Watching Swarm Dynamics from Above: A Framework for Advanced Object Tracking in Drone Videos

Duc Pham, Matthew Hansen, Félicie Dhellemmens, Jens Krause, Pia Bideau

Watching Swarm Dynamics from Above: A Framework for Advanced Object Tracking in Drone Videos

Duc Pham[1] , Matthew Hansen[3] , Félicie Dhellemmens[2], Jens Krause[3,4], Pia Bideau[2,4*]
[1]Technical University Berlin, [2]Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LJK,
[3]Leibniz Institute for Freshwater Ecology and Inland Fisheries, Berlin
[4]Science of Intelligence, Research Cluster of Excellence, Berlin

Abstract

Easily accessible sensors, like drones with diverse onboard sensors, have greatly expanded studying animal behavior in natural environments. Yet, analyzing vast, unlabeled video data, often spanning hours, remains a challenge for machine learning, especially in computer vision. Existing approaches often analyze only a few frames. Our focus is on long-term animal behavior analysis. To address this challenge, we utilize classical probabilistic methods for state estimation, such as particle filtering. By incorporating recent advancements in semantic object segmentation, we enable continuous tracking of rapidly evolving object formations, even in scenarios with limited data availability. Particle filters offer a provably optimal algorithmic structure for recursively adding new incoming information. We propose a novel approach for tracking schools of fish in the open ocean from drone videos. Our framework not only performs classical object tracking in 2D, instead it tracks the position and spatial expansion of the fish school in world coordinates by fusing video data and the drone's on board sensor information (GPS and IMU). The presented framework for the first time allows researchers to study collective behavior of fish schools in its natural social and environmental context in a non-invasive and scalable way.

1. Introduction

Schools of fish, flocks or birds or herds of sheep - in nature, collectives exhibit remarkable behaviors as they seemingly synchronize their movements within the group. Flexibly the collective changes their formation, density, speed and may even change their formation completely by smoothly dividing or merging [1]. These phenomena lead to marvellous visual spectacles that we can observe in the sky or in water. Both - aerial and marine environments, allow researchers to study collective behavior and the interaction between multiple collectives in natural environments that are minimally influenced by constraints imposed by the terrestrial structure. Therefore these ecological environments present an unique opportunity to study behavior of animal groups consisting of several thousand individuals [12]. The complexity of such group-motions go far beyond patterns that can be described with traditional representations like skeletons, key-points or bounding boxes mimicking the spatial configuration of an object [18]. Latest research mostly focused on tracking behavior of one or more individuals [10, 13, 21–23, 25, 27, 28, 30]. In this work, we tackle the problem of tracking collective behavior. The analysis of collective behavior shifts the perspective from previously specific observations of individual behavior to a global view encompassing group dynamics involving multiple individuals, often of the same species. Our specific goal is to track a school of fish as a cohesive group, capturing key features like position, spatial extent, and shape of the entire collective. We propose a method that leverages classical probabilistic state estimation via particle filtering by integrating

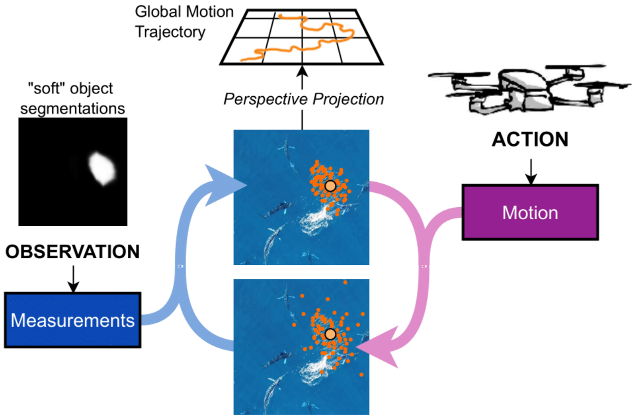


Figure 1. **Swarm Dynamics from Above (SwDA),** a framework for tracking collective behavior from drone videos. The recursive architecture of a particle filter, coupled with frame-by-frame semantic segmentation allows tracking over long time horizons.

[*]Corresponding Author: pia.bideau@inria.fr

Method	Object (Animal) Tracker	Animal Group Representation	Moving Camera	2D Trajectory	3D Trajectory	Temporal Scope	Tracking Environmnent
DeepLabCut [23]	Point/Skeleton	multiple individuals	✓	✓	✗	long-term	terrestrial/marine
Particle Video [13]	Point	-	✓	✓	✗	short-term	terrestrial
Follow Anything [22]	Point/Shape	-	✓	✓	✗	long-term	terrestrial/marine
Quantifying Behavior with a Drone [21]	Point/Bounding Box	multiple individuals	✓	✓	✓	long-term	terrestrial
Swarm Dynamics from Above (SwDA)	Point/Shape	swarm	✓	✓	✓	long-term	marine

Table 1. **Literature review** on approaches for tracking animal behavior captured by a moving camera. Methods deviate in their object representation (point/skeleton/bounding box), non of existing methods directly track a pixel accurate object shape over long time horizons. SwAD is the only method designed for marine environments.

advances in semantic object segmentation. The particle filter allows for recursively adding new incoming information over long time horizons. We continuously track the fish school while integrating learned segmentation masks and the drone's movement (GPS and IMU). Our contributions can be summarized as follows:

i. For the first time, an approach is introduced that enables the analysis of animal behavior in real-world coordinates, even in demanding environments lacking geographical structures or unique landmarks.
ii. The proposed methodology combines learning-based approaches for semantic segmentation with recursive Bayesian filtering, resulting in high performance even in low-data regimes.
iii. Its temporal robustness is demonstrated in the tasks of trajectory and shape estimation. Both of which are evaluated over extended time periods (~5 minutes).
iv. Code and the new dataset for tracking collective animal behavior will be released a long an extended version of this workshop paper.

2. Related Work

Algorithms have been developed for tracking animal behavior in well-controlled laboratory experiments using stationary cameras [5, 23], but also for handling challenging field experiments [19, 21, 22, 26]. While the ultimate goal is to analyze movements of single animals or groups in the wild, the natural environment significantly determines the capability of extracting accurate information about the animal's movement. Table 1 provides an exemplary overview of the diversity of existing techniques and their applications. Techniques range from high-precision tracking in structured environments to more adaptable methods suited for the complex and variable conditions like marine environments. In terrestrial environments well-identifiable landmarks help to reconstruct the 3D trajectory from video sources [11, 21]. The ability to extract 3D behavioral trajectories significantly degrades with the opportunity of identifying high-quality landmarks. Estimating 3D trajectories in marine environments has not been widely addressed [22, 24]. The primary reasons are the lack of geometric structure needed to reconstruct 3D movement trajectories, making it a challenging vision and robotics problem. Additionally, the accessibility of animals found far off the coast makes the use of technical communication devices that rely on a fixed base (receiver) impractical. This work for the first time processes visual data (videos) and GPS/IMU information recorded by a drone to extract real world trajectories of a fish school in the Pacific Ocean 10-30 km offshore Baja California Sur.

3. Visual Tracking of Swarm Dynamics

We present a novel framework for tracking swarm dynamics from drone recordings. In Section 3.1 we introduce the recursive structure of the particle filter that unifies accurate frame-wise object detections and the drone's sensor data. Building upon the algorithmic structure of the particle filter, Section 3.2 introduces *Swarm Dynamics from Above*.

3.1. Background: Particle Filter

We examine the task of inferring a hidden state s from a sequence of observations o and performed actions a, i.e. the probability distribution of $p(s_t|o_{0:t}, a_{0:t}) = \text{bel}(s_t)$. In other words, we wish to localize the object in world coordinates from frame-wise object detections and the drone's pose provided by its GPS and IMU data. Following the principle of Bayesian filtering, a belief over the object's location evolves over time through two sequential steps that iterativly track the object's location: 1) prediction using action a_t and 2) update using observation o_t:

$$\overline{\text{bel}}(s_t) = \int p(s_t|s_{t-1}, a_t)\text{bel}(s_{t-1})\, ds_{t-1} \tag{1}$$

$$\text{bel}(s_t) = \eta p(o_t|s_t)\overline{\text{bel}}(s_t) \tag{2}$$

The Bayes filter computes $\text{bel}(s_t)$ recursively from $\text{bel}(s_{t-1})$ while incorporating the new information contained in a_t and o_t. *How to represent this believe?* Particle filters are a way to efficiently represent an arbitrary (non-Gaussian) distribution. In case of a particle filter a set of particles $s_t^{[0]}, ..., s_t^{[N]}$ and weights $w_t^{[0]}, ..., w_t^{[N]}$ serve as an approximation to the probability distribution to be estimated. More specifically particles are iteratively moved, weighted and resampled. The particle filter implements the prediction step by moving each particle stochastically, which is achieved by sampling from a generative motion model, $s_t^{[i]} \sim p(s_t^{[i]}|a_t, s_{t-1}^{[i]})$. During the measurement update the weight of each particle i is set to the observation likelihood, $w_t^{[i]} = p(o_t|s_t^{[i]})$ and particles are resampled accordingly. Following the underlying recursive structure of the particle filter we introduce *Swarm Dynamics from Above*, a model that allows robust tracking of animal behavior in real world-coordinates, even in demanding environments that lack geometric structures or unique landmarks. In these highly demanding environments any classical tracking algorithms that solely rely on motion estimates through optical flow are typically prone to errors [7, 15].

3.2. Tracking of Swarm Dynamics in Drone Videos

The drone is equipped with a high resolution camera, gimbal for image stabilization and on board sensors - IMU and GPS, that provide the absolute drone pose (in geographic coordinates), its translational velocity measured in m/s and its rotation (pitch, yaw and roll) in degree. Measurements are converted into cartesian coordinates with its origin being at the position of the drone projected onto the ground. The camera's pose is determined from provided sensor measurements using a standard Kalman filtering approach. Given an accurate estimate of the drone's pose the following paragraphs outline the particle tracking framework with its motion model and measurement model that keep track over the objects position on the image plane over long time horizons. At time $t = 0$ the particle filter is initialized with a uniformly distributed set of P particles. Section 3.2.1 summarizes the conversion of the 2D trajectory into a global motion trajectory. It's worth noting that 2D tracking encompasses information about the drone's movement.

Motion Model. The complex open ocean environment, lacking clear geometrical structures and often featuring sun reflections due to wind or animal movements near the water surface, hinders direct motion estimation from optical flow. Optical flow relies on the brightness consistency assumption [16], which doesn't hold aforementioned cases. Instead of estimating particle movement from video frames, we infer the particle movement induced by the camera's motion. Let the camera rotation be defined by $[A, B, C]$ and its translation by [U, V, W], following rules of perspective projection the flow can be geometrically determined via:

$$\vec{v} = \frac{1}{z}\begin{bmatrix}-fU + xW\\ -fV + yW\end{bmatrix} + \begin{bmatrix}\frac{A}{f}xy - Bf - \frac{B}{f}x^2 + Cy\\ Af + \frac{A}{f}y^2 - \frac{B}{f}xy - Cx\end{bmatrix}, \quad (3)$$

here f is the camera's focal length in pixels and z the camera's height (distance to the captured scene) [15]. Each particle is displaced by $\vec{v}$, and Gaussian noise is introduced by resampling from a Gaussian distribution with the mean equal to the updated particle position (Fig. 2).

Measurement Model. Incoming new observations o_t continuously update the belief over the object's pose using Bayes' rule. We estimate *soft* segmentation masks o_t by first training DeeplabV3 [4] with a MobileNet backbone architecture [17] on the new dataset for swarm tracking. For training we minimize the binary cross-entropy loss, which is equivalent to maximizing the likelihood of the data. This process yields soft segmentation masks obtained during network inference. Particles are weighted according to estimated segmentation masks and resampled accordingly. We utilize roulette wheel sampling for particle resampling, a technique that incorporates elements from evolutionary computation methodologies [9].

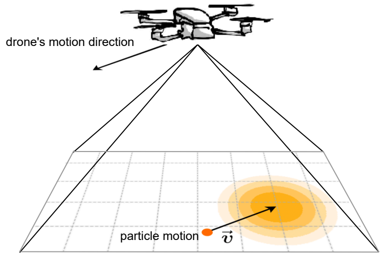


Figure 2. **Illustration of the motion model.** Each particle is displaced by the induced motion vector due to the drone's movement

Starting from initially uniformly distributed particles, particles quickly adapt to the swarm's shape and are capable to capture its dynamics - displacement and shape deformation. To achieve this goal, we introduced a model that iteratively conducts *prediction* by considering the drone's actions and then *updates* the prediction using learned object segmentation masks (Fig. 1). This improves tracking accuracy, particularly in low data regimes, a prevalent challenge frequently encountered in animal behavior research.

3.2.1 Estimating the Global Movement Trajectories

Global movement trajectories are computed following the physical principles of image formation. Given the camera's pose, the camera's extrinsic parameters R and $\boldsymbol{t}$ determining the 3D camera motion can be determined. Intrinsic camera parameters K such as focal length and principle point offset are known. Therefore the global movement trajectory can be obtained as follows: $\boldsymbol{p} = K[R|\boldsymbol{t}]\boldsymbol{P}$, where $\boldsymbol{P}$ denotes the world point and $\boldsymbol{p}$ denotes its projection onto the image plane [14]. We define the origin of the global coordinate system to be at the initial drone pose projected onto the ground. Furthermore it is assumed that the camera's height is equal to the distance to the object being tracked, implying that the swarm is situated at the water surface.

4. Experiments

We present novel data for tracking the dynamics of a school of fish being pursued by predators. The strong interaction between predator and prey results in highly expressive and unique swarm dynamics. We include an overview of the experimental setup, evaluation metrics covering track-

Num Training Samples	400			16		
Method	$S = 30$	$S = 20$	$S = 10$	$S = 30$	$S = 20$	$S = 10$
Follow Anything [22]	37.96	32.76	20.33	38.37	33.10	20.42
DeepLabV3	69.31	63.65	45.06	31.63	16.81	6.61
SwDA	**84.40**	**77.93**	**50.31**	**76.66**	**69.78**	**40.22**

Table 2. **Tracking Accuracy.** We compare the quality of estimated 2D trajectories with *Follow Anything*, a tracker that similarly tracks an object from a moving drone and *DeepLabV3*. DeepLabV3 segments each frame in a sequence (such as a video) independently of the others, resulting in instantaneous segmentation masks. The mean point of each mask is tracked.

ing, shape segmentation accuracy, and the reconstruction of world coordinates from image detections.

Implementation Details. Data was recorded with the drone DJI Phantom 4. For Network training we utilized DeeplabV3 pretrained on ImageNet [6] and further train using the BCEWithLogitLoss and AdamW optimizer. The networks learning rate was set to 1e-3. The network was trained for 50 epochs and the model that reaches lowest error on validation set was chosen. During particle tracking the objects shape was approximated with 1000 particles.

Data. Recordings picture highly dynamic predator and prey interactions in the Pacific Ocean 10-30 km offshore Baja California Sur. Data is captured by a DJI Phantom 4 pro drone at 60fps. Videos show a schooling prey during group hunts of striped marlins. Prey fish schools typically consist of populations ranging from approximately 100 to over 3000 individuals. Each video shows a different school of fish during predator attack. This wide range not only affects the school's appearance in terms of its size but also has a significant impact on the school's dynamics and behavior, especially in response to predatory attacks. Examples are shown in Figure 4. Each video file is accompanied with accurately synchronized sensor measurements of the drones on board sensors. In total 40min of video data is provided - 8 videos of 5min each. The videos are split into 4-folds. Per fold 4 videos are used for training, 2 videos are used for validation and testing respectively. Per video 100 frames distributed over the full duration of each video are annotated with pixel accurate segmentation masks outlining the fish school's shape, resulting in a total of 800 segmentation masks - per k-fold 400 samples are used for training, 200 for validation and 200 for testing. The fish school's movement is tracked throughout the full video sequence with ground truth point annotations. Data will be made available for further analysis and research purposes.

4.1. Evaluation

This section evaluates and discusses multiple aspects of our tracking framework: estimation of movement trajectories, tracking of shape and localisation in world coordinates.

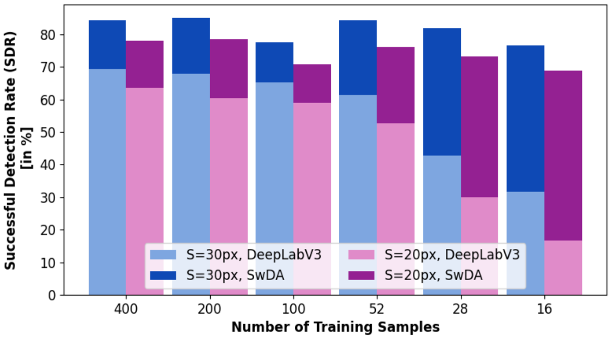
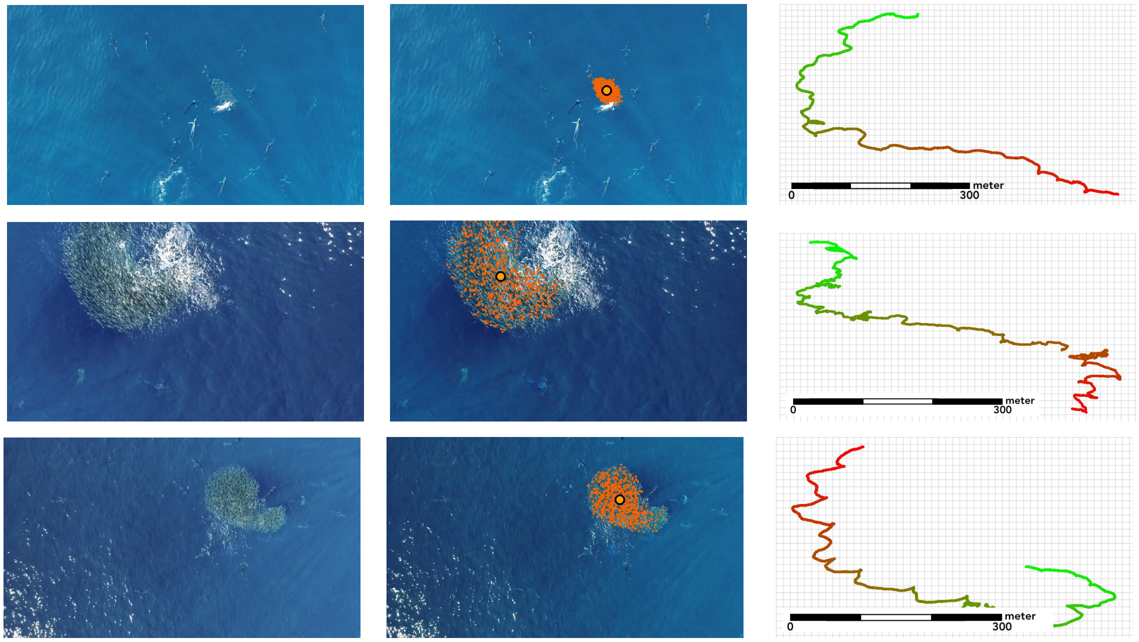


Figure 3. **Tracking Accuracy.** Evaluation of the Tracking Accuracy for different amount of labeled training data. Accuracy is measured via the successful detection rate (SDR). Results for two different precision ranges are shown: SDR within a radius of 30 pixels (blue), SDR within a radius of 20 pixels (purple).

Method	Intersection over Union	Precision	Recall	F1-measure
Follow Anything [22]	42.5	52.0	51.1	49.8
DeepLabV3	73.8	83.1	77.4	78.9
SwDA	71.4	76.2	85.4	78.9

Table 3. **Shape Segmentation Accuracy.** To compare the quality of the swarm's shape, we convert the set of tracked particles to segmentation masks. Reconstructing the shape of a 2D point cloud on the plane is inferred through its corresponding α-shape.

Movement Trajectories An often used measure to assess the 2D tracking accuracy is the average trajectory error [32]. A significant drawback of this measure however becomes apparent when the tracker loses the target, the output location can be random and the average error value may not measure the tracking performance correctly [2, 31]. Instead, a widely accepted measure is the successful detection rate (SDR), a measure that reports the percentage of accurate detections within a predefined precision range. We compare our approach SwDA with Follow Anything and DeepLabV3. Follow Anything leverages foundation models like CLIP [29], DINO [3], and SAM [20] to compute segmentation masks that best align with the queried objects, therefore its performance is independent upon the amount of training data as one can see in Table 2. The queried object is identified using ground truth segmentation masks. DeepLabV3 relies frame-wise estimates - not exploiting temporal consistency. We follow a similar protocol as introduced in DeepLabCut [23] to extract trajectories from frame-wise segmentation masks. Fig 3 shows the tracking accuracy averaged over all four training folds and for different amounts of training data. SwDA - our particle set tracker, shows significantly stronger tracking performance also in low-data regimes. While the accuracy of a pure learning based approach drops to 31.6% and 38.4% respectively (precision range of S=30px) when trained with only 16 training samples, the particle tracker only slightly



(a) Original frame (b) SwDA - 2D particle tracking (c) 3D motion trajectory (birds-eye view)

Figure 4. **Qualitative Results** of three different videos are shown: (a) Original frame, (b) Swarm detection via particle tracking and (c) Global movement trajectory of the swarm, 10m/grid cell, color visualises time. The overall video length is ~5min. Full videos are provided in the accompanying suppl. material.

reduces performance compared to training with all available training data. We improve upon recent advancements in semantic object segmentation by integrating particle filter-based tracking. This approach leverages physical regularities such as temporal consistency commonly found in behavioral animal data.

Tracking of Collective Formation Patterns. Shape segmentation accuracy is evaluated using four distinct accuracy metrics: intersection over union, precision, recall, and the F1-measure. In SwDA, soft segmentation masks from DeepLabV3 serve as frame-wise observations o_t, and particles are resampled accordingly. It is anticipated that this process will roughly maintain the segmentation quality expected from DeepLabV3. Given a set of particles, we approximate pixelwise segmentation masks by computing its corresponding α-shape [8]. This leads to a slightly more spatially expanded segmentation regions, which can be seen in the higher recall of SwDA (Tab. 3).

3D Tracking and Localization. Based solely on maritime drone recordings without unique landmarks it is impossible to achieve a good estimate of the algorithms ability to extract real-world coordinates from 2D trajectories on the image plane. Therefor we record aerial video data capturing a simple terrestrial environment, with clear and easy to identify markers. The ground truth 3D position of markers are measured using real-time kinematic positioning (RTK). Eight markers are located at different positions and recorded by a moving drone. While the detection of the target position is meant to be kept as simple as possible, the goal of this experiment is to evaluate the algorithms ability to retrieve accurate relative distances between detected makers. The accuracy of reconstructing a 3D position of a static landmark is reported in Tab. 4. The drone's sensor measurements are fused to estimate the drone's pose.

Method	absolute error [in m]	standard deviation
GPS	0.41	0.28
IMU	0.96	0.65
IMU+GPS (Kalman filter)	**0.32**	**0.21**

Table 4. **Tracking Accuracy in 3D.** We compare different localization approaches via GPS, IMU or their combination. SwDA integrates sensor measurements from IMU and GPS via Kalman filtering. The relative distances between markers is evaluated. Absolut distances of markers vary between 14m and 23m.

5. Conclusion

This work unifies learning and probabilistic modeling within a coherent algorithmic structure. By exploiting temporal consistency which can be found in behavioral data, our proposed method can handle challenging training scenarios where not much annotated training data is available. Furthermore since tracking relies on both visual appearance features and measurements of physical sensors the proposed algorithm is capable to tracking animal behavior in highly demanding environments, such as the open ocean.

Acknowledgements

We thank Captain Marco and his crew at Magdalene Bay Whale Tours for assistance obtaining video of striped marlin. We thank Felipe Galván-Magaña from CICIMAR for logistical support. This project has been funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC 2002/1 "Science of Intelligence" – project number 390523135. In addition, this work has been partially supported by MIAI@Grenoble Alpes, (ANR-19-P3IA-0003)

Ethics

Field data was conducted under permits SGPA/DGVS/02460/18, SGPA/DGVS/01643/19 and SGPA/DGVS/08074/21, and we followed the ASAB ethics (Guidelines for the treatment of animals in behavioural research and teaching 2020, Animal Behaviour 159, i-xi) recommendations for fieldwork.

References

[1] The principles of collective animal behaviour. *Philosophical transactions of the royal society B: Biological Sciences*, 361 (1465):5–22, 2006. 1

[2] Boris Babenko, Ming-Hsuan Yang, and Serge Belongie. Robust object tracking with online multiple instance learning. *IEEE transactions on pattern analysis and machine intelligence*, 33(8):1619–1632, 2010. 4

[3] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, pages 9650–9660, 2021. 4

[4] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017. 3

[5] Anthony I Dell, John A Bender, Kristin Branson, Iain D Couzin, Gonzalo G de Polavieja, Lucas PJJ Noldus, Alfonso Pérez-Escudero, Pietro Perona, Andrew D Straw, Martin Wikelski, et al. Automated image-based tracking and its application in ecology. *Trends in ecology & evolution*, 29(7): 417–428, 2014. 2

[6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 4

[7] Konstantinos G Derpanis, Matthieu Lecce, Kostas Daniilidis, and Richard P Wildes. Dynamic scene understanding: The role of orientation features in space and time in scene classification. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1306–1313. IEEE, 2012. 3

[8] Herbert Edelsbrunner, David Kirkpatrick, and Raimund Seidel. On the shape of a set of points in the plane. *IEEE Transactions on information theory*, 29(4):551–559, 1983. 5

[9] David E Golberg. Genetic algorithms in search, optimization, and machine learning. addion wesley. *Reading*, 673, 1989. 3

[10] Jacob M Graving, Daniel Chae, Hemal Naik, Liang Li, Benjamin Koger, Blair R Costelloe, and Iain D Couzin. Deepposekit, a software toolkit for fast and robust animal pose estimation using deep learning. *Elife*, 8:e47994, 2019. 1

[11] Lars Haalck, Michael Mangan, Antoine Wystrach, Leo Clement, Barbara Webb, and Benjamin Risse. Cater: Combined animal tracking & environment reconstruction. *Science Advances*, 9(16):eadg2094, 2023. 2

[12] MJ Hansen, S Krause, F Dhellemmes, K Pacher, RHJM Kurvers, P Domenici, and J Krause. Mechanisms of prey division in striped marlin, a marine group hunting predator. *Communications Biology*, 5(1):1161, 2022. 1

[13] Adam W. Harley, Zhaoyuan Fang, and Katerina Fragkiadaki. Particle video revisited: Tracking through occlusions using point trajectories. In *Computer Vision – ECCV 2022*, pages 59–75, Cham, 2022. Springer Nature Switzerland. 1, 2

[14] Richard Hartley and Andrew Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003. 3

[15] Berthold Horn. *Robot vision*. MIT press, 1986. 3

[16] Berthold KP Horn and Brian G Schunck. Determining optical flow. *Artificial intelligence*, 17(1-3):185–203, 1981. 3

[17] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017. 3

[18] Lacey F Hughey, Andrew M Hein, Ariana Strandburg-Peshkin, and Frants H Jensen. Challenges and solutions for studying collective animal behaviour in the wild. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 373(1746):20170005, 2018. 1

[19] Roland Kays, Margaret C Crofoot, Walter Jetz, and Martin Wikelski. Terrestrial animal tracking as an eye on life and planet. *Science*, 348(6240):aaa2478, 2015. 2

[20] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *ICCV*, pages 4015–4026, 2023. 4

[21] Benjamin Koger, Adwait Deshpande, Jeffrey T. Kerby, Jacob M. Graving, Blair R. Costelloe, and Iain D. Couzin. Quantifying the movement, behaviour and environmental context of group-living animals using drones and computer vision. *Journal of Animal Ecology*, 92(7):1357–1371, 2023. 1, 2

[22] Alaa Maalouf, Ninad Jadhav, Krishna Murthy Jatavallabhula, Makram Chahine, Daniel M Vogt, Robert J Wood, Antonio Torralba, and Daniela Rus. Follow anything: Open-set detection, tracking, and following in real-time. *IEEE Robotics and Automation Letters*, 9(4):3283–3290, 2024. 2, 4

[23] Alexander Mathis, Pranav Mamidanna, Kevin M. Cury, Taiga Abe, Venkatesh N. Murthy, Mackenzie Weygandt Mathis, and Matthias Bethge. Deeplabcut: markerless pose

estimation of user-defined body parts with deep learning. *Nature Neuroscience*, 21(9):1281–1289, 2018. 1, 2, 4

[24] Jordan K Matley, Natalie V Klinard, Ana P Barbosa Martins, Kim Aarestrup, Eneko Aspillaga, Steven J Cooke, Paul D Cowley, Michelle R Heupel, Christopher G Lowe, Susan K Lowerre-Barbieri, et al. Global trends in aquatic animal tracking with acoustic telemetry. *Trends in Ecology & Evolution*, 37(1):79–94, 2022. 2

[25] Tanmay Nath, Alexander Mathis, An Chi Chen, Amir Patel, Matthias Bethge, and Mackenzie Weygandt Mathis. Using deeplabcut for 3d markerless pose estimation across species and behaviors. *Nature protocols*, 14(7):2152–2176, 2019. 1

[26] Ran Nathan, Christopher T Monk, Robert Arlinghaus, Timo Adam, Josep Alós, Michael Assaf, Henrik Baktoft, Christine E Beardsworth, Michael G Bertram, Allert I Bijleveld, et al. Big-data approaches lead to an increased understanding of the ecology of animal movement. *Science*, 375(6582): eabg1780, 2022. 2

[27] Talmo D Pereira, Nathaniel Tabris, Arie Matsliah, David M Turner, Junyu Li, Shruthi Ravindranath, Eleni S Papadoyannis, Edna Normand, David S Deutsch, Z Yan Wang, et al. Sleap: A deep learning system for multi-animal pose tracking. *Nature methods*, 19(4):486–495, 2022. 1

[28] Alfonso Pérez-Escudero, Julián Vicente-Page, Robert C Hinz, Sara Arganda, and Gonzalo G De Polavieja. idtracker: tracking individuals in a group by automatic identification of unmarked animals. *Nature methods*, 11(7):743–748, 2014. 1

[29] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *Int. Conf. on Machine Learning*, pages 8748–8763. PMLR, 2021. 4

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

[31] Yi Wu, Jongwoo Lim, and Ming-Hsuan Yang. Online object tracking: A benchmark. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2411–2418, 2013. 4

[32] Zichao Zhang and Davide Scaramuzza. A tutorial on quantitative trajectory evaluation for visual (-inertial) odometry. In *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 7244–7251. IEEE, 2018. 4