Tracking Different Ant Species: An Unsupervised Domain Adaptation Framework and a Dataset for Multi-object Tracking

Chamath Abeysinghe[1] , **Chris Reid**[2] , **Hamid Rezatofighi**[1] and **Bernd Meyer**[1]

[1]Dept. of Data Science and Artificial Intelligence, Monash University

[2]School of Natural Sciences, Macquarie University

{chamath.abeysinghe, hamid.rezatofighi, bernd.meyer}@monash.edu, chris.reid@mq.edu.au

Abstract

Tracking individuals is a vital part of many experiments conducted to understand collective behaviour. Ants are the paradigmatic model system for such experiments but their lack of individually distinguishing visual features and their high colony densities make it extremely difficult to perform reliable tracking automatically. Additionally, the wide diversity of their species' appearances makes a generalized approach even harder. In this paper, we propose a data-driven multi-object tracker that, for the first time, employs domain adaptation to achieve the required generalisation. This approach is built upon a joint-detection-and-tracking framework that is extended by a set of domain discriminator modules integrating an adversarial training strategy in addition to the tracking loss. In addition to this novel domain-adaptive tracking framework, we present a new dataset and a benchmark for the ant tracking problem. The dataset contains 57 video sequences with full trajectory annotation, including 30k frames captured from two different ant species moving on different background patterns. It comprises 33 and 24 sequences for source and target domains, respectively. We compare our proposed framework against other domain-adaptive and non-domain-adaptive multi-object tracking baselines using this dataset and show that incorporating domain adaptation at multiple levels of the tracking pipeline yields significant improvements. The code and the dataset are available at https://github.com/chamathabeysinghe/da-tracker.

1 Introduction

Many biologists and ecologists are interested in understanding social insect behaviours, in particular, that of ants to gain insights into how social systems collaborate in nature. This important research is unfortunately slowed down by the difficulty of automatically tracking ants of many different species in very diverse experimental environments.

Compared to other multi-object tracking (MOT) problems in computer vision, ant tracking imposes unique challenges; individual ants are visually extremely similar and move
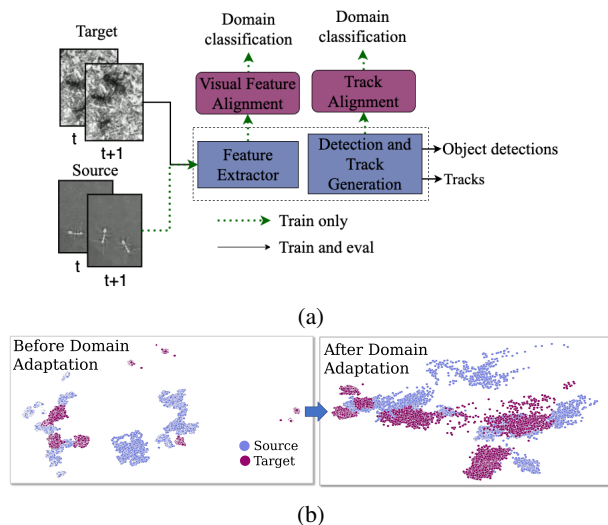

Figure 1: (a) A schematic of our proposed unsupervised domain adaptation framework for multi-object tracking, (b) A visualization, representing the source and target feature distributions before and after applying our domain adaptation strategy.

in highly crowded environments with complex interactions. This leads to a significant level of occlusions, overlapping movements *etc*. Furthermore, ants comprise a great variety of species with broadly differing appearances between species. The main problem for data-driven approaches arising from this diversity is the transferability between species. State-of-the-art data-driven tracking methods such as [Meinhardt *et al.*, 2022; Sun *et al.*, 2020] trained on a dataset of a specific ant species and environment do not perform well on datasets with different ant species and environments, as shown in the experiment section of this paper. Consequently, the most popular approaches used for ant tracking rely on model-based detection and tracking techniques [Pérez-Escudero *et al.*, 2014; Naiser *et al.*, 2018]. To the best of our knowledge, no species-diverse large-scale dataset exists that can be used for training and comparing state-of-the-art data-driven detection and tracking approaches, e.g., [Ren *et al.*, 2015; Meinhardt *et al.*, 2022; Sun *et al.*, 2020].

In this paper, we propose a data-driven domain adaptive multi-object tracking framework, named DA-Tracker, which

is capable of transferring knowledge from training with one ant species (source) to another ant species in a different environment (target). To the best of our knowledge, our proposed approach is the first MOT framework to tackle this unsupervised domain adaptation problem in an end-to-end trainable manner. Fig. 1 shows a high-level design of the proposed framework and its performance in unifying the source and target feature representation.

Our proposed end-to-end multi-object tracking approach builds upon [Meinhardt *et al.*, 2022] and extends this to perform domain adaptation. Our domain adaptation modules are based on adversarial training strategies that generate similarity between the encoded input and decoded output representations from source and target data by enforcing highly overlapping feature distributions. To this end, domain adaptation is applied to image level features and track level information separately. To train DA-Tracker, we use two different categories of losses: 1) supervised losses, used for source domain data only, and 2) discriminator losses, used for both source and target domain data. The multi-object tracking module learns to localize objects and generate tracks using the supervised losses from the annotated outputs available in the source data. The three discriminators, connected to the intermediate encoding and decoding layers in the tracking pipeline, ensure that the tracking module generates highly similar feature distributions for source and target domains data across the input and output representations.

As a second contribution, we introduce and make publicly available a new large-scale dataset for ant tracking with unsupervised domain adaptation in realistic experimental setups. We use this to evaluate our framework in this paper. Our dataset includes 57 video sequences (more than $30K$ image frames) captured in a laboratory environment from two ant species *Oecophylla smaragdina* (weaver ants) and *Camponotus aeneopilosus* (carpenter ants) with very different appearance and behaviour moving on different background structures (ranging from plain to grassy backgrounds) under various ant population density, lighting and the camera zoom conditions (Fig. 3). The dataset is provided with about $700K$ and $2K$ high-quality bounding boxes and tracks, respectively. To establish a standard benchmark, we divide both the target and source domain data into three splits: training, validation and test. The training split comprises about 50% of data; test and validation about 25% each. To evaluate our method, we adopt standard multi-object tracking metrics such as MOT-Clear [Bernardin and Stiefelhagen, 2008], IDF [Ristani *et al.*, 2016] and HOTA [Luiten *et al.*, 2021].

In summary, our main contributions are to: (1) propose an unsupervised domain adaptation method for multi-object tracking in an end-to-end trainable framework, (2) introduce a large-scale ant dataset and benchmark for unsupervised domain adaptation in multi-object tracking, (3) comprehensively evaluate and compare our framework against the state-of-the-art detection based and data-driven MOT approaches.

2 Related Work

Multi-object tracking in computer vision: In multi-object tracking problems, data-driven approaches have shown state-of-the-art performance on many publicly available MOT datasets such as [Milan *et al.*, 2016; Geiger *et al.*, 2012]. A main paradigm in multi-object tracking is tracking-by-detection where the problem is dived into a two-step process: (1) object detection and (2) track generation. In the first step, a deep learning-based object detector, e.g., [Ren *et al.*, 2015; He *et al.*, 2017; Carion *et al.*, 2020], is used to localize all the objects of interest in all the frame sequences. Next, a tracking technique is applied to generate tracks, e.g., a filtering-based framework [Bewley *et al.*, 2016; Wojke *et al.*, 2017; Rezatofighi *et al.*, 2015], an optimisation-based technique [Schulter *et al.*, 2017; Lan *et al.*, 2016] or another approach [Oh *et al.*, 2009; Choi, 2015] using a similarity/distance measure between detection and hypothetical tracks, that is generally based on appearance information [Wojke *et al.*, 2017; Liu *et al.*, 2020], motion information [Leal-Taixé *et al.*, 2011; Saleh *et al.*, 2021] or both [Sadeghian *et al.*, 2017; Bae and Yoon, 2014].

A different multi-object tracking paradigm that has recently become popular is joint-detection-and-tracking [Meinhardt *et al.*, 2022; Sun *et al.*, 2020; Bergmann *et al.*, 2019; Feichtenhofer *et al.*, 2017]. Joint-detection-and-tracking performs object detection and tracking simultaneously in a single step. Integrating everything into a single step allows us to efficiently exchange information between object detection and tracking. Our proposed model is built on top of this category of MOT approaches. In this paper, we use Trackformer [Meinhardt *et al.*, 2022] as an end-to-end trainable MOT framework as our baseline. Trackformer [Meinhardt *et al.*, 2022] is a joint-detection-and-tracking method inspired by the transformer object detector, DETR [Carion *et al.*, 2020]. Trackformer has two object detectors working on consecutive frames. It learns to initialize new tracks (i.e. new detected objects), and terminate the exiting tracks or associate them to the next frame by predicting their next state (i.e. bounding box and confidence scores) in an auto-regressive scheme between two consecutive frames.

Ant tracking: Most existing ant tracking frameworks follow model-based approaches [Pérez-Escudero *et al.*, 2014; Naiser *et al.*, 2018]. A popular monitoring tool used for ant tracking is iDTracker. As originally described in [Pérez-Escudero *et al.*, 2014], this tool generates object detections using colour intensity thresholds and assigns a unique ID using a classification network. The original approach is model-based and does not need training data. It can efficiently work under simple tracking scenarios. However, it fails to track reliably in highly crowded environments. We note that a more recent extension, termed iDTracker.ai, uses deep learning (CNNs) for individual identification [Romero-Ferrero *et al.*, 2019]. This approach relies on individuals being sufficiently distinguishable and no results for ants have been published in [Romero-Ferrero *et al.*, 2019]. Another tracking-by-detection approach using deep learning for object detection (Mask R-CNN [He *et al.*, 2017]) is described in [Imirzian *et al.*, 2019]. This approach uses Earth Mover's Distance (EMD) [Chen *et al.*, 2014] to link detections into tracks. The performance of these tracking-by-detection approaches heavily relies on the supervised object detectors which need to be

re-trained for every experimental setting.

Unsupervised domain adaptation: Unsupervised domain adaptation is a process of generalizing a model to work on other input data than the labeled training data. There are many approaches to unsupervised domain adaptation, e.g. adversarial-based methods [Saito *et al.*, 2019; Chen *et al.*, 2018], discrepancy-based methods [Yan *et al.*, 2017; Long *et al.*, 2017], and reconstruction-based methods [Bousmalis *et al.*, 2016; Ghifary *et al.*, 2015]. Domain adaptation is a very well-studied problem in a few machine learning/computer vision tasks such as image classification [Ganin and Lempitsky, 2015; Jiang *et al.*, 2020] and object detection [Saito *et al.*, 2019; Chen *et al.*, 2018; Zhang *et al.*, 2021; Huang *et al.*, 2022]. However, it has been barely extended to higher-level tasks, such as multi-object tracking, due to the complexity of such problem.

Many domain adaptation methods [Chen *et al.*, 2018; Saito *et al.*, 2019; Zhang *et al.*, 2021; Huang *et al.*, 2022] in object detection are adversarial-based methods that train image domain classifiers alongside the object detector. In adversarial-based domain adaptation, model generalization is achieved by a discriminator forcing the feature extractor to translate feature representations for different domain data into one common distribution. [Chen *et al.*, 2018] have proposed a Faster R-CNN with an adversarial domain adaptive approach. In this method, a Faster R-CNN connects to domain classification layers at two levels: the image level and the instance level, and translates features into a common distribution. [Saito *et al.*, 2019] propose improvements to this DA-Faster R-CNN method by introducing a new loss function that estimates a cost based on the classification's difficulty.

Only a few studies of domain adaptation in the multi-object tracking exist [Gaidon and Vig, 2015]. One way to improve tracking's accuracy for an unseen domain is to use a domain-adaptive object detector followed by a model-based tracking approach. The baseline models to which we compare in this study work in this way. In contrast to this and previous work, we apply domain adaptation techniques on both object detection and track generation in a data-driven multi-object tracker.

Dataset and benchmark: Several large-scale standard datasets and benchmarks for multi-object tracking problems exist, e.g. for pedestrian/human tracking [Milan *et al.*, 2016; Kumar *et al.*, 2020; Sun *et al.*, 2022], vehicle tracking [Geiger *et al.*, 2012; Du *et al.*, 2018], and animal tracking [Ray and Stopfer, 2022]. Having these public datasets and benchmarks helps to develop state-of-the-art solutions in respective fields.

For ant tracking, we have only a few publicly available datasets [Imirzian *et al.*, 2019; Wu *et al.*, 2022]. The dataset from [Wu *et al.*, 2022] is relatively small with about $5k$ images. The dataset published in [Imirzian *et al.*, 2019] has $20K$ annotated images but only sparse object detections with 1.8 average detections per frame. Due to this sparseness we cannot observe complex motions and frequently overlapping tracks. Both of these datasets contain videos from only one ant species. Our proposed large-scale dataset has two ant species and higher density of 25.6 detections per frames allowing us to assess complex tracking scenarios. Furthermore, we establish standard benchmark tasks for this dataset.
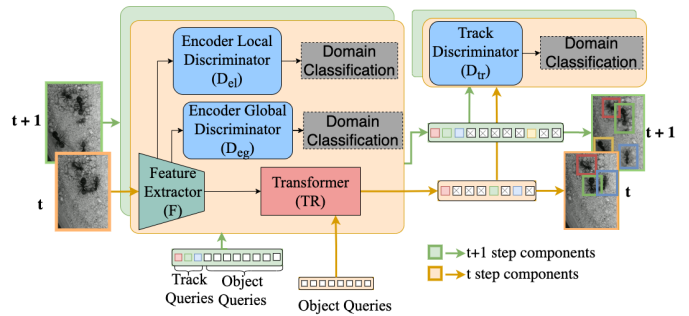

Figure 2: A schematic of our proposed network architecture. The multi-object tracking module predicts detected boxes and associate them between two frames in a unified network. To ensure our proposed framework can adapt to the target data distribution, the intermediate layer feature representations are enforced to follow a common data distribution. The conversion happens at two levels: Image encoded features(D_{eg} and D_{el}), track level (D_{tr}).

3 Framework

We present an unsupervised domain adaptation method for a multi-object tracking network that translates source and target image features into a common distribution. Mapping the target feature distribution to the source distribution, we can expect a tracker trained only on the source domain tracker to retain its performance in the target domain. Our proposed approach relies on several adversarial losses for classifying image-level and track-level features between source and target data. As depicted in Fig. 2, our adversarial approach relies on three discriminators (blue boxes): two of these operate on the encoded visual features, while the third one operates on the decoded features representing the outputs/trajectories.

3.1 Multi-Object Tracker

Our approach is based on the recently proposed end-to-end multi-object tracker, known as Trackformer [Meinhardt *et al.*, 2022]. In this framework, the tracker component uses two extended object detectors that work on consecutive frames and links the individual predictions using *track queries*. These track queries are essentially a summary of information about object detections in the previous frame that is passed to a predictor for the current frame. To train the tracker on the source domain, we use the supervised loss, $\mathcal{L}_{MOT,S}$, suggested by [Meinhardt *et al.*, 2022], which is a combination of losses for track initiation and localization for new objects and state prediction for the existing ones.

3.2 Domain Adaptation in Layers

To extend this tracking framework for unsupervised domain-adaption, we adopt the adversarial training strategy by integrating three different discriminators, trained simultaneously with the multi-object tracker. These discriminators enforce similarity between the encoded/decoded features from source and domain data. Their details are elaborated below:

Encoded visual feature alignment: Feature alignment works at frame-level feature extraction layers in the backbone F. One discriminator (D_{el} in Figure 2) operates on

local features, the other one (D_{eg}) on global features (e.g. backgrounds, scene layouts). At the global level, domain-specific attributes may vary significantly, and attempts to align them may negatively impact model performance [Saito *et al.*, 2019]. By incorporating two discriminators, the degree of feature alignment can be tailored to the specific layer.

D_{el} takes its input of height H and width W from the backbone layer (F') and produces a domain classification prediction of the same shape. A pixel level loss is estimated from this prediction for each source domain frame, x_i^S, and target domain frame, x_i^T, to establish a strong alignment cost for local features. This local feature classification cost, $\mathcal{L}_{local}$ is calculated over two consecutive time steps, $t = 1, 2$:

$$\mathcal{L}_{loc_{S,t}} = \frac{1}{n_S H W} \sum_{i=1}^{n_S} D_{el}(F'(x_{i,t}^S))^2 \qquad (1)$$

$$\mathcal{L}_{loc_{T,t}} = \frac{1}{n_T H W} \sum_{i=1}^{n_T} (1 - D_{el}(F'(x_{i,t}^T)))^2 \qquad (2)$$

$$\mathcal{L}_{local} = \frac{1}{4} \sum_{t=1,2} \mathcal{L}_{loc_{S,t}} + \mathcal{L}_{loc_{T,t}} \qquad (3)$$

D_{eg} similarly predicts the domain category at the global feature level (Equation 6). This loss, a modified version of cross-entropy loss, assigns a higher value for hard-to-classify examples [Saito *et al.*, 2019]. Equation 4 and 5 calculate this loss for source domain and target domain respectively.

$$\mathcal{L}_{gl_{S,t}} = -\frac{1}{n_S} \sum_{i=1}^{n_S} (1 - D_{eg}(F''(x_{i,t}^S))^\gamma \\ \times \log(D_{eg}(F_2(x_{i,t}^s))) \qquad (4)$$

$$\mathcal{L}_{gl_{T,t}} = -\frac{1}{n_T} \sum_{i=1}^{n_T} D_{eg}(F''(x_{i,t}^T)^\gamma \\ \times \log(1 - D_{eg}(F''(x_{i,t}^T))) \qquad (5)$$

$$\mathcal{L}_{global} = \frac{1}{4} \sum_{t=1,2} \mathcal{L}_{gl_{S,t}} + \mathcal{L}_{gl_{T,t}} \qquad (6)$$

Our ablation experiments provided in Sec. 5 show that the visual feature alignment by itself already yields a significant improvement in the tracking performance for the target domain data, compared to the supervised baseline. This performance improvement is mainly due to better object detection in the target domain data. To compensate domain shift between the outputs (tracks), we also add a track-level alignment component by an additional track discriminator.

Track level alignment: We introduce a track discriminator, D_{tr} at the final output layer of the Transformer (TR) to compensate the domain shift between the output (track) distributions. The output from the transformer, $q_{i,j}$ contains an embedding of both track queries and object queries. These are used as the initialisation for the continuation of the track. Domain alignment at the track level thus addresses track continuation and track termination.

The track discriminator, D_{tr}, ensures that the decoded (output-level) features from the source and domain data are not distinguishable by classifying their domain individually in an adversarial learning setting using the following losses:

$$\mathcal{L}_{tr_{S,t}} = \frac{1}{n_S(n_{tr} + n_{ob})} \sum_{i=1}^{n_S} \sum_{j=1}^{n_{tr}+n_{ob}} (1 - D_{tr}(q_{i,j}))^\gamma \\ \times \log(D_{tr}(q_{i,j})) \qquad (7)$$

$$\mathcal{L}_{tr_{T,t}} = \frac{1}{n_T(n_{tr} + n_{ob})} \sum_{i=1}^{n_T} \sum_{j=1}^{n_{tr}+n_{ob}} (D_{tr}(q_{i,j}))^\gamma \\ \times \log(1 - D_{tr}(q_{i,j})) \qquad (8)$$

$$\mathcal{L}_{track} = \frac{1}{4} \sum_{t=1,2} \mathcal{L}_{tr_{S,t}} + \mathcal{L}_{tr_{T,t}} \qquad (9)$$

3.3 End-to-End Trainable Domain Adaptation

Finally, Equation 10 combines the loss of all three discriminators with the supervised tracking loss, $L_{MOT,S}$. To train the model with all the losses together in an end-to-end trainable fashion we integrate a gradient reverse layer [Ganin and Lempitsky, 2015] between the discriminators and the tracker. The gradient reverse layer introduces negative feedback when the source and target domains features are easily distinguishable. The rationale is that easy discrimination can be assumed to be detrimental to the domain transfer of tracking. In this way the model can be trained for multiple objectives simultaneously.

$$\mathcal{L}_{total} = \mathcal{L}_{MOT,S} + \lambda_1 \mathcal{L}_{local} + \lambda_2 \mathcal{L}_{global} + \lambda_3 \mathcal{L}_{track} \quad (10)$$

4 Dataset and Benchmark

4.1 Annotated Dataset

Our benchmark dataset contains 57 video sequences recorded in a standard laboratory environment in the context of a real biological experiment. It comprises 33 videos in the source domain and 24 videos in the target domain for a total of 36k frames and 700k object detections. Each video contains between 10-50 ants on average. The videos were recorded with different backgrounds, zoom levels and lighting conditions. Table 1 details the dataset metrics. Our dataset comprises videos of two very different species of ants: Weaver

	Source				Target			
	Train	Val	Test	All	Train	Val	Test	All
Tot. videos	20	6	7	33	12	6	6	24
Tot. frames	12200	3440	4160	19800	8200	4100	4100	16400
Tot. detections	286K	82K	103K	471K	135K	53K	60K	249K
Tot. tracks	635	217	199	1051	455	215	203	873
Avg. track len.	450.0	379.8	518.7	448.5	297.0	249.8	298.5	285.7

Table 1: Our dataset has two parts: Source and Target. For domain adaptation applications, we assume we have labels only for source domain data during training. To establish a benchmark, we divide both source and the target data into training, validation and test splits.

(a) Highly crowded target domain sample

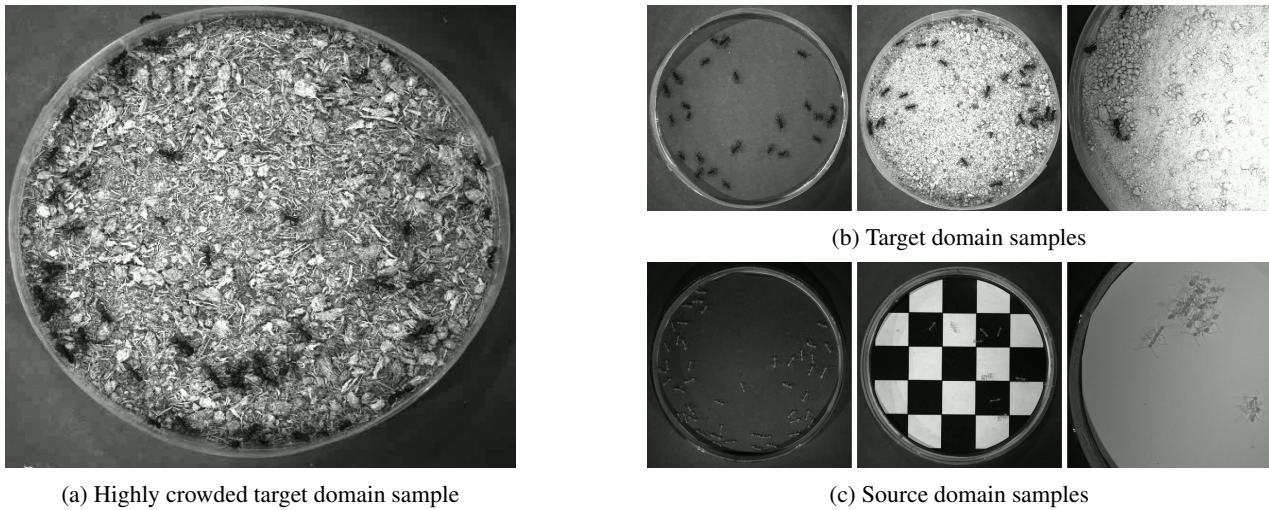(b) Target domain samples

(c) Source domain samples

Figure 3: Samples taken from our dataset to show the diversity. Dataset has two ant species divided as target domain and source domain. Target domain has four videos in four different backgrounds and source domain has 3 different backgrounds.

(a) Speed distribution

(b) Travel distance distribution

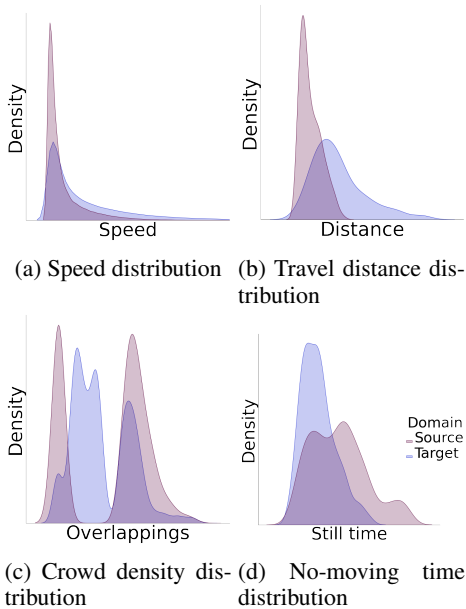(c) Crowd density distribution

(d) No-moving time distribution

Figure 4: Distribution difference in target and source domain. Here we illustrate behavioural pattern difference between two ant species.

ants (*Oecophylla smaragdina*) and Carpenter ants (*Camponotus aeneopilosus*) [Hölldobler *et al.*, 1990]. As shown in Figs. 3 and 4, these two ant species have clearly distinguishable physical appearances and behavioural patterns, making this dataset a proper practical benchmark for domain adaptation in multi-object tracking.

We produced the ground-truth track using a two-step process: (1) a simple tracking algorithm was used to produce a tentative, partial and noisy ground truth (2) all these initial annotations were manually corrected by a human expert.

For the initial phase, we first trained a Faster R-CNN detector [Ren *et al.*, 2015] using the detected ant locations by an off-the-shelf QR tracker [Boenisch *et al.*, 2018] on an exter-

nal dataset including ants tagged with micro-QR codes. We then used this trained detector on a second set of videos including untagged ants, i.e. our final benchmark dataset. Next, a basic multi-object tracking approach (Kalman filtering followed by Hungarian matching), was applied to link detections between successive frames and to produce the initial track annotations. Due to the fact that the object detector was trained for QR-tagged ants and due to the simplistic tracking linking, this "pseudo ground-truth" is obviously not useable as a real ground-truth. Thus, in the second phase, the "pseudo ground-truth" was manually and carefully corrected by human inspection to generate a fully valid ground-truth.

4.2 Benchmark

Source and target domains: The main application of our benchmark is to assess multi-object tracking frameworks in an unsupervised domain adaptation experimental setting. For this benchmark, we consider the weaver ant videos as the source domain and the carpenter ant videos as the target domain data, respectively.

Train, validation and test split: We divided both the target and source domain data into three splits: training, validation and test. The training split comprises about 50% of data; test and validation about 25% each. We tried our best to ensure the statistics of each split reflects the similar distribution in the terms of background types, zoom levels, density of ants *etc*. We will use the training and validation sets of the source (inputs and their annotations) and target domains (inputs only) for optimizing the model parameters and hyperparameter tuning, respectively. The final results are evaluated on the target domain's test split.

Evaluation metric: Tracking performances in the biological/ecological literature are reported in a variety of different ways that do not allow direct comparison. To overcome this, we adopt standards of the multi-object tracking literature: MOTA, HOTA and IDF1. In addition to these, we include other key metrics like IDSW, MT, ML in the bench-

Detector	Tracker	MOTA	IDF1	HOTA	MT	ML	Frag	FP(/frame)	ID Sw.
DETR [Carion *et al.*, 2020][†]	SORT [Bewley *et al.*, 2016]	-3.759	0.107	0.203	54	69	673	61.75	582
DETR [Carion *et al.*, 2020][†]	EMD tracker [Imirzian *et al.*, 2019]	-3.830	0.110	0.205	61	65	731	63.20	640
Transtrack [Sun *et al.*, 2020][†] (joint detection and tracking)		-3.606	0.066	0.132	60	84	3210	59.76	2546
Trackformer [Meinhardt *et al.*, 2022][†] (joint detection and tracking)		-1.02	0.187	0.262	57	118	1148	20.87	1351
DA FRCNN [Saito *et al.*, 2019][‡]	SORT [Bewley *et al.*, 2016]	0.216	0.303	0.288	38	57	1874	4.53	1086
DA FRCNN [Saito *et al.*, 2019][‡]	EMD tracker [Imirzian *et al.*, 2019]	0.225	0.333	0.316	52	31	3402	5.79	1223
Adapt DETR[‡]	SORT [Bewley *et al.*, 2016]	0.411	0.254	0.284	96	40	1851	2.70	1723
Adapt DETR[‡]	EMD tracker [Imirzian *et al.*, 2019]	0.413	0.275	0.305	**112**	**21**	2513	4.05	2085
DA-Tracker[‡] (joint detection and tracking)		**0.493**	**0.494**	**0.433**	42	42	**755**	**1.261**	**495**
Transtrack [Sun *et al.*, 2020][††] (joint detection and tracking)		0.918	0.584	0.584	200	1	819	0.491	823
Trackformer [Meinhardt *et al.*, 2022][††] (joint detection and tracking)		0.912	0.646	0.610	180	4	615	0.309	463

Table 2: Comparison of multi-object trackers on the target domain with and without domain adaptation techniques. We report three sets of baselines. In the first set we train non-adaptive multi-object trackers only on source domain data and evaluate on target data. In the second set of baselines, we do the same with domain-adaptive object trackers. Finally, as an upper bound, we evaluate fully supervised trackers directly trained on the target domain data. Highlighted methods are our contributions. † - Trained on source domain labels and images ‡ - Trained on source domain labels, images and target domain images †† - Trained on target domain labels and images.

mark. An explanation of the full details and intricacies of these metrics is beyond the scope of this paper. Full details are given in [Luiten *et al.*, 2021; Ristani *et al.*, 2016; Bernardin and Stiefelhagen, 2008]. In broad sketch terms, tracking requires so solve two sub-tasks (object localization and detection association) and these metrics emphasize the performance on these two sub-tasks to different extents. IDF1 is biased toward tracking accuracy, while MOTA is biased toward detection accuracy. HOTA provides an overall assessment of both of these factors.

5 Experiments

Implementation: Our proposed architecture incorporates a multi-object tracker that is connected to a set of discriminators at different levels. The MOT component of this network is based on [Meinhardt *et al.*, 2022], which is a joint-detector-tracker that has a feature extraction backbone and a transformer encoder-decoder network. Our method employs ResNet101 as the backbone for feature extraction. The transformer network, which takes both the backbone output and a positional embedding as input, is composed of a 6-layer encoder and a 6-layer decoder. This network includes three discriminators: two image feature discriminators and a track discriminator. The image feature discriminators are convolutional networks that use the output from the first and third layers of the backbone as local-level and global-level input features, respectively. They have three convolutional layers each. The track discriminator is an Multi-layer perceptron with two fully connected layers. It classifies the output embedding from the transformer decoder. We use gradient reverse layers to integrate the discriminators. Full details of the implementation are available in supplementary materials.

Experimental evaluation: We compare our method with a selection of representative baseline models on our dataset. We train our unsupervised domain adaptation MOT model on the training set of the source domain using the provided annotations and the training set of target domain (images only). Then we evaluated the model on the target domain test split.

For a direct comparison with our proposed domain adaptation tracking method, we did not find any relevant literature based on deep learning to address the domain adaptation for MOT problem directly. Therefore, we compare detection-based trackers that use domain adaptation methods in the object detection step. DA-FRCNN [Saito *et al.*, 2019] is a version of Faster R-CNN which uses domain adaptation techniques. We create tracks from these detections by using SORT [Bewley *et al.*, 2016] (Kalman filtering and Hungarian matching) and an earth-mover distance-based [Imirzian *et al.*, 2019] trackers. We produce an additional not previously published baseline by modifying DETR [Carion *et al.*, 2020] to perform domain adaptation using two discriminators working on image-level features similar to our proposed approach. This domain adaptive detection method is named as "Adapt DETR" in our experiments. We also use SORT and EMD to generate tracks using this domain adaptive detector.

For completeness, we also compare our proposed approach to baselines that do not use domain adaptation. These are Trackformer [Meinhardt *et al.*, 2022], and standard DETR [Carion *et al.*, 2020] with tracking using SORT [Bewley *et al.*, 2016] and EMD [Imirzian *et al.*, 2019]. The latter approach is closest to what has recently been proposed as a state-of-the-art tracker for real-world ant tracking in the biological literature [Imirzian *et al.*, 2019].

Tab. 2 summarizes our results. To measure the improvement achieved by applying domain adaptation, we compare against the models that do not use domain adaptation. Compared to Trackformer, we achieve more than 40% improvement in HOTA, MOTA and IDF1 metrics. Fig. 6 illustrates the difference between predictions in these two experiments. We note similar significant improvement against other non-domain adaptive approaches.

Compared with detection-based trackers that use domain adaptation in the detection component alone, our methods shows more than 27% improvement in HOTA, MOTA and IDF1. Compared to the best performing baseline, Adapt DETR, i.e. DETR with integrated domain adaptation strategy,

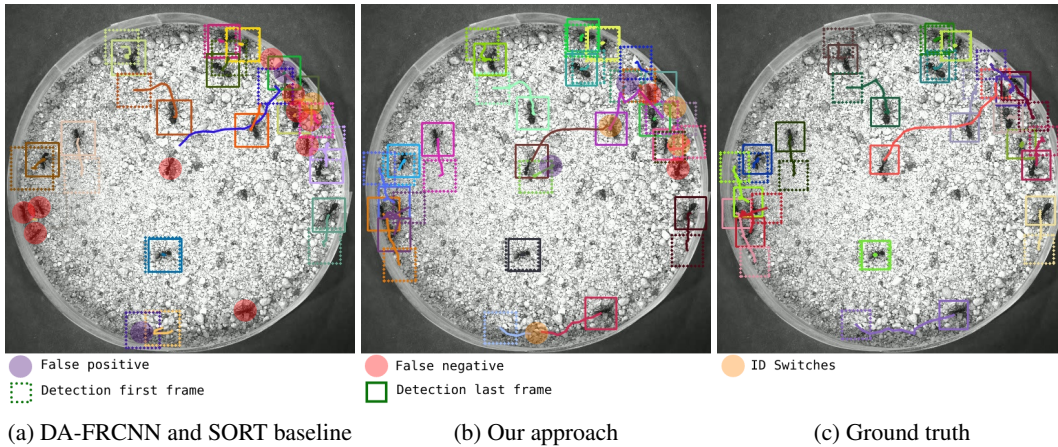| (a) DA-FRCNN and SORT baseline | (b) Our approach | (c) Ground truth |

Figure 5: Visualization of tracks between for twenty consecutive frames. Domain adaptive detection based tracker in 5a has many incorrect predictions in crowded areas compared to our proposed approach.

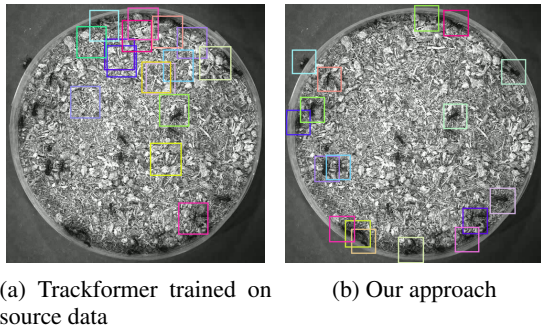(a) Trackformer trained on source data

(b) Our approach

Figure 6: We achieved a significant tracking improvement by applying our proposed domain adaptation techniques. This illustrates detections without (a) and with (b) domain adaptation.

D_{el}	D_{eg}	D_{tr}	MOTA	IDF1	HOTA	MT	ML	Frag	FP	ID Sw.
–	–	–	-1.02	0.187	0.262	57	118	1148	20.87	1351
✓	–	–	0.160	0.305	0.365	70	88	1916	5.021	5684
–	✓	–	0.301	0.325	0.331	65	48	1621	4.625	1996
✓	✓	–	0.381	0.453	**0.445**	**118**	**10**	1089	5.406	880
✓	✓	✓	**0.493**	**0.494**	0.433	42	42	**755**	**1.261**	**495**

Table 3: We conduct an ablation study to evaluate the accuracy improvement of domain adaptation components: Encoded feature local discriminator(D_{el}), Encoded feature global discriminator(D_{eg}) and Track discriminator(D_{tr}). Model without any of these components is equivalent to baseline Trackformer model. Combining D_{el} and D_{eg} yields significant improvement compared to individual components, but it results in many false detections due to incorrect predictions propagating to future predictions. D_{tr} address this issue by converting transformer decoder output embeddings.

followed by SORT or EMD tracker, our method still shows more than 16% improvement. Fig. 5 shows the difference between baseline and proposed approach track predictions.

To report an upper bound on the performance, i.e. an overly optimistic best case scenario, we trained a separate model directly on the target domain's training split in a supervised way assuming the annotated data is available. As expected, this approach performs better than our proposed framework.

Ablation study: Table 3 shows the result of the ablation study conducted to understand the contribution of each component. The results demonstrate the each discriminator's contribution to the final performance. Using only the visual feature discriminators, the performance can be improved considerably; However it still produces many false positives. This higher false positive rate is a result of object queries and track queries not being properly optimised for the target domain data. The integration of the track-level discriminator significantly reduces the number of false positives.

6 Conclusion

We have introduced a new multi-object tracking model capable of domain transfer. This was achieved by integrating multiple unsupervised adversarial discriminators at different processing stages into a joint-detection-and-tracking model.

Our experiments have shown that our approach can achieve noticeable performance improvements when tracking objects in a new target domain data with different visual appearances and shifted output distributions. This is exactly the case for many ant experiments in collective behaviour studies that use different ant species and different experimental assays.

Multi-object tracking of ant experiments is a core component of experiments in collective behaviour research but training a new model for every different setup and species is not practically feasible, since generating training data is as laborious and costly as tracking the experiment manually and in some cases even more. It is mostly for this reason that many experiments have to fall back on manual tracking. This severely limits experiment sizes and data throughput and thus the value and reach of such experiments. Our technique can clearly help to alleviate this bottleneck.

To encourage other researchers to work in the same space, we have provided a new benchmark dataset for ant tracking based on laboratory experiments. It is clear that this is only the beginning of making untagged multi-insect tracking a routine component in such experiments. We are continuing to expand this proposed model and specifically, we plan to next address the transfer between multiple domains. This will be reflected in the dataset that will be extended accordingly.

References

[Bae and Yoon, 2014] Seung-Hwan Bae and Kuk-Jin Yoon. Robust online multi-object tracking based on tracklet confidence and online discriminative appearance learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1218–1225, 2014.

[Bergmann *et al.*, 2019] Philipp Bergmann, Tim Meinhardt, and Laura Leal-Taixe. Tracking without bells and whistles. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 941–951, 2019.

[Bernardin and Stiefelhagen, 2008] Keni Bernardin and Rainer Stiefelhagen. Evaluating multiple object tracking performance: the clear mot metrics. *EURASIP Journal on Image and Video Processing*, 2008:1–10, 2008.

[Bewley *et al.*, 2016] Alex Bewley, Zongyuan Ge, Lionel Ott, Fabio Ramos, and Ben Upcroft. Simple online and realtime tracking. In *2016 IEEE international conference on image processing (ICIP)*, pages 3464–3468. IEEE, 2016.

[Boenisch *et al.*, 2018] Franziska Boenisch, Benjamin Rosemann, Benjamin Wild, David Dormagen, Fernando Wario, and Tim Landgraf. Tracking all members of a honey bee colony over their lifetime using learned models of correspondence. *Frontiers in Robotics and AI*, 5:35, 2018.

[Bousmalis *et al.*, 2016] Konstantinos Bousmalis, George Trigeorgis, Nathan Silberman, Dilip Krishnan, and Dumitru Erhan. Domain separation networks. *Advances in neural information processing systems*, 29, 2016.

[Carion *et al.*, 2020] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020.

[Chen *et al.*, 2014] Jianxu Chen, Cameron W Harvey, Mark S Alber, and Danny Z Chen. A matching model based on earth mover's distance for tracking myxococcus xanthus. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 113–120. Springer, 2014.

[Chen *et al.*, 2018] Yuhua Chen, Wen Li, Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Domain adaptive faster r-cnn for object detection in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3339–3348, 2018.

[Choi, 2015] Wongun Choi. Near-online multi-target tracking with aggregated local flow descriptor. In *Proceedings of the IEEE international conference on computer vision*, pages 3029–3037, 2015.

[Du *et al.*, 2018] Dawei Du, Yuankai Qi, Hongyang Yu, Yifan Yang, Kaiwen Duan, Guorong Li, Weigang Zhang, Qingming Huang, and Qi Tian. The unmanned aerial vehicle benchmark: Object detection and tracking. In *Proceedings of the European conference on computer vision (ECCV)*, pages 370–386, 2018.

[Feichtenhofer *et al.*, 2017] Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman. Detect to track and track

to detect. In *Proceedings of the IEEE international conference on computer vision*, pages 3038–3046, 2017.

[Gaidon and Vig, 2015] Adrien Gaidon and Eleonora Vig. Online domain adaptation for multi-object tracking. *arXiv preprint arXiv:1508.00776*, 2015.

[Ganin and Lempitsky, 2015] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, pages 1180–1189. PMLR, 2015.

[Geiger *et al.*, 2012] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *2012 IEEE conference on computer vision and pattern recognition*, pages 3354–3361. IEEE, 2012.

[Ghifary *et al.*, 2015] Muhammad Ghifary, W Bastiaan Kleijn, Mengjie Zhang, and David Balduzzi. Domain generalization for object recognition with multi-task autoencoders. In *Proceedings of the IEEE international conference on computer vision*, pages 2551–2559, 2015.

[He *et al.*, 2017] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[Hölldobler *et al.*, 1990] Bert Hölldobler, Edward O Wilson, et al. *The ants*. Harvard University Press, 1990.

[Huang *et al.*, 2022] Wei-Jie Huang, Yu-Lin Lu, Shih-Yao Lin, Yusheng Xie, and Yen-Yu Lin. Aqt: Adversarial query transformers for domain adaptive object detection. In *31st International Joint Conference on Artificial Intelligence, IJCAI 2022*, pages 972–979. International Joint Conferences on Artificial Intelligence, 2022.

[Imirzian *et al.*, 2019] Natalie Imirzian, Yizhe Zhang, Christoph Kurze, Raquel G Loreto, Danny Z Chen, and David P Hughes. Automated tracking and analysis of ant trajectories shows variation in forager exploration. *Scientific reports*, 9(1):1–10, 2019.

[Jiang *et al.*, 2020] Pin Jiang, Aming Wu, Yahong Han, Yunfeng Shao, Meiyu Qi, and Bingshuai Li. Bidirectional adversarial training for semi-supervised domain adaptation. In *IJCAI*, pages 934–940, 2020.

[Kumar *et al.*, 2020] SV Aruna Kumar, Ehsan Yaghoubi, Abhijit Das, BS Harish, and Hugo Proença. The p-destre: A fully annotated dataset for pedestrian detection, tracking, and short/long-term re-identification from aerial devices. *IEEE Transactions on Information Forensics and Security*, 16:1696–1708, 2020.

[Lan *et al.*, 2016] Long Lan, Dacheng Tao, Chen Gong, Naiyang Guan, and Zhigang Luo. Online multi-object tracking by quadratic pseudo-boolean optimization. In *IJCAI*, pages 3396–3402, 2016.

[Leal-Taixé *et al.*, 2011] Laura Leal-Taixé, Gerard Pons-Moll, and Bodo Rosenhahn. Everybody needs somebody: Modeling social and grouping behavior on a linear programming multiple people tracker. In *2011 IEEE inter-*

national conference on computer vision workshops (ICCV workshops), pages 120–127. IEEE, 2011.

[Liu *et al.*, 2020] Qiankun Liu, Qi Chu, Bin Liu, and Nenghai Yu. Gsm: Graph similarity model for multi-object tracking. In *IJCAI*, pages 530–536, 2020.

[Long *et al.*, 2017] Mingsheng Long, Han Zhu, Jianmin Wang, and Michael I Jordan. Deep transfer learning with joint adaptation networks. In *International conference on machine learning*, pages 2208–2217. PMLR, 2017.

[Luiten *et al.*, 2021] Jonathon Luiten, Aljosa Osep, Patrick Dendorfer, Philip Torr, Andreas Geiger, Laura Leal-Taixé, and Bastian Leibe. Hota: A higher order metric for evaluating multi-object tracking. *International journal of computer vision*, 129(2):548–578, 2021.

[Meinhardt *et al.*, 2022] Tim Meinhardt, Alexander Kirillov, Laura Leal-Taixe, and Christoph Feichtenhofer. Trackformer: Multi-object tracking with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8844–8854, 2022.

[Milan *et al.*, 2016] Anton Milan, Laura Leal-Taixé, Ian Reid, Stefan Roth, and Konrad Schindler. Mot16: A benchmark for multi-object tracking. *arXiv preprint arXiv:1603.00831*, 2016.

[Naiser *et al.*, 2018] Filip Naiser, Matěj Šmíd, and Jiří Matas. Tracking and re-identification system for multiple laboratory animals. In *Visual observation and analysis of vertebrate and insect behavior workshop at international conference on pattern recognition (ICPR)*, 2018.

[Oh *et al.*, 2009] Songhwai Oh, Stuart Russell, and Shankar Sastry. Markov chain monte carlo data association for multi-target tracking. *IEEE Transactions on Automatic Control*, 54(3):481–497, 2009.

[Pérez-Escudero *et al.*, 2014] 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.

[Ray and Stopfer, 2022] Subhasis Ray and Mark A Stopfer. Argos: A toolkit for tracking multiple animals in complex visual environments. *Methods in Ecology and Evolution*, 13(3):585–595, 2022.

[Ren *et al.*, 2015] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015.

[Rezatofighi *et al.*, 2015] Seyed Hamid Rezatofighi, Anton Milan, Zhen Zhang, Qinfeng Shi, Anthony Dick, and Ian Reid. Joint probabilistic data association revisited. In *Proceedings of the IEEE international conference on computer vision*, pages 3047–3055, 2015.

[Ristani *et al.*, 2016] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *European conference on computer vision*, pages 17–35. Springer, 2016.

[Romero-Ferrero *et al.*, 2019] Francisco Romero-Ferrero, Mattia G Bergomi, Robert C Hinz, Francisco JH Heras, and Gonzalo G De Polavieja. Idtracker. ai: tracking all individuals in small or large collectives of unmarked animals. *Nature methods*, 16(2):179–182, 2019.

[Sadeghian *et al.*, 2017] Amir Sadeghian, Alexandre Alahi, and Silvio Savarese. Tracking the untrackable: Learning to track multiple cues with long-term dependencies. In *Proceedings of the IEEE international conference on computer vision*, pages 300–311, 2017.

[Saito *et al.*, 2019] Kuniaki Saito, Yoshitaka Ushiku, Tatsuya Harada, and Kate Saenko. Strong-weak distribution alignment for adaptive object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6956–6965, 2019.

[Saleh *et al.*, 2021] Fatemeh Saleh, Sadegh Aliakbarian, Hamid Rezatofighi, Mathieu Salzmann, and Stephen Gould. Probabilistic tracklet scoring and inpainting for multiple object tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14329–14339, 2021.

[Schulter *et al.*, 2017] Samuel Schulter, Paul Vernaza, Wongun Choi, and Manmohan Chandraker. Deep network flow for multi-object tracking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6951–6960, 2017.

[Sun *et al.*, 2020] Peize Sun, Jinkun Cao, Yi Jiang, Rufeng Zhang, Enze Xie, Zehuan Yuan, Changhu Wang, and Ping Luo. Transtrack: Multiple object tracking with transformer. *arXiv preprint arXiv:2012.15460*, 2020.

[Sun *et al.*, 2022] Peize Sun, Jinkun Cao, Yi Jiang, Zehuan Yuan, Song Bai, Kris Kitani, and Ping Luo. Dancetrack: Multi-object tracking in uniform appearance and diverse motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20993–21002, 2022.

[Wojke *et al.*, 2017] Nicolai Wojke, Alex Bewley, and Dietrich Paulus. Simple online and realtime tracking with a deep association metric. In *2017 IEEE international conference on image processing (ICIP)*, pages 3645–3649. IEEE, 2017.

[Wu *et al.*, 2022] Meihong Wu, Xiaoyan Cao, Xiaoyu Cao, and Shihui Guo. A dataset of ant colonies motion trajectories in indoor and outdoor scenes for social cluster behavior study. *arXiv preprint arXiv:2204.04380*, 2022.

[Yan *et al.*, 2017] Hongliang Yan, Yukang Ding, Peihua Li, Qilong Wang, Yong Xu, and Wangmeng Zuo. Mind the class weight bias: Weighted maximum mean discrepancy for unsupervised domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2272–2281, 2017.

[Zhang *et al.*, 2021] Jingyi Zhang, Jiaxing Huang, Zhipeng Luo, Gongjie Zhang, and Shijian Lu. Da-detr: Domain adaptive detection transformer by hybrid attention. *arXiv preprint arXiv:2103.17084*, 2021.