NetTrack: Tracking Highly Dynamic Objects with a Net

Guangze Zheng[1,2] Shijie Lin[1,2] Haobo Zuo[1,2] Changhong Fu [3] Jia Pan[1,2*]
[1]The University of Hong Kong [2]Centre for Transformative Garment Production [3]Tongji University
{guangze, lsj2048, haobozuo}@connect.hku.hk, changhongfu@tongji.edu.cn, jpan@cs.hku.hk
https://george-zhuang.github.io/nettrack

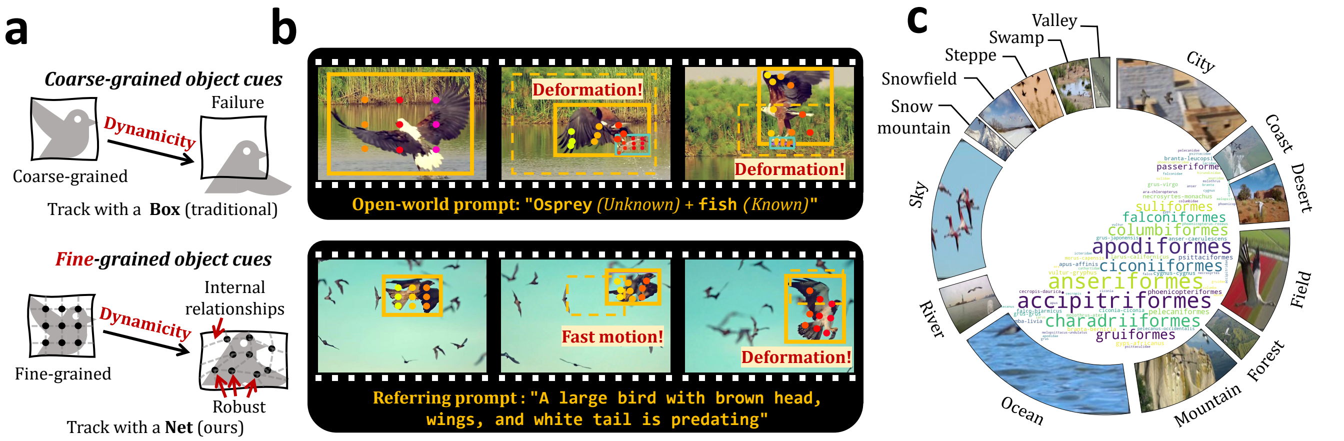

Figure 1. **a** The visualization of the proposed NetTrack is similar to a **Net**. Object dynamicity distorts the internal relationships of the object, presenting challenges for traditional coarse-grained tracking methods that rely solely on bounding boxes. While NetTrack introduces fine-grained Nets that are robust to dynamicity. **b** Qualitative results of NetTrack tracking highly dynamic objects under *open-world tracking* and *referring expression comprehension* settings. Dynamicity like deformation and fast motion results in drastic changes in the coarse-grained representation, while the fine-grained Nets can contract robustly. The dashed boxes represent the object position from the previous time step. **c** We propose a challenging benchmark named BFT, dedicated to evaluating highly dynamic object tracking with abundant scenarios shown in the external circular and diverse species shown in the central word cloud.

Abstract

The complex dynamicity of open-world objects presents non-negligible challenges for multi-object tracking (MOT), often manifested as severe deformations, fast motion, and occlusions. Most methods that solely depend on coarse-grained object cues, such as boxes and the overall appearance of the object, are susceptible to degradation due to distorted internal relationships of dynamic objects. To address this problem, this work proposes ***NetTrack****, an efficient, generic, and affordable tracking framework to introduce fine-grained learning that is robust to dynamicity. Specifically, NetTrack constructs a dynamicity-aware association with a fine-grained* ***Net****, leveraging point-level visual cues. Correspondingly, a fine-grained sampler and matching method have been incorporated. Furthermore, NetTrack learns object-text correspondence for fine-grained localization. To evaluate MOT in extremely dynamic open-world scenarios, a bird flock tracking (****BFT****) dataset is constructed, which exhibits high dynamicity with diverse species and open-world scenarios. Comprehensive evaluation on BFT validates the effectiveness of fine-grained learning on object dynamicity, and thorough transfer experiments on challenging open-world benchmarks, i.e., TAO, TAO-OW, AnimalTrack, and GMOT-40, validate the strong generalization ability of NetTrack even without finetuning.*

1. Introduction

Multiple object tracking (MOT) [7, 23, 34, 42, 45] aims to maintain continuous visual perception of objects of interest in videos and the real world. Traditional MOT methods often assume objects as coarse-grained entities because in classical MOT tasks [9, 61, 67], the dynamicity of specific object categories [10] and scenes is not significant, and

[*] Corresponding author.

the internal relationships within objects are relatively stable. However, the demand for tracking arbitrary objects, especially highly dynamic objects, in open-world MOT tasks [31, 39, 47] severely challenges this assumption.

The high dynamicity of open-world objects, manifested as severe deformation, fast motion, and frequent occlusion, poses challenges for existing methods in two major aspects:

1) **Association** For most methods relying solely on coarse-grained visual representations, the high dynamicity renders the temporal continuity fragile in terms of association, since the internal relationships in the objects are distorted. These methods typically represent the overall object as coarse-grained bounding boxes [7, 71] or the corresponding features [23, 34], and the dynamicity significantly reduces the similarity of such representations across different time steps, as shown in Fig. 1-b.

2) **Localization** The high dynamicity also poses challenges to establishing accurate object-text correspondence for localization. State-of-the-art (SoTA) methods [23, 34] typically learn the coarse-grained correspondence between the entire image and text in pre-training. For severely deformed or occluded objects, these methods often struggle to localize.

In this work, we propose NetTrack, introducing fine-grained learning to address the above two aspects. Regarding association, NetTrack utilizes physical points on the object's appearance that are less susceptible to object dynamicity and form fine-grained visual cues. For localization, grounded pre-training is utilized to learn fine-grained correspondences between objects and text. Therefore, our primary contributions are outlined as follows:

Fine-grained Net for dynamicity-aware association Instead of viewing the object as a coarse-grained entity, this work tracks the object with a fine-grained Net, which leverages points of interest (POIs) on the surface of object appearance. The dynamicity, such as deformations, distorted internal relationships between POIs by altering global relative position and appearance feature distribution, while the fine-grained representations of the points themselves, such as local appearance color and relationships with neighboring points, are seldom affected and exhibit robustness, as shown in Fig. 1-b. Following this viewpoint, we design a fine-grained sampler to discover potential POIs and utilize fine-grained visual cues of these points, along with the emerging physical point tracking methods [12, 22, 28], for robust tracking. Subsequently, a simple yet effective fine-grained similarity calculation method is proposed to determine the containment relationship between the tracked POIs and candidate objects. The proposed fine-grained similarity scores are combined with the existing coarse-grained to achieve more robust association of dynamic objects.

Object-text correspondence for fine-grained localization To discover and localize highly dynamic objects of interest in tracking, this work adopts a pre-training method to tracking by phrase grounding [32, 41, 68] for fine-grained object-text correspondence. Compared to CLIP-based tracking methods [23, 34] that utilize coarse-grained image-text correspondence, NetTrack can more effectively distinguish highly dynamic objects, as shown in Fig. 2. Furthermore, by embedding descriptors (GPT-3.5 [6] in Fig. 2) within the framework, the proposed framework learns contextual information, such as provided professional application and knowledge context, to mitigate background interference and achieve practical real-world applications for efficient dynamic object tracking.

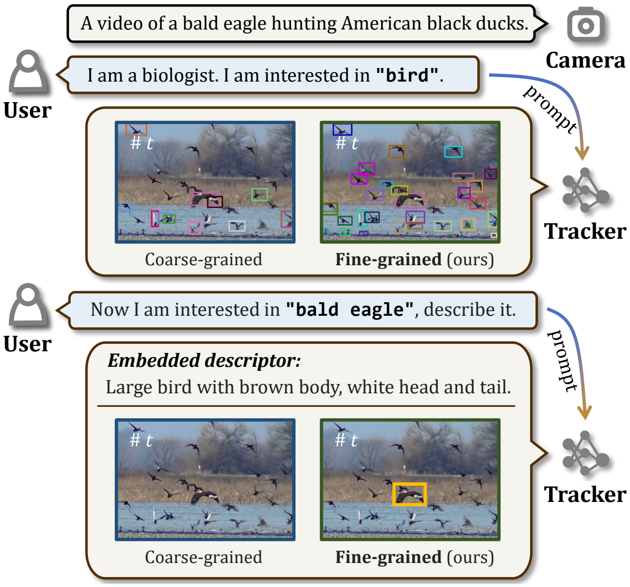
A video of a bald eagle hunting American black ducks.
Camera
User
I am a biologist. I am interested in "bird".
prompt
# t
# t
Tracker
Coarse-grained
Fine-grained (ours)
User
Now I am interested in "bald eagle", describe it.
prompt
Embedded descriptor:
Large bird with brown body, white head and tail.
# t
# t
Tracker
Coarse-grained
Fine-grained (ours)

Figure 2. Comparison between the localization method [23, 50] based on coarse-grained object-text correspondence and our fine-grained method. Our fine-grained approach localizes dynamic objects better and can leverage professional descriptions from embedded descriptors (GPT-3.5 [6] in the example) with a better understanding of context.

A highly dynamic benchmark and transfer experiments on diverse scenarios This work introduces a highly dynamic open-world MOT dataset, named bird flock tracking (BFT), to evaluate the performance of tracking methods in tracking highly dynamic objects. BFT is particularly notable for the complex and unpredictable dynamicity of 22 bird species for three main reasons: 1) Fast motion caused by the three-dimensional activity space. 2) Deformations resulting from frequent flapping of wings [37]. 3) Occlusions arising from the collective behavior of birds [38, 39] in the flock. Furthermore, BFT comprises 14 distinct open-world scenes and 22 species in 106 sequences, showcasing a rich diversity, as depicted in Fig. 1-c. In evaluation, the proposed NetTrack framework reaches SoTA performance in tracking highly dynamic objects in BFT. Besides, comprehensive zero-shot transfer experiments show NetTrack surpasses tracking baselines on several challenging open-world MOT benchmarks, *e.g.*, TAO [8, 34], TAO-OW [42], AnimalTrack [69], and GMOT-40 [1]. The introduced fine-grained learning contributes to stronger generalization ability of NetTrack even without finetuning. As an efficient, generic, and affordable tracking framework, NetTrack also exhibits potential in open-world application scenarios, further highlighting its suitability for downstream tasks.

2. Related Work

Open-world multi-object tracking methods Tracking-by-detection [4, 5, 7, 48, 65, 71] has been the most popular framework in MOT, which includes localizing potential objects and associating them over time [45]. Traditional MOT methods typically focus on limited scenes and object categories, such as pedestrians [4, 59, 65, 70] in public places or vehicles [53, 71] in autonomous driving scenarios. Comparatively, open-world tracking tasks require trackers to have the ability to track any object in complex and dynamic scenes. The rise of CLIP [54]-based open-set object detection [19, 50, 73] has promoted this task, inspiring advanced open-world tracking baselines [34, 50] to utilize CLIP-style pre-training to achieve generalization by leveraging the correspondence between text and images. However, these mainstream tracking methods [4, 7, 34, 59, 71] usually view objects as coarse-grained bounding boxes, but the high dynamicity of open-world objects can often disrupt the temporal similarity of this coarse representation. Moreover, compared to the shallow-fused vision-language features used in CLIP-like pre-training, localizing dynamic objects often requires establishing fine-grained correspondences between the object and text to counteract the appearance distortion or impairment of the objects.

The recent emergence of physical point tracking methods [11, 12, 22, 28, 63] has inspired this work to introduce fine-grained visual cues of objects. These methods aim to track arbitrary physical points over video clips, relying on point-level appearance representation rather than coarsely propagating the overall object, therefore holding promise to maintain good generalization for dynamic objects. Additionally, the pre-training approach based on phrase grounding [21] has also been applied in open-set object detection tasks [32, 41, 68], and its potential benefits for dynamic object tracking are anticipated due to object-level, language-aware, and semantic-rich visual representations.

Open-world multi-object tracking benchmarks Classical MOT benchmarks mainly focus on limited object categories and scenarios, where objects typically maintain stable appearances or poses and undergo relatively simple motion, *e.g.*, tracking pedestrian [9, 10, 30, 49] or vehicles [18, 60, 67]. With increasing demands for open-world tracking applications, MOT benchmarks that focus on a wider range of scenarios and object classes have emerged. TAO [8] includes numerous *unseen* objects in massive data, GMOT-40 [1] focuses on tracking *unseen* object categories, and AnimalTrack [69] places emphasis on tracking wildlife. Later, TAO-OW [42] defines *known* and *unknown* object categories in an open-world setting, and Li *et al.* [34] divide object categories into *base* and *novel* on the TAO benchmark in an open-vocabulary setting. In diverse open-world MOT tasks, while learning *unseen* classes is of paramount importance, the dynamicity resulting from potential severe deformations and fast motion of these *unseen* objects are equally crucial, necessitating a comprehensive evaluation.

3. Method

The proposed NetTrack framework introduces a fine-grained Net for dynamicity-aware object association and fine-grained object-text correspondence for dynamicity-aware localization. Sec. 3.1 describes structuring the object into fine-grained Nets using sampling and performing association. Sec. 3.2 primarily discusses how fine-grained object-text correspondence positively affects the localization of dynamic objects.

3.1. Fine-Grained Net

The proposed dynamicity-aware association utilizes fine-grained Nets to construct robust visual cues for object dynamicity. It mainly consists of a fine-grained sampler and a matching method. The overall process is shown in Fig. 3.

Fine-grained sampler This work introduces point-level visual cues to form fine-grained Nets with points of interest (POIs). Ideally, sampling POIs should accurately capture every valuable point on the surfaces of every interested object, avoiding background interference or the redundant computational burden. A straightforward thought is to sample POIs inside boxes of tracked objects and update points frame by frame. However, such an approach can impose a certain computational burden, ignorance of false negative samples, and insufficient visual context. Therefore, a fine-grained sampler is proposed to sample cross-frame POIs.

Denote the expected distribution of POIs as $f(\mathbf{x})$, where $\mathbf{x}$ refers to a point in the image I. The object motion is estimated based on Kalman Filter [27] as in [4, 5, 7, 71]. Such estimation acts as the coarse distribution of novel objects in a certain period of $\mathcal{S}$ frames. The distribution can then be transformed to a point-level form as $p(\mathbf{x}|\mathcal{T}_\text{o}^\text{coarse}, \{\mathbf{I}\}_{i=1}^{\mathcal{S}})$, where $\mathcal{T}_\text{o}^\text{coarse}$ is the coarsely estimated coarse-grained trajectory of objects and $p(\cdot)$ is a binary distribution that discover the potential POIs. This distribution serves as an importance weight to sample the POIs. Given point number K, the expected POIs can then be formulated using importance sampling [62] as:

$$\mathbb{E}_{\mathbf{x}\sim p(\mathbf{x}|\{\mathbf{I}\}_{i=1}^{\mathcal{S}})}[f(\mathbf{x})] = \frac{1}{K}\sum_{i=1}^{K}\frac{f(\mathbf{x}_i)}{p(\mathbf{x}_i|\mathcal{T}_\text{o}^\text{coarse}, \{\mathbf{I}\}_{i=1}^{\mathcal{S}})}\,. \quad (1)$$

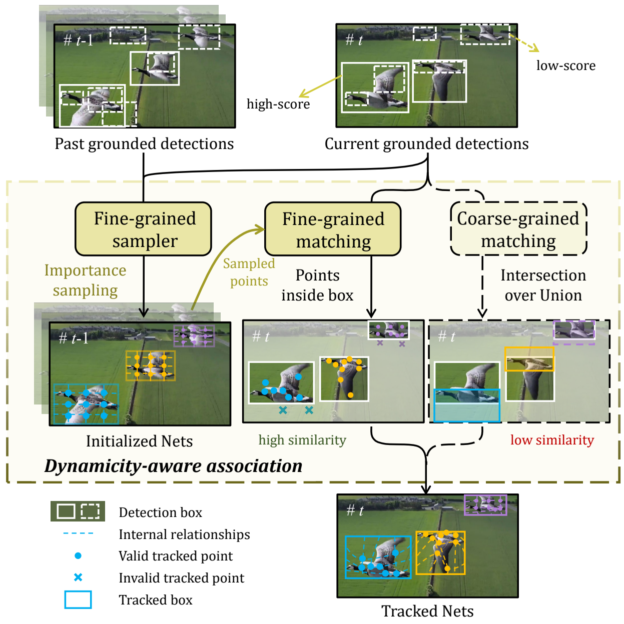

Figure 3. Dynamicity-aware association in the NetTrack framework. Unlike the coarse-grained association methods that only learn the box motion or overall appearance, dynamicity-aware association benefits from fine-grained Nets which are robust against the open-world dynamicity and exhibit stronger generalization ability.

Therefore, the fine-grained POIs are determined at frame #t-1 and estimated at frame #t with point tracking models.

Fine-grained matching Utilizing fine-grained Nets for tracking requires matching the memorized POIs with the current detection results based on temporal similarity. Given the point tracker model $\mathtt{Tr}_{\mathrm{p}}$, the estimated point trajectories $\mathcal{T}_{\mathrm{p}}$ can be obtained in the aforementioned period. After acquiring the detection results $\mathcal{D}_t$ for the current frame $\#t$, the fine-grained matching method calculates the number of estimated points from a Net that fall inside a candidate detection box as fine-grained similarity. Suppose N is the number of tracked objects in frame $\#t-1$, the element $\mathbf{S}_{i,j}$ of the matching fine-grained score matrix $\mathbf{S}$ of N Nets $\{\mathbf{P}_i\}_{i=1}^{N}$ and M detection boxes $\{\mathbf{b}_j\}_{j=1}^{M}$ can be expressed as:

$$\mathbf{S}_{i,j} = w_{i,j}\frac{|\mathbf{P}_i \cap \mathbf{b}_j|}{|\mathbf{P}_i|} \;, \quad w_{i,j} = \min\{1, \frac{\mathcal{A}(\hat{\mathbf{b}}_i)}{\mathcal{A}(\mathbf{b}_j)}\} \;, \tag{2}$$

where w is a weight to penalize candidate detection boxes with excessively large areas, as larger areas often result in predicted points prone to fall within the box, leading to potential misjudgments. $|\mathbf{P}_i \cap \mathbf{b}_j|$ refers to the number of points from Net $\mathbf{P}_i$ positioning inside $\mathbf{b}_j$, depicted as valid points in Fig. 3, and $|\mathbf{P}_i|$ is the number of points in Net $\mathbf{P}_i$. $\mathcal{A}(\cdot)$ refers to area of a box, and $\hat{\mathbf{b}}$ is the predicted box of a tracked object in frame $\#t$ using [27]. Afterward, combined with the coarse-grained similarity score, the overall matching score can be obtained. As shown in Fig. 3, object dynamicity often leads to a decrease in coarse-grained similarity in Intersection over Union (IOU), while fine-grained association remains robust. The matching process is then carried out using the Hungarian algorithm [29]. Details of the method are described in Algorithm 1.

3.2. Fine-Grained Object-Text Correspondence

To learn fine-grained object-text correspondence for localization, this work introduces a pre-training strategy based on phrase grounding to track dynamic objects and mitigates the adverse effects of object dynamics with a deep fusion of textual and object features. Different from SoTA tracking methods [23, 34] that utilizes CLIP [54]-based pre-training, we follow [32, 41, 68] to identify the correspondence between phrases in sentences and objects in images to formulate fine-grained object-text correspondence. Given the input image $\mathbf{I}$ and language prompt $\mathbf{P}$, corresponding object features $\mathbf{F}_{\mathrm{O}}$ and language features $\mathbf{F}_{\mathrm{L}}$ can be obtained with a visual encoder $\mathrm{Enc}_{\mathrm{V}}$ and a language encoder $\mathrm{Enc}_{\mathrm{L}}$, respectively. Afterward, we can get fused features $\mathbf{F}'_{\mathrm{O}}$ and $\mathbf{F}'_{\mathrm{L}}$ by deep fusion, and further obtain the object-text correspondence score $\mathbf{S}_{\mathrm{ground}}$. The formula for this process is:

$$\begin{aligned} &\mathbf{F}_{\mathrm{O}} = \mathrm{Enc}_{\mathrm{V}}(\mathbf{I}), \;\; \mathbf{F}_{\mathrm{L}} = \mathrm{Enc}_{\mathrm{L}}(\mathbf{P}), \\ &\mathbf{F}'_{\mathrm{O}}, \; \mathbf{F}'_{\mathrm{L}} = \mathrm{Fuse}(\mathbf{F}_{\mathrm{O}}, \mathbf{F}_{\mathrm{L}}), \;\; \mathbf{S}_{\mathrm{ground}} = \mathbf{F}'_{\mathrm{O}}\mathbf{F}'^{\top}_{\mathrm{L}}. \end{aligned} \tag{3}$$

From a visual perspective, fine-grained object-text correspondence enhances the language awareness of visual features, thereby enabling better discernment of deformed objects. From a language view, such correspondence learns contextualized representations at the word or sub-sentence level during pre-training [41], avoiding biases caused by unnecessary word interactions. The proposed framework also allows for a more detailed understanding of the object with an embedded descriptor, *e.g.*, large language models [6, 52]. Consequently, such fine-grained correspondence is better suited for capturing more specific contextual information in professional scenarios, as illustrated in Fig. 2 and Fig. 10.

4. BFT Dataset

Data collection Bird flocks are among the most dynamic objects to track in the open world and thus are considered ideal subjects for this work. The dynamicity of birds is mainly attributed to three phenomena: 1) Bird flocks exhibit higher maneuverability compared to ground objects due to the three-dimensional activity space and an additional degree of freedom. In addition, the inertia of birds is relatively small, allowing them to accelerate, decelerate, and change direction more flexibly. The complex aerodynamic effects [16] also make the motion of flying bird flocks more difficult to predict. 2) Birds generally experience frequent and intense deformation during flight, mainly due to wingbeat [37]. 3) Collective behavior [38, 39] is widespread in

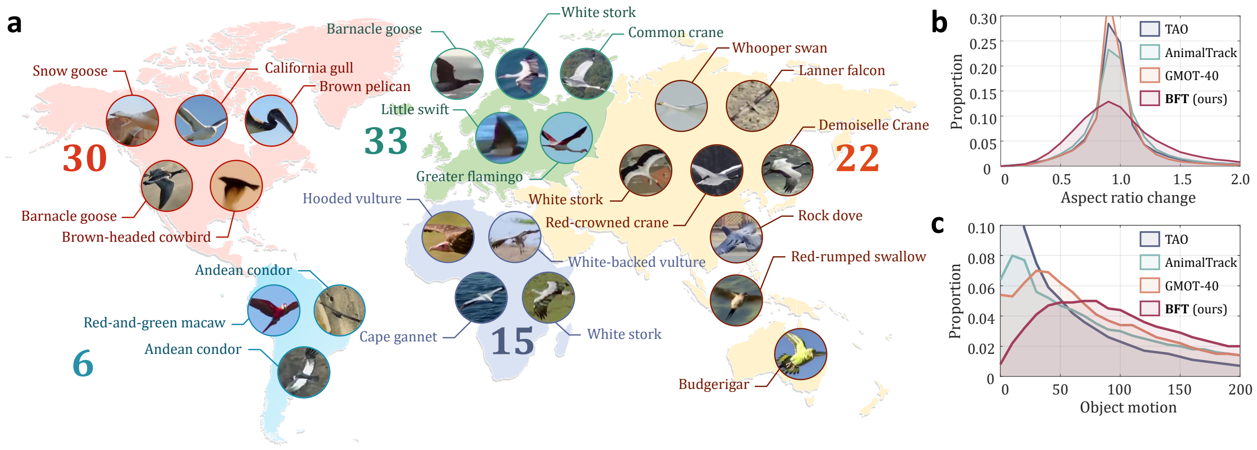

Figure 4. **a** The diverse geographical distribution of some representative flying bird species exhibits the diversity of BFT. The numbers on the map represent the number of videos in each corresponding area, *e.g.*, 30 videos from North America. **b** Dynamicity comparison between BFT and other datasets on aspect ratio change. The more dispersed distribution means more frequent object deformation and occlusion in BFT. **c** Dynamicity comparison between BFT and other datasets on object motion. The larger object motion in BFT represents the faster motion of objects.

many species of bird flocks. This often results in a dense distribution of bird flocks within a limited space, making it visually susceptible to occlusion. In addition to the aforementioned dynamic challenges, birds often have similar appearances in flocks, which also adds to the difficulty of visual discrimination.

To showcase the diversity of open-world scenarios and the variety of species, the BFT dataset incorporates 22 bird species and 14 common natural and cultural scenes, covering six continents as illustrated in Fig. 4-a and Fig. 1-c. Detailed corresponding order, family, genus, and species of bird are in Fig. 7. The primary data source is the BBC nature documentary series *Earthflight* [25]. 106 carefully selected clips were extracted from around 6 hours of video, which were further divided into a training set with 35 videos, a validation set with 25 videos, and a test set with 36 videos. All the data underwent meticulous annotation by experts and multiple rounds of review by tracking domain experts, as well as verification by experts in the field of biology. The frame rate of both the videos and annotations is commonly set at 25 frames per second (FPS).

High dynamicity The higher dynamicity in BFT includes more severe deformations, faster motion, and more frequent occlusions. Quantitatively, Fig. 4-b,c compare the dynamicity of BFT with other open-world MOT datasets [1, 8, 69] from two aspects. Specifically, aspect ratio change (ARC) [15, 51] is a commonly used tracking attribute, which measures the frequency and severity of object deformations or occlusions. Object motion is another attribute to measure the displacement of an object between two consecutive time steps. Detailed statistics are shown in Sec. 8. Due to the more dispersed ARC distribution of BFT and the larger values of the motion distribution, BFT represents stronger dynamicity compared to other dataset.

5. Experiments

The experimental section aims to validate the following core conclusions of this work:

1) High dynamicity of open-world objects poses significant challenges for MOT.
2) NetTrack outstandingly handles dynamic objects and exhibits strong generalization abilities on diverse open-world tracking datasets without finetuning.
3) The proposed fine-grained learning shows stronger generalization abilities for tracking dynamic objects compared to coarse-grained methods.

5.1. Settings

Dataset BFT is utilized to assess the performance of trackers in highly dynamic open-world scenarios. In zero-shot transfer evaluation, the validation sets of the large-scale TAO-OW [42] and TAO [8] are employed for extensive generalization ability assessment. Specifically, the evaluation of TAO follows the description in [34], where an open-vocabulary setting is adopted for *base* and *novel* categories, and the classification ability of trackers is evaluated. *Novel* classes are the classes defined as rare in the LVIS [20] dataset. Differently, object classes of TAO-OW are divided into *known* and *unknown* based on whether they belong to the 80 categories in COCO [36]. In the ablation experiments, in addition to TAO and TAO-OW, AnimalTrack [69] and GMOT-40 [1] are also included as references and evaluated in an open-world setting following TAO-OW. Regarding AnimalTrack, 8 out of 10 classes are outside the COCO categories. Similarly, 12 out of 18 classes in GMOT-40 are

Table 1. Overall evaluation on the highly dynamic BFT. * denotes comprehensive evaluation metrics. Finetuned results are in gray, and the best results in each setting are in **bold**. † denotes the offline setting.

Detector	Method	BFT benchmark evaluation										Line
		OWTA*↑	D. Re.↑	D. Acc.↑	D. Pr.↑	A. Acc.↑	A. Re.↑	A. Pr.↑	HOTA*↑	MOTA*↑	IDF1*↑	
a) Finetuned on BFT dataset												
YOLO-v3 [55]	JDE [64]	33.0	46.9	40.9	61.0	23.4	24.7	66.7	30.7	35.4	37.4	1
CenterNet [74]	CenterTrack [75]	61.6	70.5	58.5	71.5	54.0	57.8	82.8	56.0	60.2	61.0	2
	FairMOT [70]	40.2	57.5	53.3	75.7	28.2	29.4	78.3	38.5	56.0	41.8	3
YOLO-v5 [24]	CSTrack [35]	34.2	49.6	47.0	79.9	23.7	24.5	81.1	33.2	46.7	34.5	4
Deformable DETR [76]	TransTrack [59]	66.8	73.9	64.2	76.9	60.3	64.6	82.1	62.1	71.4	71.4	5
	TrackFormer [48]	67.4	74.5	66.0	78.9	61.1	65.8	82.8	63.3	74.1	72.4	6
	P3AFormer [72]	42.3	40.9	38.1	71.9	44.0	46.9	75.7	40.7	43.8	52.9	7
	TransCenter [66]	63.5	73.2	65.8	77.6	55.3	58.5	82.8	60.0	76.4	68.6	8
YOLOX [17]	SORT [4]	63.2	64.2	60.6	78.7	62.3	65.5	82.0	61.2	75.5	77.2	9
	IOUTracker [5]	**70.5**	**75.2**	**67.5**	79.0	66.3	70.2	85.5	66.6	**78.5**	76.4	10
	ByteTrack [71]	65.2	66.3	61.2	75.3	64.1	69.0	77.1	62.5	77.2	**82.3**	11
	OC-SORT [7]	68.9	69.2	65.4	**83.8**	**68.7**	**72.1**	**86.8**	**66.8**	77.1	79.3	12
b) Zero-shot setting												
YOLOX [17]	SORT [4]	54.2	55.4	52.2	79.5	53.0	55.7	82.8	52.5	60.6	63.6	14
	IOUTracker [5]	55.6	57.3	54.1	84.9	53.9	57.4	84.1	53.9	60.1	59.1	15
	DeepSORT [65]	44.7	53.1	48.1	71.0	37.8	40.4	74.8	42.3	51.3	49.9	16
	Tracktor++ [2]	29.0	65.8	60.9	82.4	12.9	20.4	29.7	27.8	35.0	26.2	17
	ByteTrack [71]	54.8	56.0	51.6	73.3	53.7	58.5	73.4	52.5	61.5	**68.4**	18
	OC-SORT [7]	58.5	57.7	55.2	**87.7**	**59.4**	**62.0**	**87.9**	57.2	61.0	66.6	19
	StrongSORT [14]	43.2	54.7	48.3	73.0	34.2	36.8	74.2	40.4	47.9	43.4	20
	StrongSORT++ [14]	42.9	54.1	44.2	61.7	34.4	37.5	69.2	38.6	39.4	42.9	21
	Deep OC-SORT [46]	33.6	26.4	25.4	81.5	42.8	45.3	84.3	32.9	25.7	39.3	22
	NetTrack (ours)	**63.3**	**70.6**	**61.2**	77.6	56.8	60.5	83.3	**58.8**	**62.5**	65.1	23
Grounding DINO [41]	SORT [4]	59.9	63.9	60.1	81.1	56.2	58.9	84.1	57.9	71.4	69.7	24
	IOUTracker [5]	70.9	77.4	62.3	71.9	65.0	**70.8**	82.4	63.5	65.8	70.7	25
	ByteTrack [71]	64.1	67.9	61.1	73.0	60.5	66.5	73.7	60.7	74.9	**78.9**	26
	OC-SORT [7]	69.0	70.9	66.8	**87.9**	**67.2**	70.1	**90.4**	66.9	73.6	76.0	27
	NetTrack (ours)	**72.5**	**80.7**	**72.6**	83.3	65.2	70.4	82.5	**68.7**	**78.9**	77.0	28

outside the COCO categories.

Metrics Open-world tracking accuracy (OWTA) [42] is an open-world MOT metric proposed for TAO-OW and is the main metric in our experiments. OWTA evaluates both *detection recall* (D. Re.) and *association accuracy* (A. Acc.), respectively. *Detection accuracy* (D. Acc.), *detection precision* (D. Pr.), *association recall* (A. Re.), and *association precision* (A. Pr.) are reference metrics. TETA [33] aims to evaluate multi-category objects and is used to evaluate the TAO dataset under an open-vocabulary setting. *Localization score* (LocA) and *association score* (AssocA) are calculated in TETA. HOTA [44], MOTA [3], and IDF1 [56] are classic metrics used for comparisons with classic MOT methods on BFT and serve as references. All evaluation processes are adopted from TrackEval [26].

Implementation details In NetTrack, the coarse-grained association adapts from BYTE [71], and the default point tracker adapts from CoTracker [28] pretrained on TAP-Vid-Kubric [11]. By default, the tracking stride is 8, lost tracks are retained for 30 frames, and the initialized point sampling is with a grid of `(3,3)`. The default detector is GroundingDINO [41] with Swin-B [43] backbone, which was pretrained on COCO [36], O365 [58], *etc*. To validate the generalization ability of NetTrack in an affordable manner for open-world MOT applications, no additional training is required for all evaluated benchmarks. The finetuning and evaluation of the publicly available SoTA trackers on BFT followed their default settings.

Table 2. Zero-shot transfer evaluation on open-vocabulary MOT comparison. * denotes comprehensive evaluation metrics, and * represents non open-world setting, *i.e.*, also trained on *novel* classes on TAO. Results of finetuning and learning *novel* classes are shown in gray, and the best results are shown in **bold**.

Method	TAO benchmark evaluation							
	Base				Novel			
	TETA*↑	LocA↑	AssocA↑	ClsA↑	TETA*↑	LocA↑	AssocA↑	ClsA↑
a) Finetuned on TAO dataset								
DeepSORT [65]	26.9	47.1	15.8	17.7	21.1	46.4	14.7	2.3
Tracktor++ [2]	28.3	47.4	20.5	17.0	22.7	46.7	19.3	2.2
QDTrack* [53]	27.1	45.6	24.7	11.0	22.5	42.7	24.4	0.4
TETer* [33]	30.3	47.4	31.6	12.1	25.7	45.9	31.1	0.2
b) Trained with LVIS dataset								
OVTrack [34]	**35.5**	**49.3**	**36.9**	20.2	27.8	48.8	**33.6**	1.5
c) Zero-shot setting								
NetTrack (ours)	33.0	45.7	28.6	**24.8**	**32.6**	**51.3**	33.0	**13.3**

5.2. High Dynamicity Evaluation

A comprehensive evaluation of NetTrack and other SoTA trackers on the highly dynamic BFT is presented in Tab. 1. The evaluation is divided into two main parts: *a)* Finetuning on the BFT dataset using closed-set trackers. *b)* Open-world MOT condition, which involves tracking under zero-shot settings. To ensure a fair evaluation of tracker performance in the highly dynamic challenges of the open-world scenarios, all text prompts for open-world conditions only include `'bird'`, consistent with the category in the COCO dataset that is used to train closed-set trackers. The

Table 3. Zero-shot transfer evaluation on TAO-OW. * denotes comprehensive evaluation metrics, and * represents non open-world setting, *i.e.*, also trained on *unknown* classes. Finetuned results are in gray, and the best results are shown in **bold**.

Method	TAO-OW benchmark evaluation					
	Known			Unknown		
	OWTA*↑	D. Re.↑	A. Acc.↑	OWTA*↑	D. Re.↑	A. Acc.↑
a) Finetuned on TAO-OW dataset						
SORT [4]	46.6	67.4	33.7	33.9	43.4	30.3
SORT-TAO* [8]	54.2	74.0	40.6	39.9	68.8	24.1
AOA* [13]	52.8	72.5	39.1	49.7	74.7	33.4
Tracktor [2]	57.9	**80.2**	42.6	22.8	54.0	10.0
OWTB [42]	60.2	77.2	47.4	39.2	46.9	34.5
Video OWL-ViT [23]	59.0	69.0	**51.5**	**45.4**	53.4	**40.5**
b) Zero-shot setting						
NetTrack (ours)	**62.7**	77.4	51.0	43.7	**58.7**	33.2

experimental results mainly demonstrate that:

1) Even in the zero-shot open-world tracking setting, NetTrack achieves superior performance compared to SoTA finetuned closed-set trackers. NetTrack improves 1.3 points on OWTA compared with the best finetuned results, confirming the zero-shot generalization ability of the proposed framework.
2) In comparison to the results after fine-tuning (lines 9-12), closed-set trackers exhibit sub-optimal zero-shot generalization ability (lines 13,14,17,18) in highly dynamic open-world scenarios, with an average decrease of 16% on OWTA, 15% on HOTA, and 21% on MOTA, which indicates closed-set trackers have suboptimal generalization ability on dynamic objects in the open world.
3) NetTrack encourages associating potential objects of interest and achieves an improvement on *detection recall* by 3.4 points. It also results in more false positive samples and adds pressure to the association with a slight decrease in A. Acc. However the comprehensive OWTA gets promoted by 1.6 points compared with the best coarse-grained association methods (lines 24-27).

5.3. Zero-Shot Transfer Evaluation

Zero-shot transfer on open-vocabulary settings In Tab. 2, zero-shot transfer on TAO with open-vocabulary MOT evaluation is shown. DeepSORT [65] and Tracktor++ [2] are with ViLD [19] as the detector. OVTrack [34] is trained on a generated dataset derived from LVIS [20], which exhibits a high level of class consistency with TAO. Compared to finetuned trackers, NetTrack significantly improves tracking classification accuracy and achieves strong zero-shot tracking accuracy. Although NetTrack is susceptible to a large number of false positive samples due to the absence of finetuning, which puts it at a slight disadvantage in the evaluation of LocA and AssocA in the *base* classes, the proposed framework achieves an 11.8-point increase in ClsA, a 2.5-point increase in LocA, comparable AssocA in the *novel* classes and a 4.5-point increase in overall TETA, further demonstrating its competitive generalization ability.

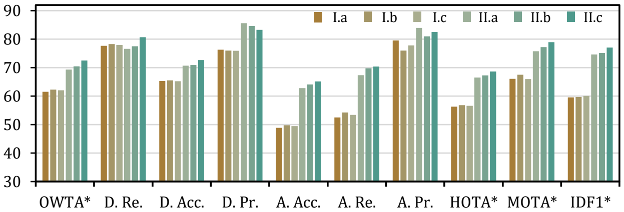

Figure 5. Comparison of detachable modules in the proposed framework, where the SoTA grounding-based detectors [32, 41] as I, II and point trackers [12, 22, 28] as a, b, c are considered. The robust performance in module variations confirms the excellent generality of the proposed framework.

Zero-shot transfer on open-world settings Zero-shot generalization of NetTrack on the TAO-OW [42] benchmark is demonstrated in Tab. 3. Apart from NetTrack, all trackers underwent fine-tuning on *known* classes of TAO-OW training set. Compared to finetuned SoTA trackers, NetTrack achieves optimal performance on *known* classes. With the D. Re. being similar to the open world tracking baseline (OWTB) [42], A. Acc. surpasses the baseline by 3.6 points, confirming the generalization ability of dynamicity-aware association. Similarly, while A. Acc. remains approximate to Video OWL-ViT [23], D. Re. shows an improvement of 8.4 points, validating the effectiveness of fine-grained localization. On *unknown* classes, the introduction of false positive samples leads to a slight decrease in A. Acc., but the overall OWTA performance is still competitive with a 5.3-point improvement on D. Re.

5.4. Ablations

Generality of fine-grained Nets In Tab. 4 and Tab. 5, the comparison between the proposed association with fine-grained Nets and coarse-grained methods [4, 7, 71] on TAO [8], TAO-OW [42], AnimalTrack [69], and GMOT-40 [1] are shown. Attributed to the proposed framework that encourages the discovery of more potential objects in open-world scenarios, NetTrack achieves significant improvements in LocA and D. Re. on both *seen* and *unseen* classes across four benchmarks. Particularly, the D. Re. of *unknown* classes on TAO-OW exhibits a remarkable increase of 18.2 points compared to the second-best performance, confirming its strong generalization. Although the introduction of false positive samples leads to a slight decrease in AssoA and A. Acc, the overall TETA and OWTA have been significantly improved in both the *seen* and *unseen* classes.

Robust framework with detachable modules To validate the generality of the proposed framework, Fig. 5 shows ablation study on detachable modules, including open-set localization methods and point trackers. Specifically, the localization methods are denoted as GLIP [32] I, Grounding

Table 4. Association comparison between fine-grained Nets and solely coarse-grained methods on TAO and TAO-OW validation benchmark in an open-vocabulary [34] and an open-world setting [42], respectively. The best results are shown in **bold**. The same detector [41] is used.

Method	TAO benchmark evaluation								TAO-OW benchmark evaluation					
	Base				Novel				Known			Unknown		
	TETA*↑	LocA↑	AssocA↑	ClsA↑	TETA*↑	LocA↑	AssocA↑	ClsA↑	OWTA*↑	D. Re.↑	A. Acc.↑	OWTA*↑	D. Re.↑	A. Acc.↑
SORT [4]	28.5	33.8	27.3	24.5	28.1	41.7	28.6	**13.7**	62.4	72.8	53.8	39.5	39.6	40.6
ByteTrack [71]	32.6	41.7	**31.2**	24.7	32.1	48.8	**34.4**	13.3	**63.3**	71.3	56.4	40.9	40.5	**42.5**
OC-SORT [7]	25.8	32.4	20.5	24.5	25.7	39.0	24.7	13.4	48.7	69.0	34.4	31.6	37.8	27.1
NetTrack (ours)	**33.0**	**45.7**	28.6	**24.8**	**32.6**	**51.3**	33.0	13.3	62.7	**77.4**	51.0	**43.7**	**58.7**	33.2

Table 5. Association comparison between fine-grained Nets and solely coarse-grained methods on AnimalTrack and GMOT-40 benchmark following the open-world setting of TAO-OW [42] for reference. The best results are shown in **bold**. The same detector [41] is used.

Method	AnimalTrack benchmark evaluation						GMOT-40 benchmark evaluation					
	Known			Unknown			Known			Unknown		
	OWTA*↑	D. Re.↑	A. Acc.↑	OWTA*↑	D. Re.↑	A. Acc.↑	OWTA*↑	D. Re.↑	A. Acc.↑	OWTA*↑	D. Re.↑	A. Acc.↑
SORT [4]	44.2	32.7	60.0	48.1	42.6	54.7	43.6	30.6	62.4	35.8	24.1	53.4
ByteTrack [71]	41.7	28.1	62.2	46.7	41.7	52.6	41.3	27.2	**63.1**	35.8	21.8	**59.2**
OC-SORT [7]	45.2	32.9	**62.6**	48.7	43.0	**55.8**	44.0	31.0	62.7	36.4	24.6	54.1
NetTrack (ours)	**48.1**	**42.0**	55.6	**51.4**	**50.8**	52.5	**45.9**	**37.8**	56.1	**36.6**	**29.3**	45.7

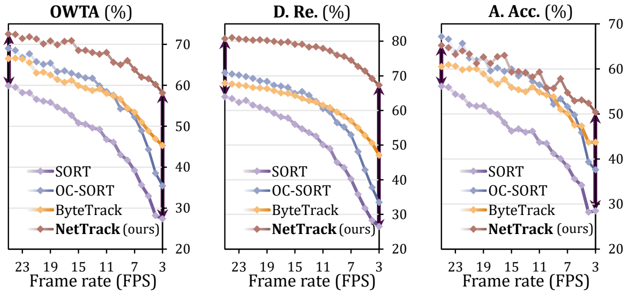

Figure 6. Stability comparison against frame-rate drops between NetTrack and coarse-grained methods [4, 7, 71] on BFT. Better stability indicates stronger generalization ability of NetTrack.

DINO [41] II, and point trackers are denoted as PIPs [22] a, TAPIR [12] b, CoTracker [28] c. The combination of Grounding DINO and CoTracker is denoted as II.c and serves as the default setting. In comparing localization ability, both methods demonstrate competitiveness in terms of D. Re. but [32] exhibits a slight performance deficit in A. Acc and overall OWTA due to the introduction of more false positives. Similarly, three point trackers exhibit approximately excellent performance. Overall, the change of modules does not significantly degrade the overall performance, thus verifying the good generalization ability of the proposed framework.

Stability against frame-rate drops In practical applications of open-world tracking, especially in scenarios related to edge devices [40], it is common to encounter reduced video frame rates due to the need to reduce computational load or save energy, further exacerbating the challenges posed by the dynamicity of open-world objects. Fig. 6 shows the tracking performance on the BFT dataset under reduced frame rates, from default frame rates (25 FPS) to one-tenth (3 FPS). Compared to other association methods [4, 7, 71], NetTrack demonstrates good stability in the face of reduced frame rates. This further illustrates the generalization performance of the proposed framework.

Remark The qualitative tracking results of diverse scenarios are shown in supplemented Sec. 9.4. The promising applications of NetTrack in video editing, open-world ecological inspection, and embedding descriptors for professional use are discussed in Sec. 10.

6. Conclusion

This work focuses on the high dynamicity in open-world MOT and proposes NetTrack to learn fine-grained object cues. Specifically, fine-grained visual cues and object-text correspondence are introduced for dynamicity-aware association and localization. This work also proposes a highly dynamic open-world MOT benchmark, BFT, and extensive evaluation with SoTA trackers proves the effectiveness of the proposed NetTrack for tracking dynamic objects. Moreover, extensive transfer experiments on several challenging open-world MOT benchmarks validate the strong generalization ability of NetTrack without finetuning. The analysis of limitations suggests that a more streamlined end-to-end manner and filtering false positive samples are promising for further improvement.

Acknowledge

This research project is partially supported by the Innovation and Technology Commission of the HKSAR Government under the InnoHK initiative, Hong Kong General Research Fund (11202119 and 11208718), Innovation and Technology Commission (GHP/126/21GD), and Guangdong, Hong Kong and Macao Joint Innovation Project (2023A0505010016). We also appreciate precious biological advice from Ming-Shan Wang.

References

[1] Hexin Bai, Wensheng Cheng, Peng Chu, Juehuan Liu, Kai Zhang, and Haibin Ling. GMOT-40: A Benchmark for Generic Multiple Object Tracking. In *CVPR*, pages 6719–6728, 2021. 3, 5, 7, 2

[2] Philipp Bergmann, Tim Meinhardt, and Laura Leal-Taixe. Tracking without Bells and Whistles. In *ICCV*, pages 941–951, 2019. 6, 7

[3] Keni Bernardin and Rainer Stiefelhagen. Evaluating Multiple Object Tracking Performance: The CLEAR MOT Metrics. *EURASIP J. Image Video Process.*, 2008:1–10, 2008. 6, 5

[4] Alex Bewley, Zongyuan Ge, Lionel Ott, Fabio Ramos, and Ben Upcroft. Simple Online and Realtime Tracking. In *ICIP*, pages 3464–3468, 2016. 3, 6, 7, 8, 1

[5] Erik Bochinski, Volker Eiselein, and Thomas Sikora. High-Speed Tracking-by-Detection without Using Image Information. In *AVSS*, pages 1–6, 2017. 3, 6

[6] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language Models Are Few-Shot Learners. In *NeurIPS*, pages 1877–1901, 2020. 2, 4, 7

[7] Jinkun Cao, Jiangmiao Pang, Xinshuo Weng, Rawal Khirodkar, and Kris Kitani. Observation-Centric SORT: Rethinking SORT for Robust Multi-Object Tracking. In *CVPR*, pages 9686–9696, 2023. 1, 2, 3, 6, 7, 8

[8] Achal Dave, Tarasha Khurana, Pavel Tokmakov, Cordelia Schmid, and Deva Ramanan. TAO: A Large-Scale Benchmark for Tracking Any Object. In *ECCV*, pages 436–454, 2020. 3, 5, 7, 2

[9] Patrick Dendorfer, Hamid Rezatofighi, Anton Milan, Javen Shi, Daniel Cremers, Ian Reid, Stefan Roth, Konrad Schindler, and Laura Leal-Taixé. MOT20: A Benchmark for Multi Object Tracking in Crowded Scenes. *arXiv preprint arXiv:2003.09003*, 2020. 1, 3

[10] Patrick Dendorfer, Aljosa Osep, Anton Milan, Konrad Schindler, Daniel Cremers, Ian Reid, Stefan Roth, and Laura Leal-Taixé. MOTChallenge: A Benchmark for Single-Camera Multiple Target Tracking. *IJCV*, 129:845–881, 2021. 1, 3

[11] Carl Doersch, Ankush Gupta, Larisa Markeeva, Adrià Recasens, Lucas Smaira, Yusuf Aytar, João Carreira, Andrew Zisserman, and Yi Yang. TAP-Vid: A Benchmark for Tracking Any Point in a Video. In *NeurIPS*, pages 13610–13626, 2022. 3, 6

[12] Carl Doersch, Yi Yang, Mel Vecerik, Dilara Gokay, Ankush Gupta, Yusuf Aytar, Joao Carreira, and Andrew Zisserman. TAPIR: Tracking Any Point with Per-frame Initialization and Temporal Refinement. In *ICCV*, pages 1–19, 2023. 2, 3, 7, 8

[13] Fei Du, Bo Xu, Jiasheng Tang, Yuqi Zhang, Fan Wang, and Hao Li. 1st Place Solution to ECCV-TAO-2020: Detect and Represent Any Object for Tracking. *arXiv preprint arXiv:2101.08040*, 2021. 7

[14] Yunhao Du, Zhicheng Zhao, Yang Song, Yanyun Zhao, Fei Su, Tao Gong, and Hongying Meng. StrongSORT: Make DeepSORT Great Again. *IEEE TMM*, 2023. 6

[15] Heng Fan, Liting Lin, Fan Yang, Peng Chu, Ge Deng, Sijia Yu, Hexin Bai, Yong Xu, Chunyuan Liao, and Haibin Ling. LaSOT: A High-Quality Benchmark for Large-Scale Single Object Tracking. In *CVPR*, pages 5374–5383, 2019. 5

[16] Steven N Fry, Rosalyn Sayaman, and Michael H Dickinson. The Aerodynamics of Free-Flight Maneuvers in Drosophila. *Science*, 300(5618):495–498, 2003. 4

[17] Zheng Ge, Songtao Liu, Feng Wang, Zeming Li, and Jian Sun. YOLOX: Exceeding YOLO Series in 2021. *arXiv preprint arXiv:2107.08430*, 2021. 6

[18] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are We Ready for Autonomous Driving? The KITTI Vision Benchmark Suite. In *CVPR*, pages 3354–3361, 2012. 3

[19] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-Vocabulary Object Detection via Vision and Language Knowledge Distillation. In *ICLR*, pages 1–21, 2021. 3, 7

[20] Agrim Gupta, Piotr Dollar, and Ross Girshick. LVIS: A Dataset for Large Vocabulary Instance Segmentation. In *CVPR*, pages 5356–5364, 2019. 5, 7, 2

[21] Tanmay Gupta, Arash Vahdat, Gal Chechik, Xiaodong Yang, Jan Kautz, and Derek Hoiem. Contrastive Learning for Weakly Supervised Phrase Grounding. In *ECCV*, pages 752–768, 2020. 3

[22] Adam W Harley, Zhaoyuan Fang, and Katerina Fragkiadaki. Particle Video Revisited: Tracking Through Occlusions Using Point Trajectories. In *ECCV*, pages 59–75, 2022. 2, 3, 7, 8

[23] Georg Heigold, Matthias Minderer, Alexey Gritsenko, Alex Bewley, Daniel Keysers, Mario Lučić, Fisher Yu, and Thomas Kipf. Video OWL-ViT: Temporally-Consistent Open-World Localization in Video. In *ICCV*, pages 13802–13811, 2023. 1, 2, 4, 7

[24] Glenn Jocher, Ayush Chaurasia, Alex Stoken, Jirka Borovec, NanoCode012, Yonghye Kwon, Kalen Michael, TaoXie, Jiacong Fang, imyhxy, Lorna, Zeng Yifu, Colin Wong, Abhiram V, Diego Montes, Zhiqiang Wang, Cristi Fati, Jebastin Nadar, Laughing, and Mrinal Jain. YOLOv5 SOTA Realtime Instance Segmentation. 6

[25] Downer John and Tennant David. Earthflight. https://www.bbc.co.uk/programmes/b018xsc1. BBC, 2011. 5

[26] Luiten Jonathon and Hoffhues Arne. Trackeval. https://github.com/JonathonLuiten/TrackEval, 2020. 6

[27] R. Kalman. A New Approach to Linear Filtering and Prediction Problems. *J. Basic Eng.*, 82:35–45, 1960. 3, 4

[28] Nikita Karaev, Ignacio Rocco, Benjamin Graham, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. Co-Tracker: It is Better to Track Together. *arXiv preprint arXiv:2307.07635*, 2023. 2, 3, 6, 7, 8

[29] Harold W Kuhn. The Hungarian Method for the Assignment Problem. *Nav. Res. Logist.*, 2(1-2):83–97, 1955. 4

[30] Laura Leal-Taixé, Anton Milan, Ian Reid, Stefan Roth, and Konrad Schindler. MOTChallenge 2015: Towards a Benchmark for Multi-Target Tracking. *arXiv preprint arXiv:1504.01942*, 2015. 3

[31] Yongseok Lee, Wonkyung Do, Hanbyeol Yoon, Jinuk Heo, WonHa Lee, and Dongjun Lee. Visual-Inertial Hand Motion Tracking with Robustness against Occlusion, Interference, and Contact. *Sci. Robot.*, 6(58):eabe1315, 2021. 2

[32] Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, et al. Grounded Language-Image Pre-Training. In *CVPR*, pages 10965–10975, 2022. 2, 3, 4, 7, 8

[33] Siyuan Li, Martin Danelljan, Henghui Ding, Thomas E Huang, and Fisher Yu. Tracking Every Thing in the Wild. In *ECCV*, pages 498–515, 2022. 6, 2

[34] Siyuan Li, Tobias Fischer, Lei Ke, Henghui Ding, Martin Danelljan, and Fisher Yu. OVTrack: Open-Vocabulary Multiple Object Tracking. In *CVPR*, pages 5567–5577, 2023. 1, 2, 3, 4, 5, 6, 7, 8

[35] Chao Liang, Zhipeng Zhang, Xue Zhou, Bing Li, Shuyuan Zhu, and Weiming Hu. Rethinking the Competition between Detection and ReID in Multiobject Tracking. *IEEE TIP*, 31: 3182–3196, 2022. 6

[36] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common Objects in Context. In *ECCV*, pages 740–755, 2014. 5, 6, 2

[37] Hangjian Ling, Guillam E Mclvor, Geoff Nagy, Sepehr MohaimenianPour, Richard T Vaughan, Alex Thornton, and Nicholas T Ouellette. Simultaneous Measurements of Three-Dimensional Trajectories and Wingbeat Frequencies of Birds in the Field. *J. R. Soc. Interface*, 15(147):20180653, 2018. 2, 4, 6, 7

[38] Hangjian Ling, Guillam E Mclvor, Kasper van der Vaart, Richard T Vaughan, Alex Thornton, and Nicholas T Ouellette. Costs and Benefits of Social Relationships in the Collective Motion of Bird Flocks. *Nat. Ecol. Evol.*, 3(6):943–948, 2019. 2, 4

[39] Hangjian Ling, Guillam E Mclvor, Joseph Westley, Kasper van der Vaart, Richard T Vaughan, Alex Thornton, and Nicholas T Ouellette. Behavioural Plasticity and the Transition to Order in Jackdaw Flocks. *Nat. Commun.*, 10(1): 5174, 2019. 2, 4

[40] Fang Liu, Guoming Tang, Youhuizi Li, Zhiping Cai, Xingzhou Zhang, and Tongqing Zhou. A Survey on Edge Computing Systems and Tools. *Proc. IEEE*, 107(8):1537–1562, 2019. 8

[41] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, et al. Grounding DINO: Marrying DINO with Grounded Pre-Training for Open-Set Object Detection. *arXiv preprint arXiv:2303.05499*, 2023. 2, 3, 4, 6, 7, 8

[42] Yang Liu, Idil Esen Zulfikar, Jonathon Luiten, Achal Dave, Deva Ramanan, Bastian Leibe, Aljoša Ošep, and Laura Leal-Taixé. Opening Up Open World Tracking. In *CVPR*, pages 19045–19055, 2022. 1, 3, 5, 6, 7, 8, 2

[43] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin Transformer: Hierarchical Vision Transformer Using Shifted Windows. In *ICCV*, pages 10012–10022, 2021. 6

[44] 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. *IJCV*, 129:548–578, 2021. 6, 5

[45] Wenhan Luo, Junliang Xing, Anton Milan, Xiaoqin Zhang, Wei Liu, and Tae-Kyun Kim. Multiple Object Tracking: A Literature Review. *Artif. Intell.*, 293:103448, 2021. 1, 3

[46] Gerard Maggiolino, Adnan Ahmad, Jinkun Cao, and Kris Kitani. Deep OC-SORT: Multi-Pedestrian Tracking by Adaptive Re-Identification. In *ICIP*, pages 1–5, 2023. 6

[47] Ivan Masmitja, Mario Martin, Tom O'Reilly, Brian Kieft, Narcís Palomeras, Joan Navarro, and Kakani Katija. Dynamic Robotic Tracking of Underwater Targets Using Reinforcement Learning. *Sci. Robot.*, 8(80):eade7811, 2023. 2

[48] Tim Meinhardt, Alexander Kirillov, Laura Leal-Taixe, and Christoph Feichtenhofer. TrackFormer: Multi-Object Tracking with Transformers. In *CVPR*, pages 8844–8854, 2022. 3, 6

[49] 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. 3, 1

[50] Matthias Minderer, Alexey Gritsenko, Austin Stone, Maxim Neumann, Dirk Weissenborn, Alexey Dosovitskiy, Aravindh Mahendran, Anurag Arnab, Mostafa Dehghani, Zhuoran Shen, et al. Simple Open-Vocabulary Object Detection. In *ECCV*, pages 728–755, 2022. 2, 3

[51] Matthias Muller, Adel Bibi, Silvio Giancola, Salman Alsubaihi, and Bernard Ghanem. TrackingNet: A Large-Scale Dataset and Benchmark for Object Tracking in the Wild. In *ECCV*, pages 300–317, 2018. 5

[52] OpenAI. GPT-4 Technical Report. *arXiv preprint arXiv:2303.08774*, pages 1–100, 2023. 4, 7

[53] Jiangmiao Pang, Linlu Qiu, Xia Li, Haofeng Chen, Qi Li, Trevor Darrell, and Fisher Yu. Quasi-Dense Similarity Learning for Multiple Object Tracking. In *CVPR*, pages 164–173, 2021. 3, 6

[54] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning Transferable Visual Models from Natural Language Supervision. In *ICML*, pages 8748–8763, 2021. 3, 4, 5

[55] Joseph Redmon and Ali Farhadi. YOLOv3: An Incremental Improvement. *arXiv preprint arXiv:1804.02767*, 2018. 6

[56] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. Performance Measures and a Dataset for Multi-Target, Multi-Camera Tracking. In *ECCV*, pages 17–35, 2016. 6, 5

[57] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-Resolution Image Synthesis with Latent Diffusion Models. In *CVPR*, pages 10684–10695, 2022. 6, 7

[58] Shuai Shao, Zeming Li, Tianyuan Zhang, Chao Peng, Gang Yu, Xiangyu Zhang, Jing Li, and Jian Sun. Objects365: A Large-Scale, High-Quality Dataset for Object Detection. In *ICCV*, pages 8430–8439, 2019. 6

[59] 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*, pages 1–11, 2020. 3, 6

[60] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. Scalability in Perception for Autonomous Driving: Waymo Open Dataset. In *CVPR*, pages 2446–2454, 2020. 3

[61] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. Scalability in Perception for Autonomous Driving: Waymo Open Dataset. In *CVPR*, pages 2446–2454, 2020. 1

[62] Surya T Tokdar and Robert E Kass. Importance Sampling: A Review. *Wiley Interdiscip. Rev. Comput. Stat.*, 2(1):54–60, 2010. 3

[63] Qianqian Wang, Yen-Yu Chang, Ruojin Cai, Zhengqi Li, Bharath Hariharan, Aleksander Holynski, and Noah Snavely. Tracking Everything Everywhere All at Once. In *ICCV*, pages 1–15, 2023. 3

[64] Zhongdao Wang, Liang Zheng, Yixuan Liu, Yali Li, and Shengjin Wang. Towards Real-Time Multi-Object Tracking. In *ECCV*, pages 107–122, 2020. 6

[65] Nicolai Wojke, Alex Bewley, and Dietrich Paulus. Simple Online and Realtime Tracking with A Deep Association Mmetric. In *ICIP*, pages 3645–3649, 2017. 3, 6, 7

[66] Yihong Xu, Yutong Ban, Guillaume Delorme, Chuang Gan, Daniela Rus, and Xavier Alameda-Pineda. TransCenter: Transformers With Dense Representations for Multiple-Object Tracking. *IEEE TPAMI*, pages 1–16, 2022. 6

[67] Fisher Yu, Haofeng Chen, Xin Wang, Wenqi Xian, Yingying Chen, Fangchen Liu, Vashisht Madhavan, and Trevor Darrell. BDD100K: A Diverse Driving Dataset for Heterogeneous Multitask Learning. In *CVPR*, pages 2636–2645, 2020. 1, 3

[68] Haotian Zhang, Pengchuan Zhang, Xiaowei Hu, Yen-Chun Chen, Liunian Li, Xiyang Dai, Lijuan Wang, Lu Yuan, Jenq-Neng Hwang, and Jianfeng Gao. GLIPv2: Unifying Localization and Vision-Language Understanding. In *NeurIPS*, pages 36067–36080, 2022. 2, 3, 4

[69] Libo Zhang, Junyuan Gao, Zhen Xiao, and Heng Fan. AnimalTrack: A Benchmark for Multi-Animal Tracking in the Wild. *IJCV*, 131(2):496–513, 2023. 3, 5, 7, 2

[70] Yifu Zhang, Chunyu Wang, Xinggang Wang, Wenjun Zeng, and Wenyu Liu. FairMOT: On the Fairness of Detection and Re-Identification in Multiple Object Tracking. *IJCV*, 129: 3069–3087, 2021. 3, 6

[71] Yifu Zhang, Peize Sun, Yi Jiang, Dongdong Yu, Fucheng Weng, Zehuan Yuan, Ping Luo, Wenyu Liu, and Xinggang Wang. ByteTrack: Multi-Object Tracking by Associating Every Detection Box. In *ECCV*, pages 1–21, 2022. 2, 3, 6, 7, 8, 1

[72] Zelin Zhao, Ze Wu, Yueqing Zhuang, Boxun Li, and Jiaya Jia. Tracking Objects as Pixel-Wise Distributions. In *ECCV*, pages 76–94, 2022. 6

[73] Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, et al. RegionCLIP: Region-Based Language-Image Pretraining. In *CVPR*, pages 16793–16803, 2022. 3

[74] Xingyi Zhou, Dequan Wang, and Philipp Krähenbühl. Objects as Points. *arXiv preprint arXiv:1904.07850*, 2019. 6

[75] Xingyi Zhou, Vladlen Koltun, and Philipp Krähenbühl. Tracking Objects as Points. In *ECCV*, pages 474–490, 2020. 6

[76] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable DETR: Deformable Transformers for End-to-End Object Detection. *arXiv preprint arXiv:2010.04159*, 2020. 6