International Journal of Computer Vision
https://doi.org/10.1007/s11263-022-01711-8

AnimalTrack: A Benchmark for Multi-Animal Tracking in the Wild

Libo Zhang[1,2,3] · **Junyuan Gao**[1,2] · **Zhen Xiao**[1] · **Heng Fan**[4]

Received: 29 April 2022 / Accepted: 7 November 2022

Abstract
Multi-animal tracking (MAT), a multi-object tracking (MOT) problem, is crucial for animal motion and behavior analysis and has many crucial applications such as biology, ecology and animal conservation. Despite its importance, MAT is largely under-explored compared to other MOT problems such as multi-human tracking due to the scarcity of dedicated benchmarks. To address this problem, we introduce *AnimalTrack*, a dedicated benchmark for multi-animal tracking in the wild. Specifically, AnimalTrack consists of 58 sequences from a diverse selection of 10 common animal categories. On average, each sequence comprises of 33 target objects for tracking. In order to ensure high quality, every frame in AnimalTrack is manually labeled with careful inspection and refinement. To our best knowledge, AnimalTrack is the *first* benchmark dedicated to multi-animal tracking. In addition, to understand how existing MOT algorithms perform on AnimalTrack and provide baselines for future comparison, we extensively evaluate 14 state-of-the-art representative trackers. The evaluation results demonstrate that, not surprisingly, most of these trackers become degenerated due to the differences between pedestrians and animals in various aspects (e.g., pose, motion, and appearance), and more efforts are desired to improve multi-animal tracking. We hope that AnimalTrack together with evaluation and analysis will foster further progress on multi-animal tracking. The dataset and evaluation as well as our analysis will be made available upon the acceptance.

Keywords Tracking · Multi-object tracking (MOT) · Multi-animal tracking (MAT) · AnimalTrack · Tracking evaluation

1 Introduction

In this paper, we are interested in multi-animal tracking (MAT), a typical kind of multi-object tracking (MOT) yet heavily under-explored. MAT is critical for understanding and analyzing animal motion and behavior, and thus has a wide range of applications in zoology, biology, ecology, and animal conservation. Despite the importance, MAT is less studied in the tracking community.

Currently, the MOT community mainly focuses on pedestrians and vehicles tracking, with numerous benchmarks introduced in recent years (Dendorfer et al., 2020; Geiger et al., 2012; Milan et al., 2016; Zhu et al., 2021). Compared with MOT on pedestrians and vehicles, MAT is challenging because of several following properties of animals:

- *Uniform appearance* Different from pedestrians and vehicles in existing MOT benchmarks that usually have distinguishable appearances (e.g., color and texture), most animals have uniform appearances that visually look extremely similar (see Fig. 1 for example). As a consequence, it is difficult to leverage their visual features

Communicated by Angjoo Kanazawa.

Libo Zhang and Junyuan Gao make equal contributions to this work. Heng Fan is the corresponding author.

✉ Heng Fan
heng.fan@unt.edu

Libo Zhang
libo@iscas.ac.cn

Junyuan Gao
2018091621016@uestc.edu.cn

Zhen Xiao
isrc_exam@iscas.ac.cn

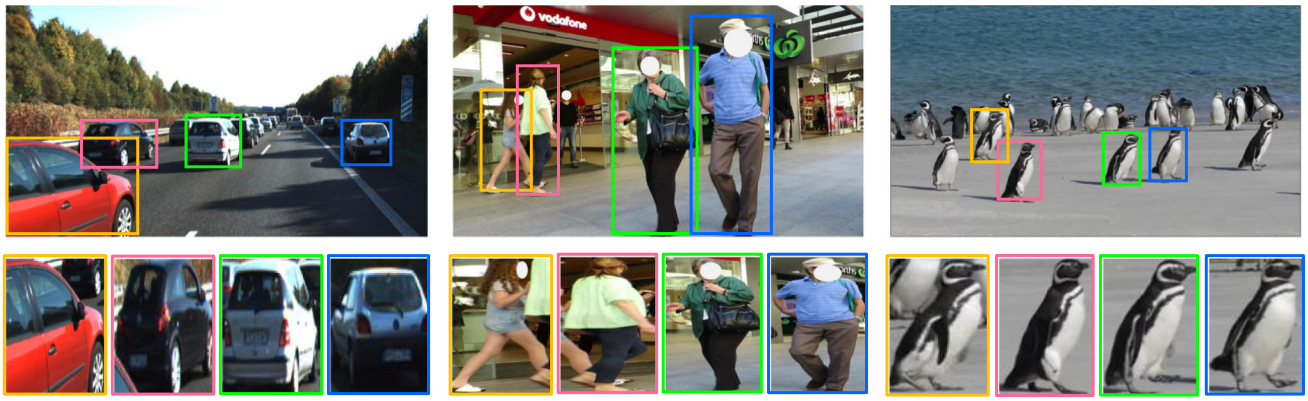
[1] State Key Laboratory of Computer Science, Institute of Software Chinese Academy of Sciences, Beijing, China

[2] Hangzhou Institute for Advanced Study, University of Chinese Academy of Sciences, Hangzhou, China

[3] Nanjing Institute of Software Technology, Nanjing, China

[4] Department of Computer Science and Engineering, University of North Texas, Denton, TX, USA

Published online: 18 November 2022



(a) Vehicle targets (partial) **(b)** Pedestrian targets (partial) **(c)** Animal (e.g., *penguin*) targets (partial)

Fig. 1 Comparison of MOT on vehicle, pedestrian and animal. **a** Shows multi-vehicle tracking from KITTI (Geiger et al., 2012), **b** Multi-pedestrian tracking from MOT17 (Milan et al., 2016) and **c** Multi-animal tracking from the proposed AnimalTrack (Please note that, we only show part of the targets in each image for simplicity). We can observe that, animals are more difficult to be distinguished due to uniform appearance compared to vehicles and pedestrians. Best viewed in color and by zooming in for all figures in this paper (Color figure online)

only to distinguish different animals by using regular association (e.g., re-identification) models.

- *Diverse pose* Animals often possess diverse poses in a video sequence. For example, a goose may *walk* or *run* on the ground, or *swim* in water, or *fly* in air, leading to significantly different poses. This diverse pose variation of animals may cause difficulties in detector design for tracking.
- *Complex motion* In addition to the aforementioned challenges, animals also have larger-range motions due to their diverse poses. For example, animals may frequently change motions from *flying* to *swimming*, or inverse. These complicated motion patters lead to higher requirement on motion modeling for when tracking animal targets.

The above properties of animals bring in technical difficulties for MAT, making it a less-touched problem. In addition, another more important reason why MAT is under-explored is the scarcity of a benchmark. Benchmark plays a crucial role in advancing multi-object tracking. As a platform, it allows researchers to develop their algorithms and fairly assess, compare and analyze different approaches for improvement. Currently, there exist many datasets (Bai et al., 2021; Dave et al., 2020; Dendorfer et al., 2020; Du et al., 2018; Geiger et al., 2012; Milan et al., 2016; Zhu et al., 2021) for MOT on different subjects in various scenarios. Nevertheless, there is *no* available benchmark dedicated for multiple animal tracking. Although some of the datasets (e.g., Bai et al., 2021; Dave et al., 2020) consist of video sequences involved with animal targets, they are limited in either video quantity and animal categories (Bai et al., 2021) or number of animal tracklets (Dave et al., 2020), which makes them not an ideal platform for studying MAT. In order to facilitate MOT on animals, a dedicated benchmark is urgently required for both designing and evaluating MAT algorithms.

Contribution Thus motivated, in this paper we make the *first* step for studying the MAT problem by introducing **AnimalTrack**, a dedicated benchmark for multi-animal tracking in the wild. Specifically, AnimalTrack consists of 58 video sequences, which are selected from 10 common animal categories in our real life. On average, each video sequence contains 33 animals for tracking. There are more than 24.7K frames in total in AnimalTrack, and every frame is manually labeled with multiple axis-aligned bounding boxes. Careful inspection and refinement are performed to ensure high-quality annotations. To the best of our knowledge, AnimalTrack is the *first* benchmark dedicated to the task of MAT.

In addition, with the goal of understanding how existing MOT algorithms perform on the newly developed AnimalTrack for future improvements, we extensively evaluate 14 popular state-of-the-art MOT algorithms. We conduct in-depth analysis on the evaluation results of these trackers. From the results, not surprisingly, we observe that, most of these trackers, designed for pedestrian or vehicle tracking, are greatly degraded when directly applied for animal tracking on AnimalTrack because of the aforementioned properties of animals. We hope that these evaluation and analysis can offer baselines for future comparison on AnimalTrack and provide guidance for tracking algorithm design.

Besides the analysis on overall performance of different tracking algorithms, we also independently study the important association techniques that are indispensable for current multi-object tracking. In particular, we compare and analyze several popular association strategies. The analysis is

expected to provide some guidance for future research when choosing appropriate association baseline for improvements.

In summary, we make the following contributions: (i) We introduce the AnimalTrack, which is, to the best of our knowledge, the first benchmark dedicated to multi-animal tracking. (ii) We extensively evaluate 14 representative state-of-the-art MOT approaches to provide future comparison on AnimalTrack. (iii) We conduct in-depth analysis for the evaluations of existing approaches, offering guidance for future algorithm design.

By releasing AnimalTrack, we hope to boost the future research and applications of multiple animal tracking. Our project with data and evaluation results will be made publicly available upon the acceptance of this work.

The rest of this paper is organized as follows. Secion 2 discusses related trackers and benchmarks of this work. Secion 3 will illustrate the proposed AnimalTrack in details. Section 4 demonstrates the evaluation results on AnimalTrack. Section 5 presents several discussions in this work, followed by conclusion in Sect. 6.

2 Related Work

MAT belongs to the problem of MOT. In this section, we will discuss related MOT algorithms and existing benchmarks that are related to AnimalTrack. Besides, we will also briefly review other animal-related vision benchmarks.

2.1 Multi-Object Tracking Algorithms

MOT is a fundamental problem in computer vision and has been actively studied for decades. In this subsection, we will briefly review some representative works and refer readers to recent surveys (Ciaparrone et al., 2020; Emami et al., 2020; Luo et al., 2021) for more tracking algorithms.

One popular paradigm is called Tracking-by-Detection which decomposes MOT into two subtasks including detecting objects (Lin et al., 2017; Ren et al., 2015) in each frame and then associating the same target to generate trajectories using optimization techniques (e.g., Hungarian algorithm Bewley et al. 2016 and network flow algorithm Dehghan et al. 2015). Within this framework, numerous approaches have been introduced (Bewley et al., 2016; Chu et al., 2019; Shuai et al., 2021; Tang et al., 2017; Wojke et al., 2017; Xu et al., 2019; Yin et al., 2020; Zhu et al., 2018). In order to improve the data association in MOT, some other works propose to directly incorporate the optimization solvers in association into learning (Brasó & Leal-Taixé, 2020; Chu & Ling, 2019; Dai et al., 2021; Schulter et al., 2017; Xu et al., 2020), which is greatly beneficial for improving tracking performance from end to end learning in deep network.

In addition to the Tracking-by-Detection framework, another MOT architecture named Joint-Detection-and-Tracking has recently drawn increasing attention in the community due to efficiency and simplicity. This framework learns to detect and associate target objects at the same time, largely simplifying the MOT framework. Many efficient approaches (Bergmann et al., 2019; Liang et al., 2022; Lu et al., 2020; Wang et al., 2020; Zhang et al., 2021b; Zhou et al., 2020) have been proposed based on this architecture. More recently, motivated by the power of Transformer (Vaswani et al., 2017), the attention mechanism has been introduced for MOT (Meinhardt et al., 2022; Sun et al., 2020) and demonstrate state-of-the-art performance.

2.2 Multi-Object Tracking Benchmarks

Benchmarks are important for the development of MOT. In recent years, many benchmarks have been propose.

PETS2009 PETS2009 (Ferryman & Shahrokni, 2009) is one of the earliest benchmarks for MOT. It contains 3 videos sequences for pedestrian tracking.

KITTI KITTI (Geiger et al., 2012) is introduced for autonomous driving. It comprises of 50 video sequences and focuses on tracking pedestrian and vehicle in traffic scenarios. Besides 2D MOT, KITTI also supports 3D MOT.

UA-DETRAC UA-DETRAC (Wen et al., 2020) includes 100 challenge sequences captured from real world traffic scenes. This dataset provides rich annotations for multi-object tracking such as illumination, occlusion, truncation ration, vehicle type and bounding box.

MOTChallenge MOTChallenge (Dendorfer et al., 2021) contains a series of benchmarks. The first version MOT15 (Leal-Taixé et al., 2015) consists of 22 sequences for tracking. Due to low difficulty of videos in MOT15, MOT16 (Milan et al., 2016) compiles 14 new and more challenging sequences compared to MOT15. MOT17 (Milan et al., 2016) uses the same videos as in MOT16 but improves the annotation and applies a different evaluation system. Later, MOT20 (Dendorfer et al., 2020) is presented with 8 new sequences, aiming at MOT in crowded scenes.

MOTS MOTS (Voigtlaender et al., 2019) is a newly introduced dataset for multi-object tracking. In addition to 2D bounding box, MOTS also provides pixel mask for each target, aiming at simultaneous tracking and segmentation.

BDD100K BDD100K (Yu et al., 2020) is recently proposed for video understanding in traffic scenes. It provides multiple tasks including multi-object tracking.

TAO TAO (Dave et al., 2020) is a large-scale dataset for tracking any objects. It consists of 2907 videos from 833

categories. TAO sparsely labels objects every 30 frames. Its average trajectories is 6.

GMOT-40 GMOT-40 (Bai et al., 2021) is a recently proposed benchmark that aims at one-shot MOT. It consists of 40 sequences from 10 categories. Each sequence provides one instance for tracking multiple targets of the same class.

UAVDT-MOT UAVDT-MOT (Du et al., 2018) consists of 100 challenging videos that are captured with a drone. These videos mainly cover pedestrian and vehicle for tracking. The goal of UAVDT-MOT is to facilitate multi-object tracking in aerial views.

VisDrone-MOT Similar to UAVDT-MOT, VisDrone-MOT (Zhu et al., 2021) also focuses on MOT with drone. The difference is VisDrone-MOT introduces more object categories, making it more challenging.

ImageNet-Vid ImageNet-Vid (Russakovsky et al., 2015) is one of the most popular benchmarks for visual recognition. It provides more than 5,000 video sequences collected from 30 categories for various visual tasks including video object detection and tracking.

YT-VIS YT-VIS (Yang et al., 2019) is a large-scale dataset containing 2,883 videos from 40 categories. It provides mask annotations for target objects and aims at facilitating the task of video instance segmentation and tracking.

DanceTrack DanceTrack (Sun et al., 2022) is a large-scale benchmark with 100 videos. The aim of DanceTrack is to explore multi-human tracking in uniform appearance and diverse motion.

Different from the above datasets for MOT on pedestrians, vehicles or other subjects, AnimalTrack focuses on dense multi-animal tracking in the wild. Although some of the benchmarks (e.g., TAO Dave et al. 2020 and GMOT-40 Bai et al. 2021) contain animal targets for tracking, they have limitations for MAT. For TAO (Dave et al., 2020), the average trajectory is 6 and even lower for animal, the average trajectory is 4. Nevertheless, in practice in the wild, it is very common to see objects moving in a dense group. The sparse trajectory in TAO may limits its usage for dense tracking case. In addition, TAO is sparsely annotated 30 frames, resulting in difficulty for trackers in learning temporal motion. Despite several animal videos, GMOT-40 (Bai et al., 2021) is limited in animal categories (4 classes) and video quantity (12 in total). Besides, GMOT-40 has a different aim for one-shot MOT. Thus, no training data is provided. Compared to TAO (Dave et al., 2020) and GMOT-40 (Bai et al., 2021), AnimalTrack is dense in trajectories and annotation (i.e., per-frame manual annotation) as well as diverse in animal classes.

We are also aware that there exist a few datasets (Betke et al., 2007; Bozek et al., 2018; Khan et al., 2004) for animal tracking. However, these datasets are usually small (e.g., with 1 or 2 video sequences) and limited to special animal category (e.g., Khan et al. 2004 for ant, Betke et al. 2007 for bat, Bozek et al. 2018 for bee), and therefore may not be suitable for animal tracking in the deep learning era. Unlike these animal tracking datasets, our AnimalTrack has more classes with more videos.

2.3 Other Animal-Related Vision Benchmarks

Our AnimalTrack is also related to many other animal-related vision benchmarks outside MOT. The work of Cao et al. (2019) introduces a large-scale benchmark for animal pose estimation, which is later extended by Yu et al. (2021) by adding more images and further increasing categories. In Mathis et al. (2021), the authors introduce a benchmark dedicated to horse pose estimation. The work of Bala et al. (2020) proposes a 3D animal pose estimation benchmark. The work of Parham et al. (2018) presents a new dataset for animal localization in the wild. A benchmark for tiger re-identification is proposed in Li et al. (2019). In Iwashita et al. (2014), the authors build a benchmark for animal activity recognition in videos. Different from these benchmarks, the proposed AnimalTrack focuses on multiple animal tracking.

3 AnimalTrack

3.1 Design Principle

AnimalTrack expects to provide the community with a new dedicated platform for studying MOT on animal. In particular, in the deep learning era, it aims at both training and evaluation for deep trackers. To this end, we follow three principles in constructing our AnimalTrack:

- *Dedicated benchmark* One motivation behind AnimalTrack is to provide a dedicated benchmark for animal tracking. Especially, considering that current deep models usually require a large mount of data for training, we hope to compile a dedicated platform containing at least 50 video sequences with at least 20K frames for animal tracking.
- *High-quality dense annotations* The annotations of a benchmark are crucial for both algorithm development and evaluation. To this end, we provide per-frame manual annotations for every sequence of AnimalTrack to ensure high annotation quality, which is different than many MOT benchmarks providing only spare annotations.
- *Dense trajectories* In real world, it is common to see animals moving in a dense group. AnimalTrack aims at

Table 1 Statistics on the proposed AnimalTrack and its comparison with several multi-object tracking benchmarks and animal videos in GMOT-40 and TAO

Benchmark	Tracking on other subjects								Tracking on animals		
	KITTI (Geiger et al., 2012)	MOT17 (Milan et al., 2016)	MOT20 (Dendorfer et al., 2020)	UAVDT-MOT (Du et al., 2018)	ImageNet-Vid (Russakovsky et al., 2015)	YT-VIS (Yang et al., 2019)	TAO (Dave et al., 2020)	GMOT-40 (Bai et al., 2021)	GMOT-40-Anim. (Bai et al., 2021)	TAO-Anim. (Dave et al., 2020)	AnimalTrack (ours)
Videos	50	14	8	100	5354	2883	2907	40	12	39	58
Categories	5	1	1	3	30	40	833	10	3	39	10
Min. len. (s)	n/a	17	17	2.8	0.2	1	n/a	3.0	3.0	1.0	6.5
Avg. len. (s)	10.0	33.0	66.8	266.7	12.1	5.6	36.8	8.9	7.1	22.0	14.2
Max. len. (s)	n/a	85.0	133.0	99.0	219.7	7.2	n/a	24.2	24.2	93.0	75.6
Total len. (s)	498.0	463.0	535.0	2,666.7	64,547.9	16,015.2	106,978.0	356.0	85.5	859.0	823.7
Avg. tracks	52	95	479	270	n/a	n/a	6	51	70	4	33
Max. tracks	n/a	222	1211	n/a	n/a	n/a	10	128	128	10	133
Total tracks	2600	1331	3833	2700	n/a	n/a	17,287	2026	837	250	1927
Frame rate	30	25	25	30	25	30	30	30	30	30	30
Ann. FPS	10	30	30	6	25	5	1	30	30	1	30
Total boxes	80K	300K	2,102K	840K	n/a	131K	333K	256K	63K	3.4K	429K
Total frames	15K	11K	13K	40K	1614K	480.5K	2674K	9K	2.6K	2.5K	24.7K

"n/a" means that the statistics can not be obtained because some of the benchmarks do not provide the test set

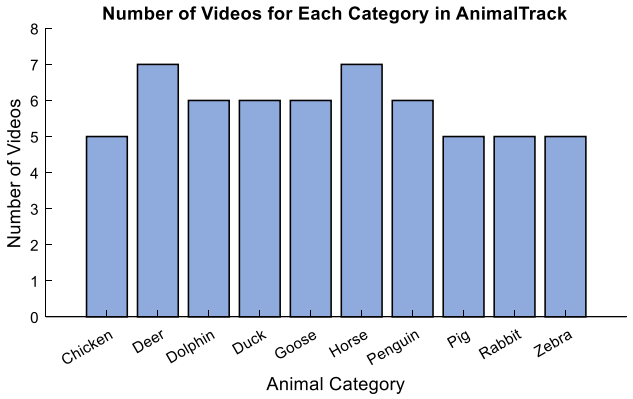


Fig. 2 Number of video sequences for each animal class in AnimalTrack. Each category consists of at least 5 and at most 7 sequences

such dense tracking on animals and expects an average video trajectory at least 25.

3.2 Data Collection

Our AnimalTrack focuses on dense multi-animal tracking. We start benchmark construction by selecting 10 common animal categories that are generally *dense* and *crowded* in the wild. These categories are *Chicken*, *Deer*, *Dolphin*, *Duck*, *Goose*, *Horse*, *Penguin*, *Pig*, *Rabbit*, and *Zebra*, perching in very different environments. Although TAO consists of more classes than ours, many categories in TAO are not available for dense multi-object tracking, which is different than our aim in this work.

After determining the animal classes, we search raw video sequences[1] of each class from YouTube (https://www.youtube.com/), the largest and the most popular video platform in the world. Initially, we have collected over 500 candidate sequences. After a joint consideration of both video quality and our principles, from these raw sequences we choose 58 video clips that are finally available for our task. For each category, there are at least 5 and at most 7 sequences, showing balance in category to some extent. Figure 2 demonstrates the number of sequences for each category in AnimalTrack. It is worth noticing that, in each single video sequence, there is only one category of animal to track. Because one of our goals is focused on dense multi-animal tracking. During the data collection, such dense-scenario videos with crowded animals usually contain one category of animals. Because of this, we decide each video consisting of one animal category for tracking in AnimalTrack.

Finally, we compile a dedicated benchmark for multi-animal tracking by collecting 58 video sequences with more than 24.7K frames and 429K boxes. The average video length is 426 frames. The longest sequence contains 2269 frames, while the shortest one consist of 196 frames. The total number of tracks in AnimalTrack is 1927, and the average number of tracks is 33. To our best knowledge, AnimalTrack is by far the largest benchmark dedicated for animal tracking. Table 1 summarizes detailed statistics on AnimalTrack and comparison with several popular MOT benchmarks and animal videos in GMOT-40 and TAO.

Table 2 Annotation format in AnimalTrack

Position	Name	Description
1	*Frame number*	Frame in which the target appears; starting from 1
2	*Identifier*	An unique ID for each trajectory
3	*Box left*	Coordinate of top-left corner of annotated object
4	*Box top*	Coordinate of top-left corner of annotated object
5	*Box width*	Width of annotated object
6	*Box height*	Height of annotated object
7	*Confidence*	Flag that indicates if the box is considered (1) or ignored (−1) for evaluation; the confidence for all targets in AnimalTrack is set to 1
8	*Class*	Type of annotated object
9	*Visibility*	Visibility ratio of object; we ignore it by setting its value to −1 in AnimalTrack

[1] Each video sequence is collected under the Creative Commons license.

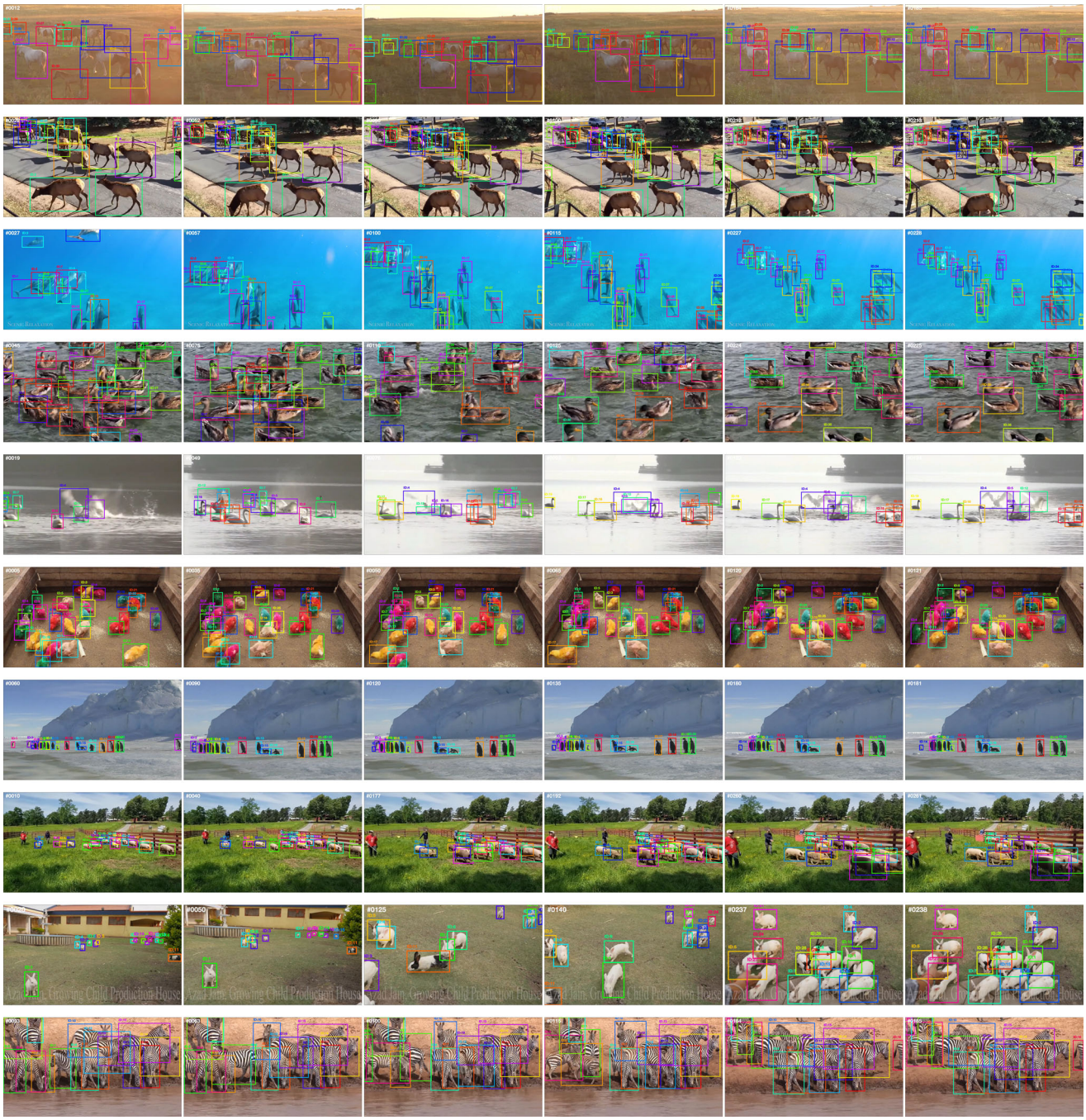


Fig. 3 Visualization of consecutive annotated example images at different annotation frequencies of 1FPS (column 1–2), 15FPS (column 3–4) and 30FPS (column 5–6) from each category in AnimalTrack (from top to bottom: horse-3, deer-4, dolphin-6, duck-4, goose-5, chicken-2, penguin-5, pig-5, rabbit-2 and zebra-4). We can observe that animals from the same class usually have uniform appearances and complex pose and motion patterns, which brings new challenges for tracking animals

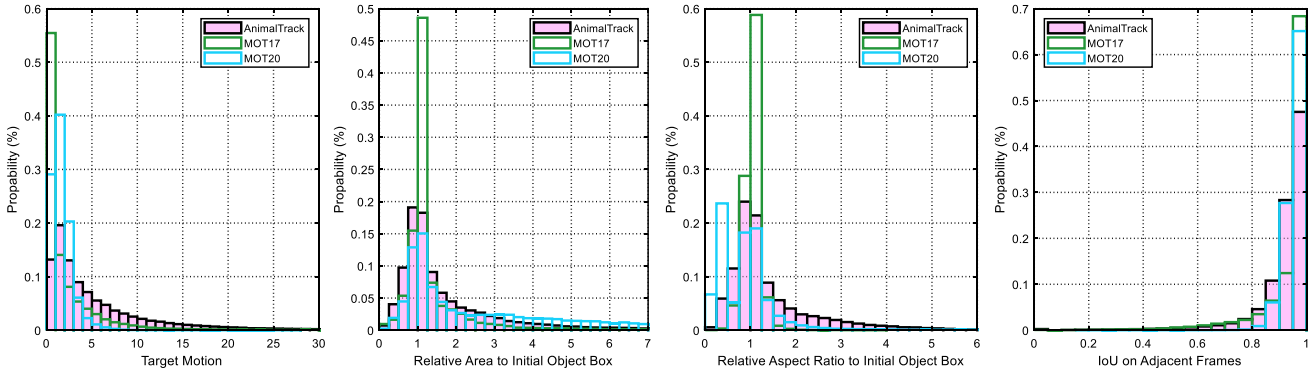


Fig. 4 Statistics of object motion, area and aspect ratio change compared to initial object and IoU on adjacent object boxes in AnimalTrack and comparison with popular pedestrian tracking benchmarks including MOT17 (Milan et al., 2016) and MOT20 (Dendorfer et al., 2020). We can observe that the animals in our benchmark have more complex pose and motions

3.3 Annotation

We use the annotation tool DarkLable[2] to annotate the videos in AnimalTrack. Following popular MOTChallenge (Dendorfer et al., 2021), we annotate each target in videos with object identifier, axis-aligned bounding box and other information. Table 2 demonstrates the annotation format for each target in AnimalTrack. Note that, slightly different from MOTChallenge, we do annotate the visibility ratio of each target because it is hard to accurately measure the visibility of the target in real world scenarios. However, we still keep it (set to -1) for padding to MOTChallenge format.

To provide consistent annotations, we follow the following labeling rules. For target object that is fully visible or partially occluded, a full-body box is annotated. If the object is under full occlusion, we do not label it. When this object re-appears in the view in future, we annotate it with the same identifier. For target objects out of view, they are assigned with new identifiers when re-entering the view.

In order to ensure the high-quality annotations of videos in AnimalTrack, we adopt a multi-round strategy. In specific, a group of volunteers who are familiar with the tracking topic and an expert (e.g., PhD student working on related areas) will first participate in manually annotating each target object in the videos. After this, a group of experts will carefully inspect the initial annotations. If these initial annotation results are not unanimously agreed by all the experts, they will be returned to the original labeling team for adjustment or refinement. We repeat this process until all annotations are satisfactorily completed.

To show the quality of our annotations, we visualize a few annotated sample from each category in AnimalTrack. In particular, we demonstrate the annotated samples from two consecutive frames at different annotation frequencies of 1 FPS, 15 FPS and 30 FPS, as shown in Fig. 3. From Fig. 3, we can see the annotations of our AnimalTrack are consistent and high-quality.

3.4 Statistics of Annotation

To better understand animal pose and motion, we show representative statistics of the annotation boxes of objects in AnimalTrack in Fig. 4. In particular, we demonstrate the object motion, relative area to initial object box, relative aspect ratio (aspect ratio is defined as ratio of width and width) and Intersection over Union (IoU) on object boxes in adjacent frames. From Fig. 4, we can clearly observe that the animal targets vary rapidly in terms of spatial pose and temporal motions.

In addition, we compare AnimalTrack and popular pedestrian tracking benchmarks including MOT17 (Milan et al., 2016) and MOT20 (Dendorfer et al., 2020). From the comparison in Fig. 4, we can see that animals have faster motion than pedestrians. Moreover, the pose variations of animals are more complex, which consequently causes new challenges in tracking animals.

3.5 Dataset Split

AnimalTrack consists of 58 video sequences. We utilize 32 out of 58 for training and the rest 26 for testing. In specific, for category with K videos, we select $K/2$ videos for training and the rest for testing if K is a even number, otherwise we choose $(K+1)/2$ videos for training and the rest for testing. During dataset splitting, we try our best to keep the distributions of training and testing set as close as possible. Table 3 compares the statistics of training/testing sets in AnimalTrack. Note that, the number of frames for the testing set is slightly more than that for the training set. The reason is that the testing

[2] The annotation tool is available at https://github.com/darkpgmr/DarkLabel.

Table 3 Comparisons between training set (i.e., $\text{AnimalTrack}_{\text{Tra}}$) and testing set (i.e., $\text{AnimalTrack}_{\text{Tst}}$) of AnimalTrack

	Videos	Categories	Min. len. (s)	Avg. len. (s)	Max. len. (s)	Total len. (s)	Avg. tracks	Total tracks	Total boxes	Total images
$\text{AnimalTrack}_{\text{Tra}}$	32	10	6.9	12.0	50.3	384.8	26	823	186K	11.5K
$\text{AnimalTrack}_{\text{Tst}}$	26	10	6.5	16.9	75.6	438.9	42	1104	243K	13.2K

set contains more longer video sequences for challenging evaluation. The detailed spilt will be released at our project website.

4 Evaluation

4.1 Evaluation Metric

For comprehensive evaluation of different tracking algorithms, we use multiple metrics. Specifically, we employ the recently proposed higher order tracking accuracy (HOTA) from Luiten et al. (2021), commonly used CLEAR metrics from Bernardin and Stiefelhagen (2008) including multiple object tracking accuracy (MOTA), mostly tracked targets (MT), mostly lost targets (ML), false positives (FP), false negatives (FN), ID switches (IDs) and number of times a trajectory is fragmented (FM) and ID metrics from Ristani et al. (2016) such as identification precision (IDP), identification recall (IDR) and related F1 score (IDF1) which is defined as the ratio of correct detections to the average number of GT and computed detections. Many previous works employ MOTA as the main metric (e.g., for ranking). Nevertheless, a recent study (Luiten et al., 2021) shows that MOTA may bias to target detection quality instead of target association accuracy. Considering this, we follow (Geiger et al., 2012; Sun et al., 2022) to adopt HOTA as the main metric in evaluation. For detailed definitions of these metrics, we refer readers to Bernardin and Stiefelhagen (2008); Ristani et al. (2016); Luiten et al. (2021).

4.2 Evaluated Trackers

Understanding how existing MOT algorithms perform on AnimalTrack is crucial for future comparison and also beneficial for tracker design. To such end, we extensively evaluate 14 state-of-the-art multi-object tracking approaches.

These approaches include SORT (Bewley et al., 2016) (ICIP'2016), DeepSort (Wojke et al., 2017) (ICIP'2017), IoUTrack (Bochinski et al., 2017) (AVSS'2017), JDE (Wang et al., 2020) (ECCV'2020), FairMOT (Zhang et al., 2021b) (IJCV'2021), CenterTrack (Zhou et al., 2020) (ECCV'2020), CTracker (Peng et al., 2020) (ECCV'2020), QDTrack (Pang et al., 2021) (CVPR'2021), ByteTrack (Zhang et al., 2021a) (arXiv'2021), Tracktor++ (Bergmann et al., 2019) (ICCV'2019), TADAM (Guo et al., 2021) (CVPR'2021), Trackformer (Meinhardt et al., 2022) (CVPR'2022), OMC (Liang et al., 2022) (AAAI'2022) and TransTrack (Sun et al., 2020) (arXiv'2020). Notably, among these approaches, TransTrack and Trackformer are two recently proposed trackers using Transformer. Despite excellent performance on pedestrian tracking, these trackers quickly degrade in tracking animals as shown in later experimental results.

In this work, following the popular MOT challenge, we adopt a private-detection setting, where each tracker is allowed to use its own detector, for performance evaluation and comparison. In particular, for all the chosen trackers, we use their architectures (including the detection component) as they are, without any modifications, but train them on our AnimalTrack. The reasons why we utilize their default architectures for training are two-fold. First, different approach may need different training strategies, which makes it difficult to optimally train each tracker for best performance. Moreover, inappropriate training settings may decrease the performance for certain trackers. Second, the original configuration for each tracker has been verified by authors. Thus, it is reasonable to assume that each tracker is able to obtain decent results even without modification. It is worth noting that, in this private setting, the detection component in each tracker is trained as well on AnimalTrack for localizing foreground target objects (i.e., objects in all animal categories in AnimalTrack) for tracking. Once training on AnimalTrack completed, these trackers will be evaluated.

4.3 Evaluation Results

In this work, the evaluation of each tracking algorithm is conducted in "*private setting*" in which each tracker should perform both object detection and target association.

4.3.1 Overall Performance

We extensively evaluate 14 state-of-the-art tracking algorithms. Table 4 shows the evaluation results and comparison.

From Table 4, we observe that QDTrack shows the best overall result by achieving 47.0% HOTA score and TransTrack the second best with 45.4% HOTA score, respectively. QDTrack densely samples numerous regions from images for similarity learning and thus can alleviate the problem of complex animal poses in detection in some degree, as evidenced by its best result of 55.7% on MOTA that focuses more on detection quality. This dense sampling strategy not only improves detection but also benefits later association, which is shown by its best 56.3% IDF1 score. TransTrack obtains the second best overall result with 45.4% HOTA score. On IDF1, it also exhibits the second best result with 53.4%. TransTrack utilizes the query-key mechanism in Transformer for multi-object tracking. The competitive performance of TransTrack shows the potential of Transformer for MOT. We notice that another Transformer-based tracker Trackformer shows poorer performance compared to TransTrack. We argue that the reason is because of its relatively weaker detection module. Tracktor++ shows the second best MOTA result with 55.2% owing to its adoption of strong Faster R-CNN (Ren et al., 2015) for detection. Com-

Table 4 Overall evaluation results and comparison of different tracking algorithms on AnimalTrack

Tracker	HOTA (%)	MOTA (%)	IDF1 (%)	IDP (%)	IDR (%)	MT	PT	ML↓	FP↓	FN↓	IDs↓	FM↓
SORT (Bewley et al., 2016)	42.8	55.6	49.2	58.5	42.4	333	470	301	19,099	86,257	2530	3730
IOUTrack (Bochinski et al., 2017)	41.6	55.7	45.7	51.9	40.7	**388**	454	**262**	25,206	**77,847**	4639	5259
DeepSORT (Wojke et al., 2017)	32.8	41.4	35.2	49.7	27.2	213	452	439	14,131	124,747	3503	4527
JDE (Wang et al., 2020)	26.8	27.3	31.0	51.0	22.0	106	414	584	17,887	155,623	3187	5031
FairMOT (Zhang et al., 2021b)	30.6	29.0	38.8	62.8	28.0	143	462	499	17,653	152,624	2335	5447
CenterTrack (Zhou et al., 2020)	9.9	1.6	7.0	8.9	5.8	265	423	416	32,050	117,614	89,655	7583
CTracker (Peng et al., 2020)	13.8	14.0	14.7	35.2	9.3	20	313	771	**13,092**	192,660	3437	8019
Tracktor++ (Bergmann et al., 2019)	44.2	*55.2*	51.0	58.5	45.1	364	472	*268*	25,477	*81,538*	*1976*	4149
ByteTrack (Zhang et al., 2021a)	40.1	38.5	51.2	*64.9*	42.3	310	465	329	31,591	116,587	**1309**	**3513**
QDTrack (Pang et al., 2021)	**47.0**	**55.7**	**56.3**	**65.6**	**49.3**	*367*	420	317	22,696	83,057	1970	5656
TADAM (Guo et al., 2021)	32.5	36.5	37.2	44.4	32.0	258	**495**	351	41,728	110,048	2538	4469
OMC (Liang et al., 2022)	43.0	53.4	50.3	61.8	42.4	324	478	302	*15,910*	92,570	4938	7162
Trackformer (Meinhardt et al., 2022)	31.0	20.4	36.5	40.9	32.8	230	*491*	383	70,404	118,724	4355	*3725*
TransTrack (Sun et al., 2020)	*45.4*	48.3	*53.4*	63.4	*46.1*	327	416	361	28,553	95,212	1978	6459

The best two results on each metric are highlighted in bold and italic fonts

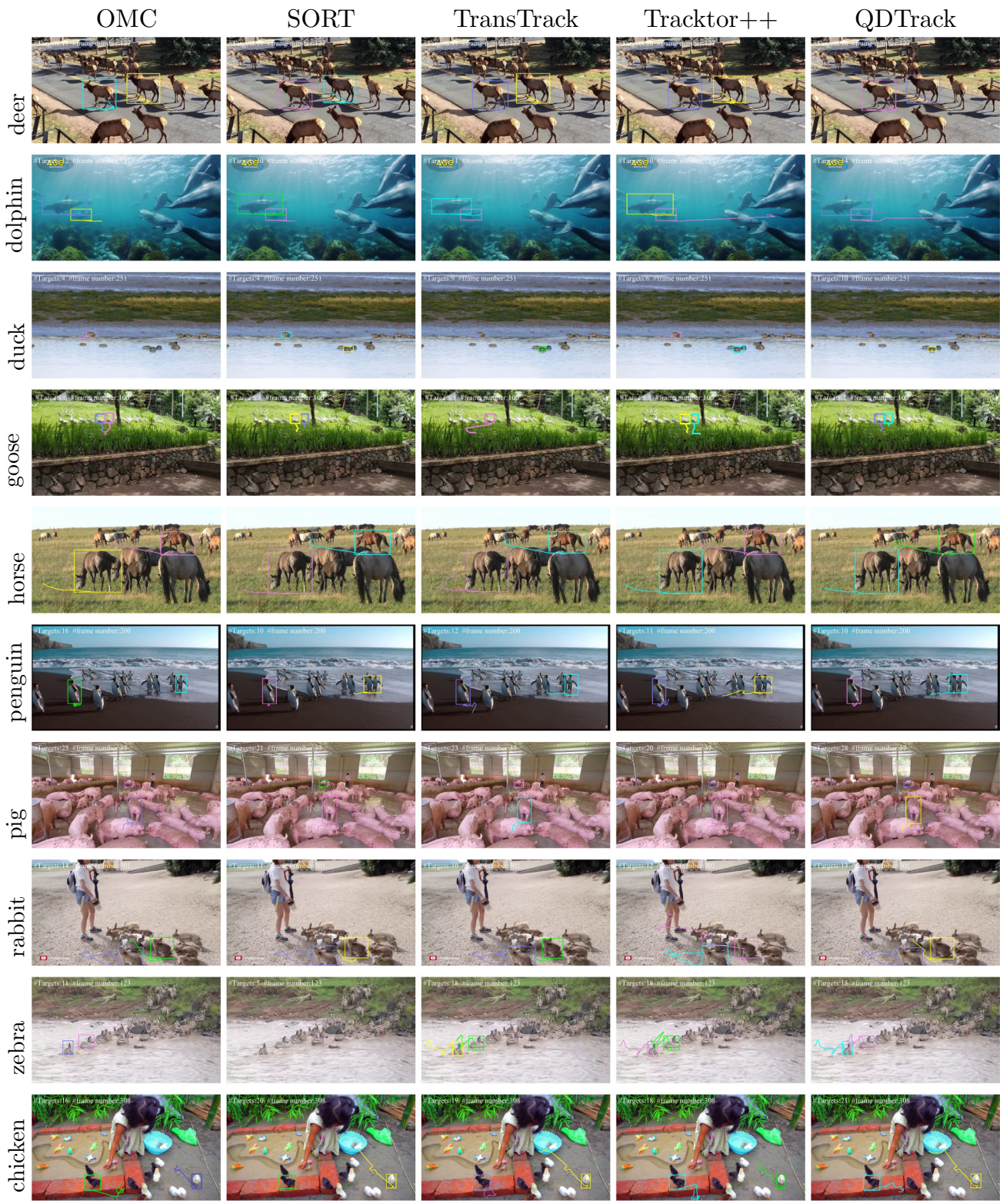
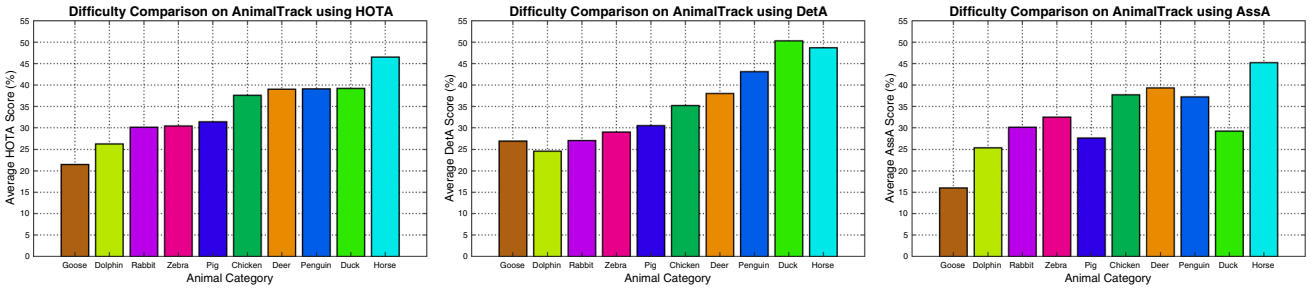


Fig. 5 Visualization of top five trackers consisting of OMC (Liang et al., 2022), SORT (Bewley et al., 2016), TransTrack (Sun et al., 2020), Tracktor++ (Bergmann et al., 2019) and QDTrack (Pang et al., 2021) based on HOTA scores on several sequences. Each color represents a tracking trajectory. Please notice that, we only show two trajectories for each tracker in the visualization for simplicity (Color figure online)



(a) Difficulty comparison using HOTA **(b)** Difficulty comparison using DetA **(c)** Difficulty comparison using AssA

Fig. 6 Difficulty comparison of different categories in AnimalTrack using different metrics including HOTA (**a**), DetA (**b**) and AssA (**c**). The larger the average score is, the less difficult the category is

pared with pedestrians, animal detection is more challenging and the usage of two-stage detectors may be more suitable.

In addition, we see some interesting findings on AnimalTrack. For example, SORT and IoUTrack are two simple trackers and outperformed by many recent approaches on pedestrian tracking benchmarks. However, we observe that, despite simplicity, these two trackers works surprisingly well on AnimalTrack. SORT and IOUTrack achieve 42.8 and 41.6% HOTA score, respectively, which surpass many recent state-of-the-arts such as JDE, FairMOT, and CTracker. This observation demonstrates that more efforts and attentions should be devoted and paid to the problem of multi-animal tracking.

Besides quantitatively evaluating and comparing different MOT approaches, we further show the qualitative results of different trackers. Due to limited space, we only demonstrate the qualitative results of top five trackers based on HOTA as in Fig. 5.

4.3.2 Difficulty Comparison of Categories

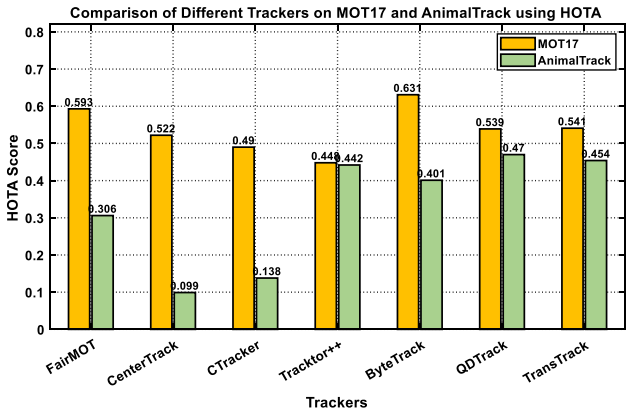
We analyze the difficulty of different animal categories in AnimalTrack. In specific, we simply average the scores of all evaluated trackers on one category to obtain the score for this category. Figure 6 shows the comparison. In Fig. 6, the larger the average score is, and the less difficult the category is. From Fig. 6, we can see that, overall, the category of *Horse* is the easiest to track while the class of *Goose* is the most difficult to track based on the average HOTA score (see Fig. 6a). We argue that *Goose* is the hardest because the gooses may have the most complex motion patters, which results in difficulties for detection (see average DetA score in Fig. 6b) and association (see average AssA score Fig. 6c). It is worth noting that, although *Goose* is easier than *Dolphin* to detect, it is much more difficult to associate. As a consequence, *Goose* is harder than *Dolphin* to track. By conducting this hardness analysis, we hope that it can guide researchers to pay more attention to the difficult categories.

4.3.3 Comparison of MOT17 and AnimalTrack

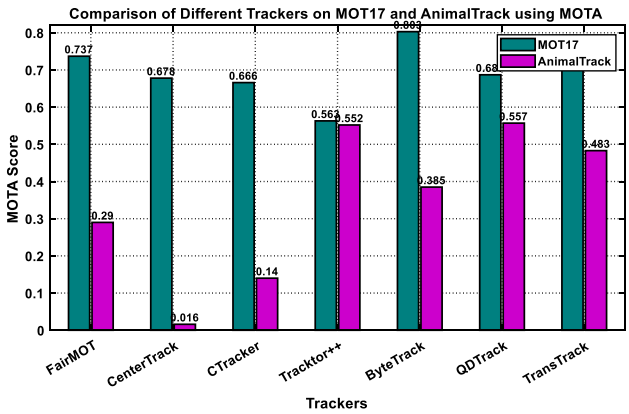
Currently, one of the main focuses in MOT community is to track pedestrians. Different from pedestrian tracking, animal tracking is more challenging because of uniform appearance of animals. In order to verify this, we compare the performance of existing state-of-the-art tracking algorithms on the popular MOT17 and the proposed AnimalTrack. Notice that, we only compare the trackers whose HOTA, MOTA and IDF1 scores are available on both MOT17 and AnimalTrack. Figure 7 displays the comparison results of these trackers.

From Fig. 7a, we can see that the best two performing trackers on MOT17 are ByteTrack and FairMOT that achieves 63.0 and 59.3% HOTA scores. Despite this, these two trackers degrade significantly when tracking animals on AnimalTrack. Specifically, their HOTA scores decrease from 63.1 to 40.1% and from 59.3 to 30.6%, showing absolute perform drops of 23.0 and 28.7%, respectively. Tracktor++ slightly performs worse on AnimalTrack than MOT17. This tracker utilizes a strong detection for tracking and shows competitive performance. Although QDTrack achieves the best HOTA result, its performs degrades on AnimalTrack compared to that on MOT17, which evidences again the challenge and difficulty we face in handling animal tracking. It is worth noting that, CenterTrack has the largest performance drop on AnimalTrack. We have carefully inspected the official implementation to ensure its correction for evaluation. After taking a close look at the implementation, we find that the features extracted in CenterTrack are not suitable for animal tracking, resulting in poor performance.

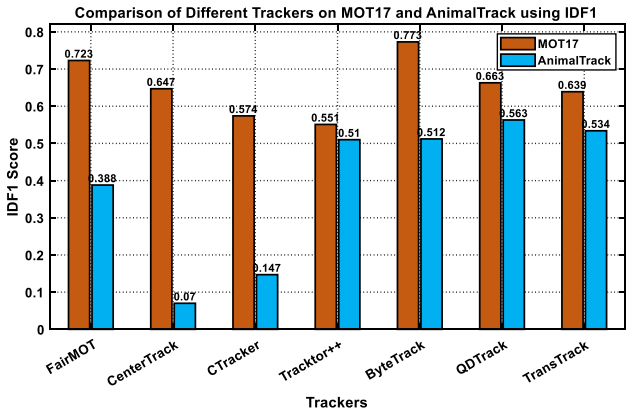
In addition to overall comparison using HOTA, we compare the MOTA score. From Fig. 7b, we can observe that the best two trackers on MOT17 are ByteTrack and TransTrack with 80.3 and 74.5% MOTA scores, respectively. Nevertheless, when tracking animals on AnimalTrack, their MOTA scores are decreased to 37.9% (42.4% absolute performance drop) and 48.3% (26.2% absolute performance drop), respectively, which shows that the animal detection is more challenging compared to human detection. Besides



(a) Comparison of MOT17 and AnimalTrack on HOTA



(b) Comparison of MOT17 and AnimalTrack on MOTA



(c) Comparison of MOT17 and AnimalTrack on IDF1

Fig. 7 Comparison of different trackers on pedestrian tracking benchmark MOT17 and the proposed AnimalTrack in terms of HOTA (**a**), MOTA (**b**) and IDF1 (**c**). We note that, compared to MOT17, all trackers become degenerated on all metrics on AnimalTrack, which shows that multi-animal tracking is more challenging than pedestrian tracking and there is a long way for improving animal tracking

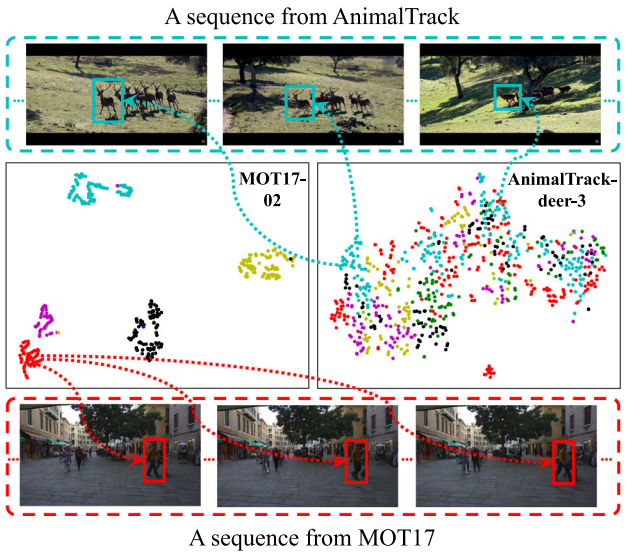


Fig. 8 Visualization and comparison of appearance features for re-identification between pedestrians and animals using t-SNE (Van der Maaten & Hinton, 2008). The same target object is represented as dots with the same color. We choose the first 30 target objects in the first 200 frames for visualization. We can clearly see that the appearance features of animals are more difficult to distinguish compared to pedestrian appearance features, resulting in new challenge for animal tracking (Color figure online)

the best two trackers on MOT17, other approaches become degenerated on AnimalTrack, which further reveals the general difficulty of detection on AnimalTrack. We notice that Tracktor++ perform consistently on both AnimalTrack than MOT17 (55.2% v.s. 56.3%). We argue that this is attributed to its powerful regressor in detection.

Moreover, we also demonstrate the comparison of IDF1 score of each tracker on MOT17 and AnimalTrack in Fig. 7c. As shown, we find that the best two trackers on MOT17 are ByteTrack and FairMOT with 77.3 and 72.3% IDF1 scores. Compared to their performance on AnimalTrack with 51.0 and 38.3% IDF1 scores, the absolute performance drops are 22.3 and 33.5%, respectively, which highlights the severe challenge in associating animals with uniform appearances. Furthermore, in addition to these two trackers, all other trackers including the best performing tracker QDTrack on AnimalTrack are actually greatly degenerated in IDF1 score, demonstrating more efforts required for solving association in animal tracking.

To further compare pedestrian and animal tracking, we analyze the appearance similarities of different pedestrians and animals on MOT17 and AnimalTrack. In particular, we train two re-identification networks with identical architectures on MOT17 and AnimalTrack, respectively. Afterwards, we extract the features of pedestrians and animals and adopt t-SNE (Van der Maaten & Hinton, 2008) to visualize these features. Figure 8 shows the visualization of appearance fea-

Table 5 Analysis on different association strategies

	Association	HOTA (%)	MOTA (%)	IDF1 (%)
❶	IOUTrack	41.6	55.7	45.7
❷	SORT	42.8	55.6	49.2
❸	DeepSORT	38.2	52.0	44.2
❹	ByteTrack	36.3	37.1	47.0
❺	QDTrack	47.0	55.7	56.1

The detection is provided by Faster R-CNN (Ren et al., 2015)

tures of pedestrians and animals. From Fig. 8, we can clearly observe that the features of animals are more complex and indistinguishable because highly similar appearances of animals compared to pedestrian appearances.

From the extensive quantitative and qualitative analysis above, we can see that tracking animals is more challenging and difficult than tracking pedestrians. Despite rapid progress on pedestrian tracking, there is a long way for improving animal tracking.

4.3.4 Analysis on Association Strategy

Association is a core component in existing MOT algorithms. In order to analyze and compare different association strategies, we conduct an independent experiment. Specifically, we adopt the classic and powerful detector Faster R-CNN (Ren et al., 2015) to provide detection results on AnimalTrack. Based on the detection results, we perform analysis on four different association strategies.

Table 5 demonstrates the comparison results. From Table 5, we can observe that QDTrack (see ❺) obtains the best performance with 47.0% HOTA score compared to trackers using other association methods, which shows that the quasi-dense matching mechanism for association is robust in dealing with animal targets with similar appearances by considering more possible regions of box examples and hard negatives. SORT (see ❶) and IOUTrack (see ❷) simply use motion information instead of appearance to perform association but achieves the second and the third best results with 42.8 and 41.6% HOTA scores. This shows that taking into consideration the motion cues in videos is beneficial for distinguishing targets with uniform appearances. Compared to SORT, DeepSORT (see ❸) adopts target appearance information for association but the performance is degraded, which once again evidences that appearance should be carefully designed when applied for associating animals. ByteTrack (see ❹) is a recently proposed approach and demonstrates state-of-the-art performance on multiple pedestrian and vehicle tracking benchmarks. The main success on these benchmarks comes from its association on all detected boxes. However, because animals have uniform appearances and it is hard to leverage their appearance information as in pedestrian or vehicles to distinguish different targets. More efforts are desired for designing appropriate association for animal targets.

4.3.5 Detection on AnimalTrack

Object detection has been a crucial component for multi-object tracking. Because of this, we have conducted an experiment with Faster R-CNN (Ren et al., 2015) using ResNet-101 (He et al., 2016) to demonstrate its detection capacity on our AnimalTrack. The reason to choose Faster R-CNN is because it is one of the most classic and popular detection frameworks and used in many multi-object tracking algorithms.

Following MS COCO (Lin et al., 2014), we adopt average precision (AP), AP_{50} and AP_{75} for detection evaluation. Definitions of these metrics can be found in Lin et al. (2014). Table 6 reports the overall and per-category detection results. From Table 6, we can see that the overall AP, AP_{50} and AP_{75} scores are 16.1, 34.4 and 13.8%, respectively. Compared to the performance of Faster R-CNN for generic object detection, there is still a large room for future improvements for animal detection on AnimalTrack.

5 Discussion

5.1 Discussion on Evaluation Metric

Evaluation metric is crucial in assessing and comparing different tracking algorithms. In this work, we leverage the common MOT metrics (see Sect. 4.1) for evaluation. However, these metrics may neglect a fact that a video may consist of too many simple tracking scenes during evaluation, which could impact the fairness in evaluating the abilities of trackers in handling hard tracking scenes. In fact, this issue does not only appear in multi-object tracking, but also in many other

Table 6 Overall and per-category detection results of Faster R-CNN (Ren et al., 2015) on AnimalTrack

	All	Chicken	Deer	Dolphin	Duck	Goose	Horse	Penguin	Pig	Rabbit	Zebra
AP (%)	16.1	25.4	2.5	13.4	49.5	5.9	16.3	16.7	12.6	6.9	11.9
AP_{50} (%)	34.4	51.3	5.2	33.1	81.5	21.8	34.7	35.9	35.4	14.0	31.6
AP_{75} (%)	13.8	23.7	2.3	8.6	56.0	0.9	13.4	12.8	6.8	5.9	7.9

Table 7 Statistics on the major animal motions

Motion	Animal category
Eat	Chicken, Duck, Horse, Pig, Zebra,
Flap	Chicken, Duck, Goose, Penguin
Walk	Chicken, Deer, Duck, Goose, Horse, Penguin, Pig, Rabbit, Zebra
Run	Deer, Horse, Pig, Rabbit
Flight	Duck, Goose
Fly	Duck, Goose
Swim	Dolphin, Duck, Goose, Penguin, Zebra
Slide	Penguin

tasks such as single-object tracking. In order to mitigate this problem, a potential solution is to provide finer annotation for the dataset for designing new metrics. For example, a group of experts (e.g., three PhD students working in related field) could offer extra weight information regarding the difficulty of scenes in each frame. The larger the difficulty of the scene is for tracking, the higher the weight is, otherwise the lower the weight is. With the weights for different difficulty degrees available, we can then design difficulty-aware metrics to improve existing measurements by paying more attention to hard tracking scenes, e.g., assigning more weight to difficult frames when computing the overall performance. In addition to the difficulty-aware overall performance, we can respectively compare different algorithms under the simple and the hard frames as we know which frames are simple and difficult, which enables in-depth analysis for different scenes. However, currently this is beyond the goal of this work. We leave it as our future work to explore more fair metrics for evaluation.

5.2 Discussion on Animal Motions

One of the reasons why tracking animals is challenging is because of their diverse motion patterns. In order to allow readers better understand animal motions in our AnimalTrack, we provide a summary as in Table 7. From Table 7, we can see that there exist eight major animal motions in AnimalTrack. Compared to existing pedestrian tracking benchmarks, the motion patterns of animals are more diverse, which results in difficulty for tracking.

6 Conclusion

In this paper, we introduce AnimalTrack, a high-quality benchmark for multi-animal tracking. Specifically, AnimalTrack consists of 58 video sequences that are selected from 10 common animal categories. To the best of our knowledge, AnimalTrack is by far the *first* and also the *largest* dataset dedicated to multi-animal tracking. By constructing AnimalTrack, we hope to provide a platform for facilitating research of MOT on animals. In addition, to provide future comparison on AnimalTrack, we extensively assess 14 popular MOT approaches with in-depth analysis. The evaluation results show that more efforts are desired for improving MAT. Furthermore, we independently study the association component for multi-animal tracking and hope that this can provide some guidance for choosing appropriate baseline for target association. Overall, we expect our dataset, along with evaluation results and our analysis, to inspire more research on multiple animal tracking using computer vision techniques.

Acknowledgements Libo Zhang was supported by the Key Research Program of Frontier Sciences, CAS, Grant No. ZDBSLY- JSC038, CAAI-Huawei MindSpore Open Fund and Youth Innovation Promotion Association, CAS (2020111).

References

Bai, H., Cheng, W., Chu, P., Liu, J., Zhang, K., & Ling, H. (2021). Gmot-40: A benchmark for generic multiple object tracking. In *IEEE international conference on computer vision and pattern recognition conference (CVPR)*.

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), 1–12.

Bergmann, P., Meinhardt, T., & Leal-Taixe, L. (2019). Tracking without bells and whistles. In *IEEE international conference on computer vision (ICCV)*.

Bernardin, K., & Stiefelhagen, R. (2008). Evaluating multiple object tracking performance: The clear mot metrics. *EURASIP Journal on Image and Video Processing, 2008*, 1–10.

Betke, M., Hirsh, D. E., Bagchi, A., Hristov, N. I., Makris, N. C., & Kunz, T. H. (2007). Tracking large variable numbers of objects in clutter. In *IEEE international conference on computer vision and pattern recognition conference (CVPR)*.

Bewley, A., Ge, Z., Ott, L., Ramos, F., & Upcroft, B. (2016). Simple online and realtime tracking. In *IEEE international conference in image processing (ICIP)*.

Bochinski, E., Eiselein, V., & Sikora, T. (2017). High-speed tracking-by-detection without using image information. In *IEEE international conference on advanced video and signal-based surveillance (AVSS)*.

Bozek, K., Hebert, L., Mikheyev, A. S., & Stephens, G. J. (2018). Towards dense object tracking in a 2d honeybee hive. In *IEEE international conference on computer vision and pattern recognition conference (CVPR)*.

Brasó, G., & Leal-Taixé, L. (2020). Learning a neural solver for multiple object tracking. In *IEEE international conference on computer vision and pattern recognition conference (CVPR)*.

Cao, J., Tang, H., Fang, H. S., Shen, X., Lu, C., & Tai, Y. W. (2019). Cross-domain adaptation for animal pose estimation. In *IEEE international conference on computer vision (ICCV)*.

Chu, P., Fan, H., Tan, C. C., & Ling, H. (2019). Online multi-object tracking with instance-aware tracker and dynamic model refresh-

ment. In *IEEE winter conference on applications of computer vision (WACV)*.
Chu, P., & Ling, H. (2019). Famnet: Joint learning of feature, affinity and multi-dimensional assignment for online multiple object tracking. In *IEEE international conference on computer vision (ICCV)*.
Ciaparrone, G., Sánchez, F. L., Tabik, S., Troiano, L., Tagliaferri, R., & Herrera, F. (2020). Deep learning in video multi-object tracking: A survey. *Neurocomputing, 381*, 61–88.
Dai, P., Weng, R., Choi, W., Zhang, C., He, Z., & Ding, W. (2021). Learning a proposal classifier for multiple object tracking. In *IEEE international conference on computer vision and pattern recognition conference (CVPR)*.
Dave, A., Khurana, T., Tokmakov, P., Schmid, C., & Ramanan, D. (2020). Tao: A large-scale benchmark for tracking any object. In *European conference on computer vision (ECCV)*.
Dehghan, A., Tian, Y., Torr, P. H., & Shah, M. (2015). Target identity-aware network flow for online multiple target tracking. In *IEEE international conference on computer vision and pattern recognition conference (CVPR)*.
Dendorfer, P., Osep, A., Milan, A., Schindler, K., Cremers, D., Reid, I., Roth, S., & Leal-Taixé, L. (2021). Motchallenge: A benchmark for single-camera multiple target tracking. *International Journal of Computer Vision, 129*(4), 845–881.
Dendorfer, P., Rezatofighi, H., Milan, A., Shi, J., Cremers, D., Reid, I., Roth, S., Schindler, K., & Leal-Taixé, L. (2020). Mot20: A benchmark for multi object tracking in crowded scenes. arXiv:2003.09003.
Du, D., Qi, Y., Yu, H., Yang, Y., Duan, K., Li, G., Zhang, W., Huang, Q., & Tian, Q. (2018). The unmanned aerial vehicle benchmark: Object detection and tracking. In *European conference on computer vision (ECCV)*.
Emami, P., Pardalos, P. M., Elefteriadou, L., & Ranka, S. (2020). Machine learning methods for data association in multi-object tracking. *ACM Computing Surveys, 53*(4), 1–34.
Ferryman, J., & Shahrokni, A. (2009). Pets2009: Dataset and challenge. In *PETS Workshop*.
Geiger, A., Lenz, P., & Urtasun, R. (2012). Are we ready for autonomous driving? the kitti vision benchmark suite. In *IEEE International conference on computer vision and pattern recognition conference (CVPR)*.
Guo, S., Wang, J., Wang, X., & Tao, D. (2021). Online multiple object tracking with cross-task synergy. In *IEEE international conference on computer vision and pattern recognition conference (CVPR)*.
He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In IEEE conference on computer vision and pattern recognition (CVPR).
Iwashita, Y., Takamine, A., Kurazume, R., & Ryoo, M. S. (2014). First-person animal activity recognition from egocentric videos. In *International conference on pattern recognition (ICPR)*.
Khan, Z., Balch, T., & Dellaert, F. (2004). An MCMC-based particle filter for tracking multiple interacting targets. In *European conference on computer vision (ECCV)*.
Leal-Taixé, L., Milan, A., Reid, I., Roth, S., & Schindler, K. (2015). Motchallenge 2015: Towards a benchmark for multi-target tracking. arXiv:1504.01942.
Li, S., Li, J., Tang, H., Qian, R., Lin, W. (2019). ATRW: A benchmark for amur tiger re-identification in the wild. In *ACM Multimedia (MM)*.
Liang, C., Zhang, Z., Zhou, X., Li, B., Lu, Y., & Hu, W. (2022). One more check: Making" fake background" be tracked again. In *Association for the advancement of artificial intelligence (AAAI)*.
Lin, T. Y., Goyal, P., Girshick, R., He, K., & Dollár, P. (2017). Focal loss for dense object detection. In *IEEE international conference on computer vision (ICCV)*.
Lin, T. Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., & Zitnick, C. L. (2014). Microsoft COCO: Common objects in context. In *European conference on computer vision (ECCV)*.
Lu, Z., Rathod, V., Votel, R., & Huang, J. (2020). Retinatrack: Online single stage joint detection and tracking. In *IEEE international conference on computer vision and pattern recognition conference (CVPR)*.
Luiten, J., Osep, A., Dendorfer, P., Torr, P., Geiger, A., Leal-Taixé, L., & Leibe, B. (2021). Hota: A higher order metric for evaluating multi-object tracking. *International Journal of Computer Vision, 129*(2), 548–578.
Luo, W., Xing, J., Milan, A., Zhang, X., Liu, W., & Kim, T. K. (2021). Multiple object tracking: A literature review. *Artificial Intelligence, 293*, 103448.
Mathis, A., Biasi, T., Schneider, S., Yuksekgonul, M., Rogers, B., Bethge, M., & Mathis, M. W. (2021). Pretraining boosts out-of-domain robustness for pose estimation. In *IEEE winter conference on applications of computer vision (WACV)*.
Meinhardt, T., Kirillov, A., Leal-Taixe, L., & Feichtenhofer, C. (2022). Trackformer: Multi-object tracking with transformers. In *IEEE international conference on computer vision and pattern recognition conference (CVPR)*.
Milan, A., Leal-Taixé, L., Reid, I., Roth, S., & Schindler, K. (2016). Mot16: A benchmark for multi-object tracking. arXiv:1603.00831.
Pang, J., Qiu, L., Li, X., Chen, H., Li, Q., Darrell, T., & Yu, F. (2021). Quasi-dense similarity learning for multiple object tracking. In *IEEE international conference on computer vision and pattern recognition conference (CVPR)*.
Parham, J., Stewart, C., Crall, J., Rubenstein, D., Holmberg, J., & Berger-Wolf, T. (2018). An animal detection pipeline for identification. In *IEEE winter conference on applications of computer vision (WACV)*.
Peng, J., Wang, C., Wan, F., Wu, Y., Wang, Y., Tai, Y., Wang, C., Li, J., Huang, F., & Fu, Y. (2020). Chained-tracker: Chaining paired attentive regression results for end-to-end joint multiple-object detection and tracking. In *European conference on computer vision (ECCV)*.
Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster R-CNN: Towards real-time object detection with region proposal networks. In *Conference on neural information processing systems (NIPS)*.
Ristani, E., Solera, F., Zou, R., Cucchiara, R., & Tomasi, C. (2016). Performance measures and a data set for multi-target, multi-camera tracking. In *European conference on computer vision (ECCV) workshop*.
Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al. (2015). Imagenet large scale visual recognition challenge. *International Journal of Computer Vision, 115*(3), 211–252.
Schulter, S., Vernaza, P., Choi, W., & Chandraker, M. (2017). Deep network flow for multi-object tracking. In *IEEE international conference on computer vision and pattern recognition conference (CVPR)*.
Shuai, B., Berneshawi, A., Li, X., Modolo, D., & Tighe, J. (2021). Siammot: Siamese multi-object tracking. In *IEEE international conference on computer vision and pattern recognition conference (CVPR)*.
Sun, P., Cao, J., Jiang, Y., Yuan, Z., Bai, S., Kitani, K., & Luo, P. (2022). Dancetrack: Multi-object tracking in uniform appearance and diverse motion. In *IEEE international conference on computer vision and pattern recognition conference (CVPR)*.
Sun, P., Cao, J., Jiang, Y., Zhang, R., Xie, E., Yuan, Z., Wang, C., & Luo, P. (2020). Transtrack: Multiple object tracking with transformer. arXiv:2012.15460.
Tang, S., Andriluka, M., Andres, B., & Schiele, B. (2017). Multiple people tracking by lifted multicut and person re-identification. In *IEEE international conference on computer vision and pattern recognition conference (CVPR)*.

Van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research, 9*(11).

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In *Conference on neural information processing systems (NIPS)*.

Voigtlaender, P., Krause, M., Osep, A., Luiten, J., Sekar, B. B. G., Geiger, A., & Leibe, B. (2019). Mots: Multi-object tracking and segmentation. In *IEEE international conference on computer vision and pattern recognition conference (CVPR)*.

Wang, Z., Zheng, L., Liu, Y., Li, Y., & Wang, S. (2020). Towards real-time multi-object tracking. In *European conference on computer vision (ECCV)*.

Wen, L., Du, D., Cai, Z., Lei, Z., Chang, M. C., Qi, H., Lim, J., Yang, M. H., & Lyu, S. (2020). UA-DETRAC: A new benchmark and protocol for multi-object detection and tracking. *Computer Vision and Image Understanding, 193*, 102907.

Wojke, N., Bewley, A., & Paulus, D. (2017). Simple online and realtime tracking with a deep association metric. In *IEEE international conference in image processing (ICIP)*.

Xu, J., Cao, Y., Zhang, Z., & Hu, H. (2019). Spatial-temporal relation networks for multi-object tracking. In *IEEE international conference on computer vision (ICCV)*.

Xu, Y., Osep, A., Ban, Y., Horaud, R., Leal-Taixé, L., & Alameda-Pineda, X. (2020). How to train your deep multi-object tracker. In *IEEE international conference on computer vision and pattern recognition conference (CVPR)*.

Yang, L., Fan, Y., & Xu, N. (2019). Video instance segmentation. In *IEEE international conference on computer vision (ICCV)*.

Yin, J., Wang, W., Meng, Q., Yang, R., & Shen, J. (2020). A unified object motion and affinity model for online multi-object tracking. In *IEEE international conference on computer vision and pattern recognition conference (CVPR)*.

Yu, F., Chen, H., Wang, X., Xian, W., Chen, Y., Liu, F., Madhavan, V., & Darrell, T. (2020). Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In *IEEE international conference on computer vision and pattern recognition conference (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. In *Conference and workshop on neural information processing systems (NeurIPS)—track on datasets and benchmarks*.

Zhang, Y., Sun, P., Jiang, Y., Yu, D., Yuan, Z., Luo, P., Liu, W., & Wang, X. (2021a). Bytetrack: Multi-object tracking by associating every detection box. arXiv:2110.06864.

Zhang, Y., Wang, C., Wang, X., Zeng, W., & Liu, W. (2021b). Fairmot: On the fairness of detection and re-identification in multiple object tracking. *International Journal of Computer Vision, 129*(11), 3069–3087.

Zhou, X., Koltun, V., & Krähenbühl, P. (2020). Tracking objects as points. In *European conference on computer vision (ECCV)*.

Zhu, J., Yang, H., Liu, N., Kim, M., Zhang, W., & Yang, M. H. (2018). Online multi-object tracking with dual matching attention networks. In *European conference on computer vision (ECCV)*.

Zhu, P., Wen, L., Du, D., Bian, X., Fan, H., Hu, Q., & Ling, H. (2022). Detection and tracking meet drones challenge. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(11), 7380–7399.