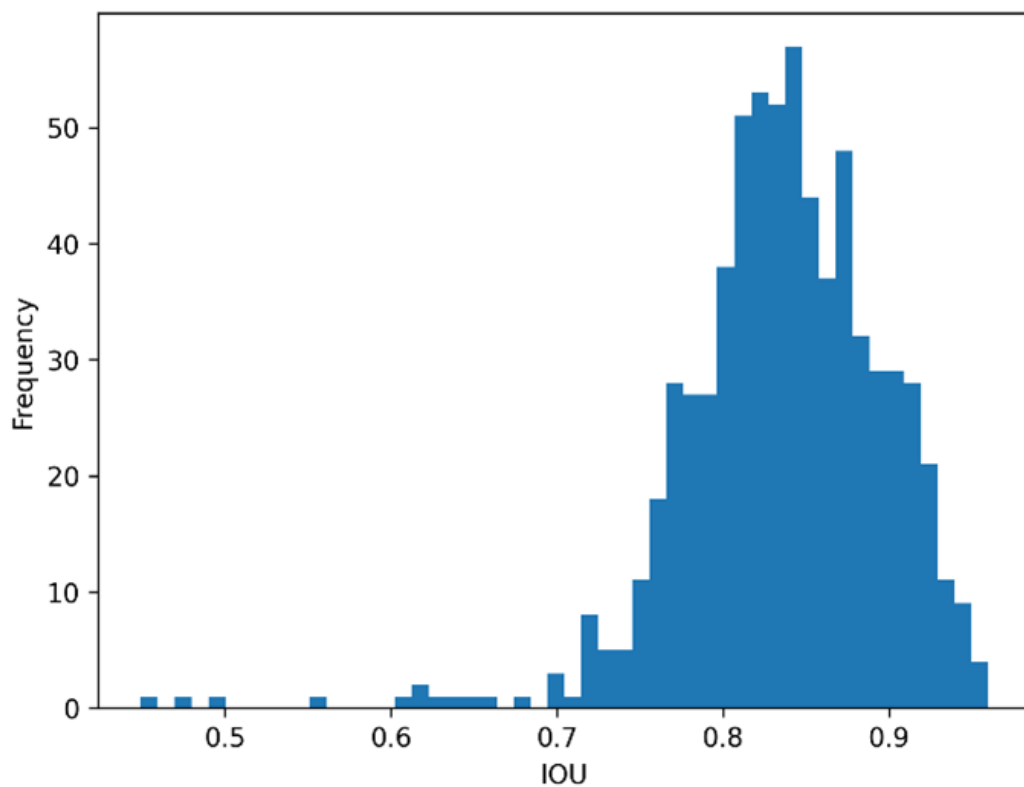
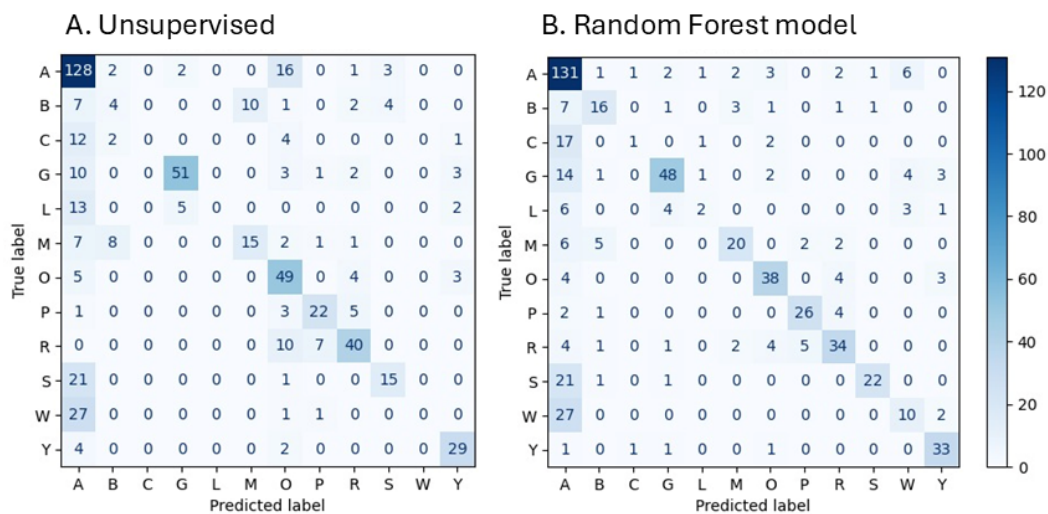


Figure S.2. IOU distribution between manually annotated segmentation masks versus masks automatically generated by SAM2, with a mean IOU of 0.84.



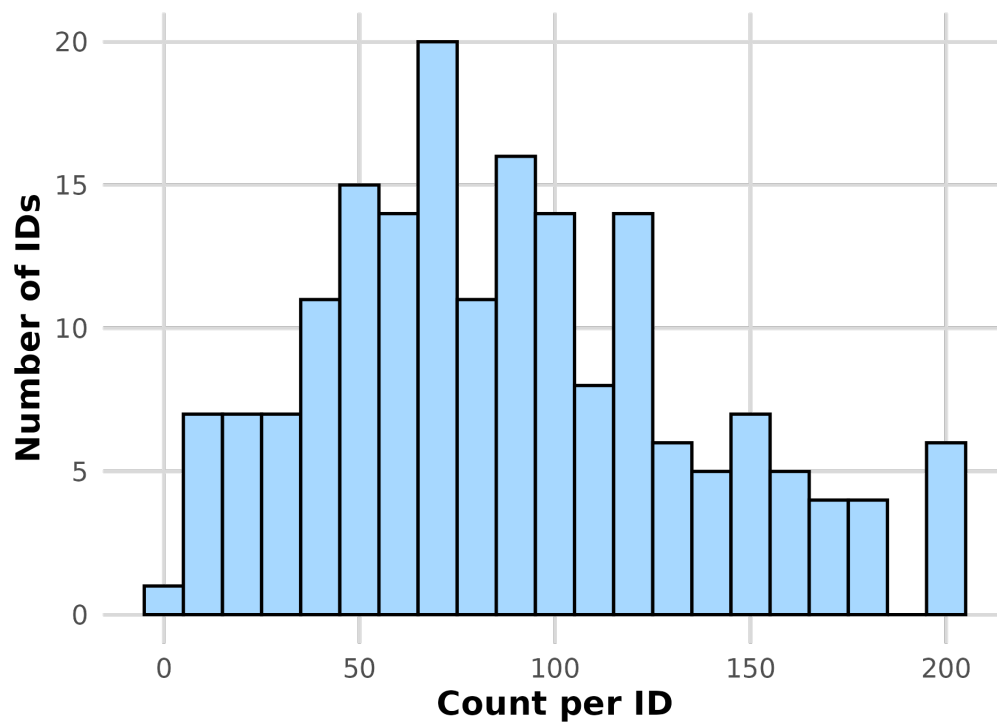


Figure S.4. Distribution of number of samples per individual in the video re-id dataset. Each sample consists of 25 frames.

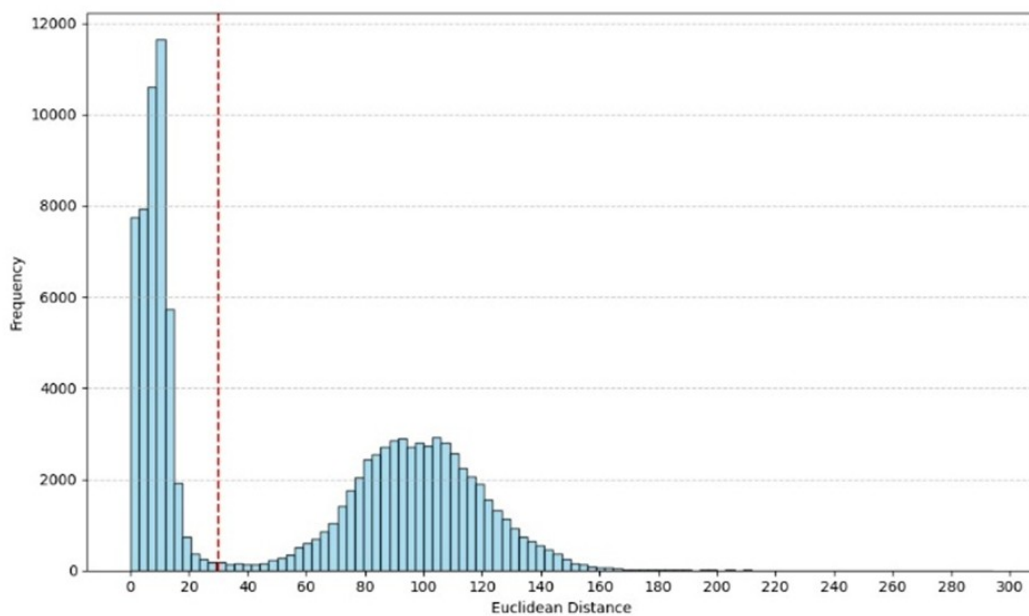


Figure S.5. Distribution of Euclidean distance between the centroid of pair of rings in a subset of the CHIRP video re-id dataset. Red dotted line refers to the 30px threshold we used to define whether two detected rings are considered as a ring pair.