SemTrack: A Large-scale Dataset for Semantic Tracking in the Wild

Pengfei Wang[1,⋆], Xiaofei Hui[1,2,⋆], Jing Wu[1,⋆], Zile Yang[1,⋆], Kian Eng Ong[1,⋆], Xinge Zhao[1], Beijia Lu[1], Dezhao Huang[3], Evan Ling[3], Weiling Chen[4,‡], Keng Teck Ma[3], Minhoe Hur[3], and Jun Liu[1,2,†]

[1] Singapore University of Technology and Design
[2] Lancaster University
[3] Hyundai Motor Group Innovation Center in Singapore (HMGICS)
[4] Tianqiao and Chrissy Chen Institute
{pengfei_wang, jing_wu, zile_yang}@sutd.edu.sg,
{xiaofei_hu, kianeng_ong}@mymail.sutd.edu.sg

Abstract. Knowing merely where the target is located is not sufficient for many real-life scenarios. In contrast, capturing rich details about the tracked target via its semantic trajectory, i.e. who/what this target is interacting with and when, where, and how they are interacting over time, is especially crucial and beneficial for various applications (e.g., customer analytics, public safety). We term such tracking as Semantic Tracking and define it as tracking the target based on the user's input and then, most importantly, capturing the semantic trajectory of this target. Acquiring such information can have significant impacts on sales, public safety, etc. However, currently, there is no dataset for such comprehensive tracking of the target. To address this gap, we create SemTrack, a large and comprehensive dataset containing annotations of the target's semantic trajectory. The dataset contains 6.7 million frames from 6961 videos, covering a wide range of 52 different interaction classes with 115 different object classes spanning 10 different supercategories in 12 types of different scenes, including both indoor and outdoor environments. We also propose SemTracker, a simple and effective method, and incorporate a meta-learning approach to better handle the challenges of this task. Our dataset and code can be found at `https://sutdcv.github.io/SemTrack`.

Keywords: Dataset · Semantic tracking · Semantic trajectory

1 Introduction

Tracking a target (*e.g.*, humans or objects) in a scene is important and useful for a wide range of applications (*e.g.*, customer analytics, crowd management).

⋆ Equal contributions
‡ Work was done at HMGICS.
† Corresponding Author (j.liu81@lancaster.ac.uk).

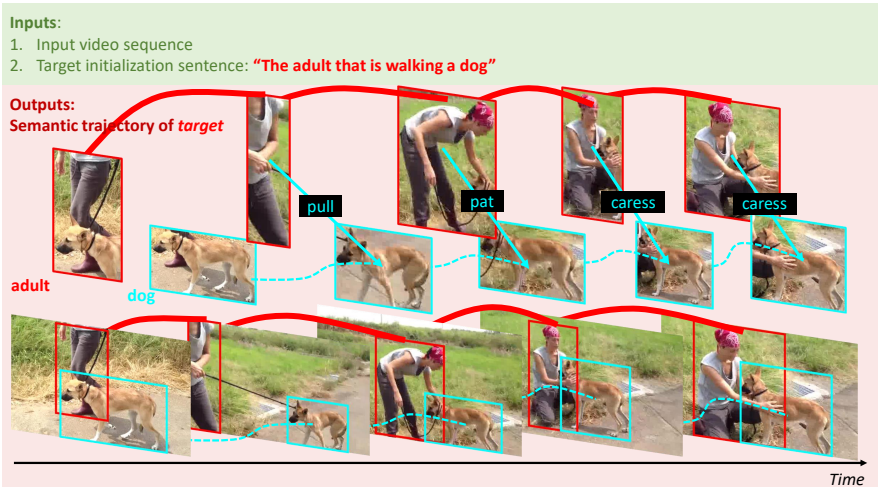

Fig. 1: In our Semantic Tracking, the model first localizes and tracks the target based on the user's input (*e.g.*, target initialization sentence), and returns the semantic trajectory of the target (*i.e.*, locations of the target, locations and classes of the surrounding affiliated objects of this target, and the interactions between the target and affiliated objects over time).

Tracking is defined in the dictionaries [1–3] as *following the movement of something* and *recording its development over time*, because we are *interested in finding out more about it*. In the computer vision domain, the existing tracking task mainly focuses on locating the target and obtaining its movement (*i.e.*, positional trajectory with only its locations over time). Given that we are interested to "find out more" about the target (*i.e.*, keeping track of both its movement and development of events), capturing just its positional trajectory is not enough to obtain a complete picture of target's movements, activities, interactions, and development over time. As such, this is not sufficient to fulfill many needs and requirements in various types of real-life scenarios.

Besides obtaining the target's positional trajectory, it can be very useful and beneficial for various scenarios if we can also obtain comprehensive details about the tracked target (*i.e.*, *who/what* this target is interacting with and *when*, *where*, and *how* this target is interacting with its surrounding objects or humans, which we collectively term as "*affiliated objects*"). Here, we use the term "*semantic trajectory*" to refer to the trajectory of the target with such rich and comprehensive logs detailing its every moment and development (*i.e.*, what is happening to the target at each timestep, including its location, the locations and classes of its surrounding affiliated objects, and its interactions with various affiliated objects). We thereby term this type of comprehensive tracking of the target's semantic trajectory as ***Semantic Tracking***. More formally, Semantic Tracking is defined as tracking a target based on the user's input and then, most importantly, capturing the rich semantic trajectory of this target. Using

Fig. 1 as an example, the user first inputs a target initialization sentence "The adult that is walking a dog" that describes the target to spatially locate it in the first frame and start tracking it. The model then obtains this target's semantic trajectory (*e.g.*, her locations and interactions with affiliated objects at each timestep such as pulling, patting, and then caressing the dog). Tracking such a rich semantic trajectory will facilitate comprehensive tracking of the target, and this will benefit a multitude of application scenarios.

One such application scenario that will benefit from Semantic Tracking is the retail sector. Knowing only where a customer is moving about in the store is not informative enough for retailers to capitalize on. In contrast, acquiring the customers' semantic trajectories in the store will offer richer and much more insightful information about their customers in terms of how they move in the store and interact with various products, as this reveals their shopping preferences that retailers are more interested in and can capitalize on. By analyzing various semantic trajectories of many different customers (*e.g.*, What products are they checking out in the store? In what sequence are they picking up the products?), retailers can derive deeper insights that can be subsequently used to customize and cater to their customers' needs such as ensuring that popular products are always in stock and certain products are strategically placed on specific aisles or shelves so that they can cross-sell or upsell these products [43]. Additionally, tracking a customer's semantic trajectory is useful in unmanned stores whereby customers can pick up and pay for the products as they conveniently walk out of the store. By enhancing the overall shopping experience, the store could benefit from greater business and sales. However, all of these would not have been possible with just the tracking of positional trajectories, but would be possible with Semantic Tracking (tracking of semantic trajectories).

Another scenario is crowd management and public safety settings. Knowing just where the specific targets (*e.g.*, VIP or suspect) are located may not be detailed enough to figure out what is happening on the ground. Instead, tracking various targets and obtaining rich logs containing vital details of the targets and their trajectories is especially crucial, particularly in time-critical life-or-death situations [28]. Once these details are acquired (*e.g.*, Where are they? What are they doing? Who are they approaching?), investigators can then rapidly analyze the semantic trajectory of the target in terms of how each target moves and interacts with others (*e.g.*, accomplice, hostage) or other objects (*e.g.*, putting down a suspicious bag).

There are other application scenarios whereby obtaining the semantic trajectories will be useful. In sports performance coaching, it can be overwhelming for the coach to *simultaneously* track each player and analyze how each player moves, acts, and interacts with objects and other players. With Semantic Tracking, each player's semantic trajectory can be easily tracked and analyzed so that detailed insights and personalized feedback can be provided to each individual and the team [63]. Besides, Semantic Tracking can also benefit live sports refereeing (*i.e.*, determining if there is any foul play) and live sports commentary highlighting the events that are happening in the targets' tracks.

In all of these real-life application scenarios and many more, Semantic Tracking can offer rich and comprehensive details about the target on top of its positional trajectory, hence allowing users to achieve a deeper and complete understanding about the tracked target and its trajectory, happenings and moments.

To achieve such comprehensive semantic tracking, a large-scale and comprehensive dataset would be required. However, currently there is no such dataset and benchmark. Existing tracking datasets [10, 11, 13, 17, 76] and tasks predict where the target is located over time but lack the rich details of its trajectory that will be useful for various domains and applications. Hence, this motivates us to create SemTrack, the *first-ever* large-scale, comprehensive, and challenging dataset for Semantic Tracking. Each video in our dataset contains annotations of the target's semantic trajectory across consecutive frames (*i.e.*, locations of the target, locations and classes of affiliated objects of this target, and the interactions between the target and affiliated objects over time). In addition, we also provide the target initialization sentence that describes the target so as to conveniently localize it and initialize tracking. Our dataset contains these annotations for a total of 6.7M frames from 6961 video samples, tracking 52 different types of interactions for 115 different types of objects in 12 different types of scenes ranging from indoor to outdoor scenes. We believe the large diversity in our dataset will attract and facilitate the community to develop, train, and evaluate various advanced methods to track and analyze the semantic trajectory of the target.

Besides constructing a new dataset SemTrack for Semantic Tracking, we also propose a simple and effective model (SemTracker) to localize and track the target and its interactions with affiliated objects, which we evaluate using our newly proposed evaluation metric (Semantic Tracking-mean Average Precision (ST-mAP)). Also, given that such comprehensive semantic tracking brings about new challenges (introduced in Sec. 4.2), we propose a meta-learning approach to better handle this challenge.

In summary, our contributions are as follows:

1. We create SemTrack, a new large, diverse, and challenging dataset to facilitate comprehensive tracking of the target's semantic trajectory.
2. We propose a simple and effective model (SemTracker) for Semantic Tracking. We also incorporate a meta-learning approach to better handle the challenges of this task.
3. We propose an evaluation metric (Semantic Tracking-mean Average Precision (ST-mAP)) for this Semantic Tracking task and use it to evaluate our model.

2 Related Works

Object tracking is important for a wide diversity of applications. We review some of the well-known tracking datasets in Tab. 1. Readers can also refer to survey papers [47, 53, 59, 80] on tracking datasets and methods.

Table 1: Key attributes of our SemTrack dataset as compared to existing tracking datasets. Our SemTrack dataset uniquely contains the semantic trajectory of the target with various types of information (*i.e.*, locations of the target, locations and classes of the affiliated objects of this target, and the interactions between the target and affiliated objects over time). In addition, as compared to other language-initialized tracking datasets, our dataset comprises more videos, frames, and target initialization sentences.

	Dataset	# Videos	# Frames	Duration	# Target initialization sentence	Locations of target	Locations of affilitated objects	Types of interaction
Bounding box Initialized	OTB2013 [71]	51	29K	16.4min	✗	✓	✗	✗
	OTB2015 [71]	100	59K	32.9min	✗	✓	✗	✗
	TC-128 [42]	128	55K	30.7min	✗	✓	✗	✗
	VOT2014 [33]	25	10K	5.7min	✗	✓	✗	✗
	VOT2017 [32]	60	21K	11.9min	✗	✓	✗	✗
	NUS-PRO [36]	365	135K	75.2min	✗	✓	✗	✗
	UAV123 [50]	123	113K	62.5min	✗	✓	✗	✗
	UAV20L [50]	20	59K	32.6min	✗	✓	✗	✗
	NfS [30]	100	383K	26.6min	✗	✓	✗	✗
	TrackingNet [51]	30,643	14.43M	140h	✗	✓	✗	✗
	GOT-10k [26]	10,000	1.5M	-	✗	✓	✗	✗
Language Initialized	OTB99-Lang [41]	99	59K	32.9min	99	✓	✗	✗
	ImageNet-Lang [41]	100	24K	13.3min	100	✓	✗	✗
	LaSOT [14]	1,400	3.52M	32.5h	1,400	✓	✗	✗
	TNL2K [67]	2,000	1.24M	11.5h	2,000	✓	✗	✗
	Web-UAV [75]	4,500	3.3M	28.9h	4,500	✓	✗	✗
	SemTrack (Ours)	6,961	6.7M	65.5h	9,760	✓	✓	52

2.1 Tracking Datasets

As the development of tracking models depends heavily on benchmark datasets [61], it is important to have large and comprehensive tracking datasets.

Existing object-tracking datasets generally only provide annotations to track the location information of the target(s). Depending on the number of targets tracked, these datasets can be categorized into Single Object Tracking (SOT) whereby only a single target is tracked [26, 30, 31, 36, 50, 51, 65, 71] and Multi-Object Tracking (MOT) whereby multiple targets are tracked [11, 12, 17, 35, 49, 69].

Many tracking datasets are designed to track specific targets, such as tracking humans (*e.g.*, MOTChallenge [11, 12, 35, 49]), vehicles (*e.g.*, UA-DETRAC [69], CityFlow [17]), vehicles and pedestrians (*e.g.*, [20, 60, 73]), and animals (*e.g.*, [76]). A number of them track different object categories/classes (*e.g.*, [10, 56]), while some track humans and objects in an indoor setting such as BEHAVE [6].

Among these datasets, a majority of them provide annotated bounding boxes to initialize the target for tracking. Some of the more recent datasets (*e.g.*, OTB99-Lang [41], ImageNet-Lang [41], LaSOT [14], and TNL2K [67]) provide annotated target initialization sentences instead. Such language-initialized tracking datasets offer a convenient and intuitive way to locate the target for tracking.

Our SemTrack dataset provides not only the description and location information of the target, it also contains the location information and classes of the affiliated objects (among a wide diversity of 115), and types of interactions between the target and its affiliated objects. Thus, our large and diverse SemTrack dataset shall facilitate the community to develop and evaluate various models for comprehensive semantic tracking of the target.

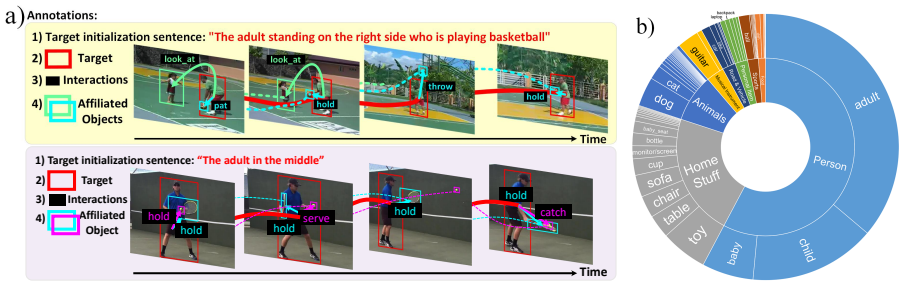

Fig. 2: **(a)** For each video in our SemTrack dataset, we annotate the target initialization sentence that describes the target. We also annotate the bounding box of the target, the bounding box of the affiliated objects of this target, and the interactions between the target and affiliated objects over time. All the comprehensive information forms the semantic trajectory of this target. **(b)** Distribution of object classes in terms of number of instances.

2.2 Object Tracking Methods

With the rapid development and popularization of tracking-related applications, the research direction of developing tracking algorithms with greater robustness and accuracy has attracted widespread attention [53].

Existing tracking tasks and methods mainly only focus on tracking the location information of the target. Many tracking methods consist of a two-stage model architecture. For the first stage, to identify and localize the target, they either require the input of the target's bounding box in the first frame or employ pre-trained object detectors (*e.g.*, Faster-RCNN [55] and YOLO series [8, 19, 37, 54]) to obtain the target's bounding box. More recently, there are some methods [9, 15, 16, 23, 40, 66, 67] that use target initialization sentence to identify and localize the target instead. These are known as language-initialized tracking methods. For the second stage, once the target is localized, the target's bounding boxes across frames (*i.e.*, trajectory) are associated using motion models or appearance and re-identification models. Motion models employ sequential analysis tools such as particle filter [4], Kalman filter (*e.g.*, SORT [5], ByteTrack [78]). In contrast, appearance models and re-identification models use deep appearance similarities (*e.g.*, DeepSort [70], JDE [68], FairMOT [79], and QDTrack [18]). Some other tracking methods do not use a two-stage model architecture but use transformer-based architecture instead (*e.g.*, TransTrack [62], TrackFormer [48], MOTR [74]).

Different from all these methods, we aim to track the target and obtain a more comprehensive information about the target (*i.e.*, its semantic trajectory). Therefore, to address such a challenging Semantic Tracking task, we develop a new method, SemTracker, to obtain its semantic trajectory. We believe that, together with our proposed dataset SemTrack, our SemTracker method can promote further development of more intelligent Semantic Tracking algorithms.

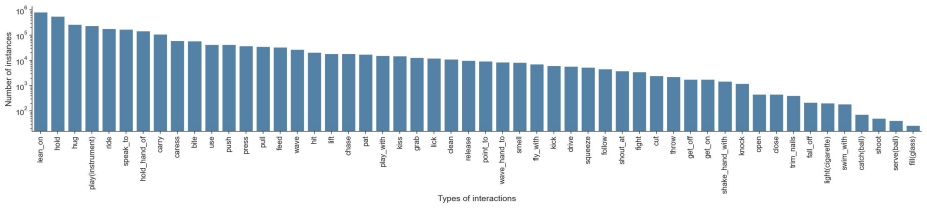

Fig. 3: Distribution of different interaction classes in terms of number of instances.

3 Our Proposed SemTrack Dataset

Our SemTrack dataset is the *first* large-scale Semantic Tracking dataset. It contains 6961 videos in total, covering a wide range of 52 different interaction classes with 115 different object classes spanning 10 different supercategories in 12 types of different scenes, ranging from indoor and outdoor scenes. To construct the SemTrack dataset, the videos are selected from various publicly available datasets capturing various types of activities and interactions: YFCC100M [64], TAO [10], ImageNet-VidVRD [58], VidOR [57], HACS [81], AVA [21], GOT-10K [26], and ILSVRC2016 [56]. 10 volunteers meticulously annotated the rich annotations of both the target initialization sentence and bounding box to localize the target, as well as detailed semantic trajectory of the target for each video over a period of 7 months, with the help of the existing annotations of the above datasets. In other words, besides inputting text, users can also input bounding box to initialize target when using our dataset, and this shall be useful for different application scenarios. Tab. 1 summarizes the distinctive features of our SemTrack dataset. Illustration examples of videos and annotations in our SemTrack dataset are shown in Fig. 2a. More details of the dataset are provided in Supplementary.

Different from existing tracking datasets, our SemTrack dataset contains annotations of a **wide range of 52 interactions of target with affiliated objects (115 different types), across 12 different types of scenes**. These interactions (Fig. 3) include various types of daily interactions of an individual ranging from movement-related and daily activities (*e.g.*, pushing, driving) to social interactions and display of affection (*e.g.*, waving, hugging). The scenes comprise both indoor scenes (*e.g.*, cooking in the kitchen) and outdoor scenes (*e.g.*, car racing in the desert, cycling in the mountain), and also interactions with animals (*e.g.*, bullfighting in the stadium, horse-riding in an Olympics venue *etc.*) and unique activities such as water sports in the river and the sea (*e.g.*, water rafting, scuba diving). Having a diverse representation of interactions and objects in a dataset that reflect the diversity in the real world is useful in developing and evaluating intelligent, robust, and generalizable tracking models that are of great practical value to handle various complex real-world scenarios. Besides this, our SemTrack dataset also contains the following characteristics:

Diverse object classes and unique target initialization sentence to localize and initialize the target. The diversity in the real world is reflected in

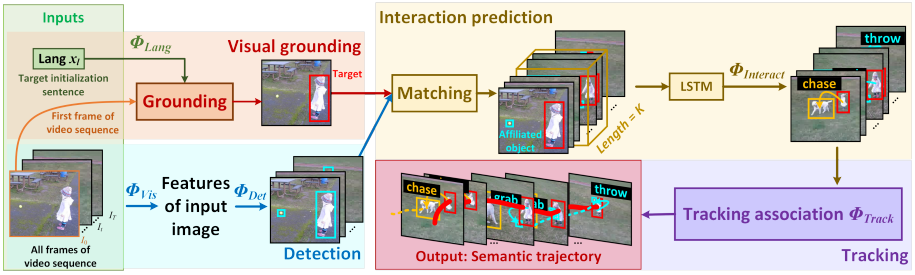

Fig. 4: Architecture of SemTracker model which consists of four modules: visual grounding, object detection, interaction prediction, and object tracking. Using the first frame of a video sequence and only the user's language description of the target (*i.e.*, target initialization sentence x_l), the visual grounding module is used to localize the target (*i.e.*, automatically generate the red bounding boxes). For each frame of the video sequence, objects are detected using the detection module (including a feature extractor Φ_{Vis} and a detection head Φ_{Det}). The location of the target in each frame is obtained by matching between patches of the detected objects. The interaction prediction module takes the information from multiple frames and adopts an LSTM structure to predict the interactions between the target and its affiliated objects. The locations of both the target and its affiliated objects across consecutive frames are then associated in the tracking module Φ_{Track}. Details about architecture are provided in Supplementary.

the diverse representation of humans and objects (Fig. 2 b), which is important in developing generalizable and robust models to localize and track targets. In total, there are 115 different types of objects, including humans, animals, vehicles, and home stuff. The various attributes (*e.g.*, color, size) of the target is uniquely and manually described so that it can be precisely localized to initialize tracking.

Various scenes, views and camera angles. Taken at different vantage points (*e.g.*, from Unmanned Aerial Vehicle, surveillance camera, panning camera) and in 12 different types of scenes (*e.g.*, indoor scenes, such as stage performances; outdoor scenes; sports scenes; and scenes with animals, including zoo and safari), our SemTrack dataset comprehensively covers a wide range of complex and challenging scenarios that are necessary to develop and evaluate robust and accurate tracking models for the real world.

Complex illumination, occlusions, and various challenging conditions. SemTrack includes videos that are taken under complex illumination and lighting conditions (*e.g.*, dark or low light conditions), including illumination changes (*e.g.*, spotlight and differently coloured stage light, sudden change of lighting), partial occlusion, viewpoint changes (including zooming, panning), abrupt slow-fast frame transition, continuous camera shaking, background blur, and soft object focus. All of these challenging scenarios shall be able to encourage the community to develop advanced strategies to overcome these practical challenges.

4 Our Proposed Semantic Tracking Method

Existing object tracking tasks and models output the positional trajectory of the target in a video sequence. Different from all these methods, we aim to track the target and its semantic trajectory. More precisely, this model first takes in a video sequence $\{I_0, I_1, ..., I_T\}$ and an input target initialization sentence x_l describing the target in I_0 (*e.g.*, to track "the adult standing on the right side who is playing basketball" as shown in Fig. 2a). The model then localizes this target and outputs its semantic trajectory.

4.1 Model Architecture

To perform Semantic Tracking, we propose a new model architecture, *i.e.*, SemTracker, which consists of a visual grounding module, an object detection module, an interaction prediction module, and a multiple object tracking module. Fig. 4 shows the model architecture of SemTracker.

Visual grounding module. The visual grounding module is used to locate the target in the first frame of the video sequence. We follow TNL2K [67] and SNLT [16] and use a pre-trained language model Φ_{Lang} (*e.g.*, BERT [29]) to extract the language embedding from the input target initialization sentence x_l that describes the target object. The language embedding and the first frame of the video are then used to localize the target.

Object detection module. The object detection module is utilized to detect the affiliated objects that potentially have interactions with the target. In this module, we follow YOLOX [19] and use the Feature Pyramid Network (FPN) [44] backbone Φ_{Vis} to extract high-dimensional features f_t from each input frame I_t (t represents the current frame index). These features are then fed into an object detection head Φ_{Det}, which outputs the detected bounding boxes $b_{t,i}$ and object classes $c_{t,i}$ for all the candidate affiliated objects in the t-th frame (i represents the index of objects in the frame). To obtain more informative representations of each detected object, we use RoI Align [25] Φ_{RoI} to extract intermediate features $f_{t,i}$ that correspond to the detected bounding boxes.

Interaction prediction module. The interaction prediction module predicts the interactions between the target and its affiliated objects. At frame I_t, the target is determined by matching the output image patch of the visual grounding module with the detected objects. The remaining detected objects are designated as candidate affiliated objects.

Both the features of the target and features of the candidate affiliated objects are fed into a Convolutional Neural Network (CNN) to predict the type of interaction between each pair. Intuitively, information in the preceding frames can be valuable in predicting the interaction in the subsequent frames. Hence, inspired by such observation, we use Long-Short Term Memory (LSTM) to take intermediate features from multiple frames as input to predict the interaction. Specifically, we concatenate the features of target to the features of the candidate affiliated objects in past K frames and then feed them to an LSTM network. The LSTM network takes the concatenated features and produces a hidden

state, which is then fed into a fully connected (FC) layer $\Phi_{Interact}$ to predict the interaction between objects in the current frame and determine which of the candidate affiliated objects have interactions with the target. Those candidate affiliated objects that have interactions with the target are now termed as the affiliated objects.

Multiple object tracking module. The purpose of the multiple object tracking module is to obtain the positional trajectories of both the target and its affiliated objects. We adopt ByteTrack [78] for this module. Given the object detection results of two consecutive frames (including bounding box $b_{t-1,i}$, $b_{t,i}$, object class $c_{t-1,i}$, $c_{t,i}$, and features $f_{t-1,i}$, $f_{t,i}$), the object tracker Φ_{Track} matches the detected objects across consecutive frames to recognize the identity of each object and obtain the corresponding trajectories.

In summary, our SemTracker method takes in a video sequence and a target initialization sentence as its input, and outputs the semantic trajectory of the target.

4.2 Meta-learning of Semantic Tracking

For Semantic Tracking task, the model localizes and tracks the target and its interactions with affiliated objects. Its performance may be influenced by the data distribution of the data used during training [27]. In the real world, there can be various scenarios with diverse data distributions that do not always match the data distribution found in the training dataset. This can lead to a drop in performance when the model is tested on different data distributions [7, 77]. For example, if the composition of target-*interaction*-affiliated object such as "adult *drink from* bottle" occurs more frequently than "adult *clean* bottle" in the training set, then the model trained with this compositional bias will tend to directly predict "*drink from*" when tracking the interaction between the target "adult" and the affiliated object "bottle" during testing, even when the interaction is actually "*clean*". Hence, by considering the different data distributions of both the object classes and types of interactions, we intend to optimize the model's performance by improving its generalization ability to diverse data distributions.

Meta Optimization Scheme. Meta-learning methods [22, 38, 52, 72] have shown to be effective in improving the model's generalization ability to new tasks, domains, *etc*. by incorporating virtual testing during training. Inspired by such *learning-to-learn* methods, we propose to improve the model's generalization ability to different data distributions with meta-learning scheme. Specifically, we first split the training set into a *virtual training set* and a set of *virtual testing set*, where each virtual testing set has different data distribution to the virtual training set. For example, if the virtual training set contains more samples of "adult *drink from* bottle", the virtual testing set can contain more samples of "adult *clean* bottle". During model training, the model's parameters are updated using the *virtual training set* and tested using the *virtual testing set*. The loss on the *virtual testing set* can offer a clear feedback for the model's generalization ability and guide the model to generalize to different data distributions. Hence, through many iterations of mimicking different testing data distributions

w.r.t. both the object classes and interaction types during training, the model is optimized to be more robust to different data distributions.

More concretely, we first split the training set D^{train} into a support set D^s (*i.e.*, *virtual training set)* and N query sets $\{D_n^q\}_{n=1}^N$ (*i.e.*, *virtual testing set*), where each of the query set has a different distribution from each other with respect to (w.r.t.) both the object classes and the types of interactions. During training, we first perform the *virtual training* step and virtually update the parameters θ of the model on the support set D^s with learning rate α via conventional gradient descent:

$$\theta' = \theta - \alpha \bigtriangledown_\theta \mathcal{L}(D^s;\ \theta) \tag{1}$$

where α denotes the learning rate of the virtual training. After that, we perform *virtual testing* where the virtually updated parameters θ' are evaluated on the query sets $\{D_n^q\}_{n=1}^N$ to test its generalization ability with diverse distributions w.r.t. object classes and types of interactions. For each query set, we compute the loss on the query set $\mathcal{L}(D_n^q;\theta')$ with the virtually updated parameters θ'. Intuitively, a model with good generalization ability towards various data distributions would have low loss values on all query sets. Thus, the loss can be used as a feedback and thus provide guidance for the model to generalize towards different data distributions. After the virtual testing, we optimize the parameters θ with *meta optimization* such that after θ is updated on the support set, the model can also generalize well (*i.e.*, obtain low $\mathcal{L}(D_n^q;\theta')$) on the N query sets. The objective of the optimization can be formulated as:

$$\begin{aligned} &\min_\theta [\mathcal{L}(D^s;\ \theta) + \sum_{n=1}^{N} \mathcal{L}(D_n^q;\ \theta')] \\ = &\min_\theta [\mathcal{L}(D^s;\ \theta) + \sum_{n=1}^{N} \mathcal{L}(D_n^q;\ \theta - \alpha \bigtriangledown_\theta \mathcal{L}(D^s;\ \theta))] \end{aligned} \tag{2}$$

where the first term indicates the general training performance, and the second term reflects the generalization ability of the model with the virtually updated parameters θ' to different data distributions. In all, we update the model parameters θ as:

$$\theta \leftarrow \theta - \beta \bigtriangledown_\theta [\mathcal{L}(D^s;\ \theta) + \sum_{n=1}^{N} \mathcal{L}(D_n^q;\ \theta - \alpha \bigtriangledown_\theta \mathcal{L}(D^s;\ \theta))] \tag{3}$$

where β denotes the learning rate of the overall optimization. By mimicking different data distribution in the query sets and perform virtual training and virtual testing, the model can be pushed to generalize towards diverse data distributions.

Construction of the Support Set and Query Sets. The protocol for such a dataset split is as follows: Initially, a part of the training set is randomly extracted from D^{train} to form the support set D^s. The frequencies of object

classes and types of interactions in D^s are calculated to obtain its data distribution. Thereafter, we generate N query sets sequentially from the remaining part of D^{train}. To maximize the differences between the m-th query set D_m^q and previously constructed sets (*i.e.*, the support set D^s and the previous $(m-1)$ query sets $\{D_n^q\}_{n=1}^{m-1}$), we select videos that maximize the sum of pairwise KL divergence between the data distribution of current query set D_m^q and each of the existing sets. In this way, we can construct a collection of query sets that have diverse data distributions w.r.t. both object classes and types of interactions.

4.3 Our Proposed Evaluation Metric

In this work, we aim to obtain and evaluate the target's semantic trajectory. However, existing metrics of tracking tasks, such as Track-mAP [10] and HOTA [46], only evaluate the estimation accuracy of the positional trajectory (*i.e.*, position of objects across multiple frames). Such metrics are not sufficient to evaluate the model's overall ability in predicting all the rich information in the semantic trajectories. To evaluate the performance of Semantic Tracking methods, we propose a new evaluation metric Semantic Tracking-mAP (ST-mAP).

Semantic Tracking-mAP. To calculate the accuracy of the predicted semantic trajectory, we consider all the following criteria: (a) correctly identify the target and correctly estimate its positional trajectory; (b) correctly identify the affiliated (interacted) objects and correctly estimate their positional trajectories; and (c) correctly predict the types of interactions between the target and its affiliated objects for each frame.

In order to effectively evaluate all the aspects, we adapt the Track-mAP metric [10] to develop a new metric, Semantic Tracking-mAP (ST-mAP). The ground truth consists of the positional trajectories of the target and an affiliated object, and the interactions between them. Similarly, the prediction is composed of the positional trajectories of the target and an affiliated object, and the interactions between them. We adopt the Hungarian algorithm [34] to match the predicted tracks and the ground truth tracks. We follow the classical mean average precision (mAP) definition [46] to calculate our ST-mAP as Equation 4.

$$\text{ST-mAP} = \frac{1}{N_{obj}} \sum_{n=1}^{N_{obj}} \text{ST-AP}_n \tag{4}$$

where N_{obj} refers to the total number of object classes. For each object class, the average precision (AP) of Semantic Tracking is calculated as indicated by Equation 5.

$$\text{ST-AP} = \frac{1}{11} \sum_{r \in 0, 0.1, \ldots, 1.0} Pr(r) \tag{5}$$

where $Pr(r)$ is an interpolated precision that takes the maximum precision over all recalls greater than r. More details about ST-mAP are provided in Supplementary.

5 Experimental Results

We perform experiments on our proposed SemTrack dataset using state-of-the-art object detection (*e.g.*, YOLOX [19]) and tracking methods (*e.g.*, SORT [5], DeepSort [70], and ByteTrack [78]). The dataset is split into training set, validation set, and testing set. In this paper, we evaluate the models on a subset of the testing set as representatives especially focusing on human (e.g., adult and baby). We release the split information in the dataset. More implementation details are in the Supplementary.

5.1 Results of SemTracker Method

The object detector used in our SemTracker model is YOLOX [19] with 3 versions (*i.e.*, YOLOX-M, YOLOX-L, and YOLOX-X). The trackers used are SORT [5], DeepSORT [70], and ByteTrack [78]. Tab. 2 shows the experimental results of our framework using different detectors and trackers. The experimental results demonstrate that the proposed framework exhibits the best performance when utilizing the YOLOX-X detector and the ByteTrack tracker.

Table 2: Comparison of different models for SemTracker.

Setting		Evaluation Metrics		
Detector	Tracker	Track-mAP ↑	HOTA ↑	ST-mAP ↑
YOLOX-M	Sort [5]	1.962	10.727	1.043
YOLOX-M	DeepSort [70]	1.6489	11.341	1.364
YOLOX-M	ByteTrack [78]	3.911	12.093	2.546
YOLOX-L	Sort [5]	2.219	11.509	1.646
YOLOX-L	DeepSort [70]	2.112	11.991	1.352
YOLOX-L	ByteTrack [78]	4.076	13.051	2.938
YOLOX-X	Sort [5]	2.039	11.746	1.241
YOLOX-X	DeepSort [70]	1.978	11.969	1.230
YOLOX-X	ByteTrack [78]	**5.875**	**13.436**	**4.625**

Similar to the observations seen in the results of other previous tracking tasks [10, 78], the performance of object tracking is affected by the performance of the object detector. To investigate how the object detector affects the overall performance of Semantic Tracking task, we provide models with the ground truth bounding box coordinates and object classes in the first frame (*i.e.*, in place of an object detector). The evaluation results are shown in Tab. 3. It is observed that by providing the model with the ground truth bounding box of the target in the first frame, the model's performance on each evaluation metric is increased (by comparing Tab. 3 with Tab. 2). Therefore, using a more accurate object detector can help improve the overall performance of Semantic Tracking.

Table 3: Results of SemTracker that is provided with the ground truth bounding box of the target in the first frame.

Tracker	Evaluation Metrics		
	Track-mAP ↑	HOTA ↑	ST-mAP ↑
Sort [5]	3.394	12.544	2.461
DeepSort [70]	3.704	13.119	3.408
ByteTrack [78]	6.521	14.429	4.792

5.2 Results of SemTracker with Meta-learning

To demonstrate the generalization ability of our proposed method, we conduct experiments with meta-learning, using the best settings obtained from using YOLOX-X and ByteTrack as shown in Tab. 2.

We implement other classical methods for handling data distribution such as the reweighting [24] and the focal loss [45], as well as the recent method (GCLLoss [39]) for comparison. In Tab. 4, we observe that our proposed meta-learning method outperforms our baseline method (SemTracker without meta-learning) and other methods. Specifically, our method achieves an obvious improvement over our baseline method. This improvement demonstrates the effectiveness of meta-learning in improving the performance of SemTracker. Additional experimental results are provided in Supplementary.

Table 4: Results of SemTracker with and without (w/o) meta-learning.

Method	Evaluation Metrics		
	Track-mAP ↑	HOTA ↑	ST-mAP ↑
SemTracker w/o Meta-learning	5.875	13.436	4.625
SemTracker + Reweighting [24]	5.950	13.868	4.714
SemTracker + Focal Loss [45]	7.934	17.492	6.869
SemTracker + GCLLoss [39]	9.664	18.831	8.211
SemTracker + Meta-learning	**10.656**	**18.854**	**8.668**

6 Conclusion

Obtaining detailed information about the target and its semantic trajectory is important for numerous applications. We create SemTrack, the first large-scale, comprehensive, and challenging dataset, which consists of annotations of the target's semantic trajectory. We also propose SemTracker, a simple and effective method, and incorporate a meta-learning approach to better handle the challenges of this task. We also propose a new evaluation metric, ST-mAP, and use it to evaluate our model. We believe that our work will inspire the community to develop, adapt, and evaluate various types of approaches for the Semantic Tracking task.

References

1. Cambridge dictionary. https://dictionary.cambridge.org/dictionary/english/track,
2. Collins dictionary. https://www.collinsdictionary.com/dictionary/english/track
3. Oxford learner's dictionary. https://www.oxfordlearnersdictionaries.com/definition/american_english/track_2
4. Arulampalam, M.S., Maskell, S., Gordon, N., Clapp, T.: A tutorial on particle filters for online nonlinear/non-gaussian bayesian tracking. IEEE Transactions on signal processing **50**(2), 174–188 (2002)
5. Bewley, A., Ge, Z., Ott, L., Ramos, F., Upcroft, B.: Simple online and realtime tracking. In: 2016 IEEE International Conference on Image Processing (ICIP). pp. 3464–3468 (2016). https://doi.org/10.1109/ICIP.2016.7533003
6. Bhatnagar, B.L., Xie, X., Petrov, I., Sminchisescu, C., Theobalt, C., Pons-Moll, G.: Behave: Dataset and method for tracking human object interactions. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE (jun 2022)
7. Bickel, S., Brückner, M., Scheffer, T.: Discriminative learning for differing training and test distributions. In: Proceedings of the 24th International Conference on Machine Learning. p. 81–88. ICML '07, Association for Computing Machinery, New York, NY, USA (2007). https://doi.org/10.1145/1273496.1273507, https://doi.org/10.1145/1273496.1273507
8. Bochkovskiy, A., Wang, C.Y., Liao, H.Y.M.: Yolov4: Optimal speed and accuracy of object detection. arXiv preprint arXiv:2004.10934 (2020)
9. Dai, Y., Ma, F., Hu, W., Zhang, F.: Spgc: Shape-prior based generated content data augmentation for remote sensing object detection. IEEE Transactions on Geoscience and Remote Sensing (2024)
10. Dave, A., Khurana, T., Tokmakov, P., Schmid, C., Ramanan, D.: TAO: A Large-Scale Benchmark for Tracking Any Object, pp. 436–454 (10 2020). https://doi.org/10.1007/978-3-030-58558-7_26
11. Dendorfer, P., Osep, A., Milan, A., Schindler, K., Cremers, D., Reid, I., Roth, S., Leal-Taixé, L.: Motchallenge: A benchmark for single-camera multiple target tracking. International Journal of Computer Vision **129**, 845–881 (2021)
12. Dendorfer, P., Rezatofighi, H., Milan, A., Shi, J., Cremers, D., Reid, I., Roth, S., Schindler, K., Leal-Taixé, L.: Mot20: A benchmark for multi object tracking in crowded scenes. arXiv preprint arXiv:2003.09003 (2020)
13. Doering, A., Chen, D., Zhang, S., Schiele, B., Gall, J.: Posetrack21: A dataset for person search, multi-object tracking and multi-person pose tracking. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20963–20972 (2022)
14. Fan, H., Lin, L., Yang, F., Chu, P., Deng, G., Yu, S., Bai, H., Xu, Y., Liao, C., Ling, H.: Lasot: A high-quality benchmark for large-scale single object tracking. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5374–5383 (2019)
15. Feng, Q., Ablavsky, V., Bai, Q., Li, G., Sclaroff, S.: Real-time visual object tracking with natural language description. pp. 689–698 (03 2020). https://doi.org/10.1109/WACV45572.2020.9093425
16. Feng, Q., Ablavsky, V., Bai, Q., Sclaroff, S.: Siamese natural language tracker: Tracking by natural language descriptions with siamese trackers. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR) (2021)

17. Feng, Q., Ablavsky, V., Sclaroff, S.: Cityflow-nl: Tracking and retrieval of vehicles at city scale by natural language descriptions. arXiv preprint arXiv:2101.04741 (2021)
18. Fischer, T., Pang, J., Huang, T.E., Qiu, L., Chen, H., Darrell, T., Yu, F.: Qdtrack: Quasi-dense similarity learning for appearance-only multiple object tracking. arXiv preprint arXiv:2210.06984 (2022)
19. Ge, Z., Liu, S., Wang, F., Li, Z., Sun, J.: Yolox: Exceeding yolo series in 2021. arXiv preprint arXiv:2107.08430 (2021)
20. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? the kitti vision benchmark suite. In: Conference on Computer Vision and Pattern Recognition (CVPR) (2012)
21. Gu, C., Sun, C., Ross, D.A., Vondrick, C., Pantofaru, C., Li, Y., Vijayanarasimhan, S., Toderici, G., Ricco, S., Sukthankar, R., Schmid, C., Malik, J.: Ava: A video dataset of spatio-temporally localized atomic visual actions. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6047–6056 (2018). https://doi.org/10.1109/CVPR.2018.00633
22. Guo, J., Zhu, X., Zhao, C., Cao, D., Lei, Z., Li, S.Z.: Learning meta face recognition in unseen domains. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6163–6172 (2020)
23. Guo, M., Zhang, Z., Fan, H., Jing, L.: Divert more attention to vision-language tracking. arXiv preprint arXiv:2207.01076 (2022)
24. He, H., Garcia, E.A.: Learning from imbalanced data. IEEE Transactions on knowledge and data engineering **21**(9), 1263–1284 (2009)
25. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask r-cnn. In: Proceedings of the IEEE international conference on computer vision. pp. 2961–2969 (2017)
26. Huang, L., Zhao, X., Huang, K.: Got-10k: A large high-diversity benchmark for generic object tracking in the wild. IEEE Transactions on Pattern Analysis and Machine Intelligence (2019)
27. Johnson, J.M., Khoshgoftaar, T.M.: Survey on deep learning with class imbalance. Journal of Big Data **6**(1), 1–54 (2019)
28. Karbalaie, A., Abtahi, F., Sjöström, M.: Event detection in surveillance videos: A review. Multimedia Tools and Applications pp. 1–39 (2022)
29. Kenton, J.D.M.W.C., Toutanova, L.K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of naacL-HLT. pp. 4171–4186 (2019)
30. Kiani Galoogahi, H., Fagg, A., Huang, C., Ramanan, D., Lucey, S.: Need for speed: A benchmark for higher frame rate object tracking. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 1125–1134 (2017)
31. Kristan, M., et. al.: The seventh visual object tracking vot2019 challenge results. In: 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW). pp. 2206–2241 (2019). https://doi.org/10.1109/ICCVW.2019.00276
32. Kristan, M., Matas, J., Leonardis, A., Vojir, T., Pflugfelder, R., Fernandez, G., Nebehay, G., Porikli, F., Čehovin, L.: A novel performance evaluation methodology for single-target trackers. IEEE Transactions on Pattern Analysis and Machine Intelligence **38**(11), 2137–2155 (Nov 2016). https://doi.org/10.1109/TPAMI.2016.2516982
33. Kristan, M., Pflugfelder, R., Leonardis, A., et al.: The visual object tracking vot2014 challenge results. In: Agapito, L., Bronstein, M.M., Rother, C. (eds.) Computer Vision - ECCV 2014 Workshops. pp. 191–217. Springer International Publishing, Cham (2015)

34. Kuhn, H.W.: The hungarian method for the assignment problem. Naval research logistics quarterly **2**(1-2), 83–97 (1955)
35. Leal-Taixé, L., Milan, A., Reid, I., Roth, S., Schindler, K.: Motchallenge 2015: Towards a benchmark for multi-target tracking. arXiv preprint arXiv:1504.01942 (2015)
36. Li, A., Lin, M., Wu, Y., Yang, M., Yan, S.: NUS-PRO: A New Visual Tracking Challenge. IEEE Transactions on Pattern Analysis and Machine Intelligence **38**(2), 335–349 (2016)
37. Li, C., Li, L., Jiang, H., Weng, K., Geng, Y., Li, L., Ke, Z., Li, Q., Cheng, M., Nie, W., et al.: Yolov6: A single-stage object detection framework for industrial applications. arXiv preprint arXiv:2209.02976 (2022)
38. Li, D., Yang, Y., Song, Y.Z., Hospedales, T.: Learning to generalize: Meta-learning for domain generalization. In: Proceedings of the AAAI conference on artificial intelligence. vol. 32 (2018)
39. Li, M., Cheung, Y.m., Lu, Y.: Long-tailed visual recognition via gaussian clouded logit adjustment. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6929–6938 (2022)
40. Li, Y., Yu, J., Cai, Z., Pan, Y.: Cross-modal target retrieval for tracking by natural language. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4931–4940 (2022)
41. Li, Z., Tao, R., Gavves, E., Snoek, C.G., Smeulders, A.W.: Tracking by natural language specification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6495–6503 (2017)
42. Liang, P., Blasch, E., Ling, H.: Encoding color information for visual tracking: Algorithms and benchmark. IEEE Transactions on Image Processing **24**(12), 5630–5644 (2015)
43. Liciotti, D., Frontoni, E., Mancini, A., Zingaretti, P.: Pervasive system for consumer behaviour analysis in retail environments. In: Video Analytics. Face and Facial Expression Recognition and Audience Measurement: Third International Workshop, VAAM 2016, and Second International Workshop, FFER 2016, Cancun, Mexico, December 4, 2016, Revised Selected Papers 2. pp. 12–23. Springer (2017)
44. Lin, T.Y., Dollár, P., Girshick, R., He, K., Hariharan, B., Belongie, S.: Feature pyramid networks for object detection. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 936–944 (2017). https://doi.org/10.1109/CVPR.2017.106
45. Lin, T.Y., Goyal, P., Girshick, R., He, K., Dollár, P.: Focal loss for dense object detection. In: Proceedings of the IEEE international conference on computer vision. pp. 2980–2988 (2017)
46. Luiten, J., Osep, A., Dendorfer, P., Torr, P., Geiger, A., Leal-Taixé, L., Leibe, B.: Hota: A higher order metric for evaluating multi-object tracking. International journal of computer vision **129**, 548–578 (2021)
47. Luo, W., Xing, J., Milan, A., Zhang, X., Liu, W., Kim, T.K.: Multiple object tracking: A literature review. Artificial intelligence **293**, 103448 (2021)
48. Meinhardt, T., Kirillov, A., Leal-Taixe, L., Feichtenhofer, C.: Trackformer: Multi-object tracking with transformers. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8844–8854 (2022)
49. Milan, A., Leal-Taixé, L., Reid, I., Roth, S., Schindler, K.: Mot16: A benchmark for multi-object tracking. arXiv preprint arXiv:1603.00831 (2016)
50. Mueller, M., Smith, N., Ghanem, B.: A benchmark and simulator for uav tracking. In: European conference on computer vision. pp. 445–461. Springer (2016)

51. Muller, M., Bibi, A., Giancola, S., Alsubaihi, S., Ghanem, B.: Trackingnet: A large-scale dataset and benchmark for object tracking in the wild. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 300–317 (2018)
52. Qu, H., Li, Y., Foo, L.G., Kuen, J., Gu, J., Liu, J.: Improving the reliability for confidence estimation. In: European Conference on Computer Vision. pp. 391–408. Springer (2022)
53. Rakai, L., Song, H., Sun, S., Zhang, W., Yang, Y.: Data association in multiple object tracking: A survey of recent techniques. Expert Systems with Applications **192**, 116300 (2022)
54. Redmon, J., Farhadi, A.: Yolov3: An incremental improvement. arXiv preprint arXiv:1804.02767 (2018)
55. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in neural information processing systems **28** (2015)
56. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L.: ImageNet Large Scale Visual Recognition Challenge. International Journal of Computer Vision (IJCV) **115**(3), 211–252 (2015). https://doi.org/10.1007/s11263-015-0816-y
57. Shang, X., Di, D., Xiao, J., Cao, Y., Yang, X., Chua, T.S.: Annotating objects and relations in user-generated videos. In: Proceedings of the 2019 on International Conference on Multimedia Retrieval. pp. 279–287. ACM (2019)
58. Shang, X., Ren, T., Guo, J., Zhang, H., Chua, T.S.: Video visual relation detection. In: ACM International Conference on Multimedia. Mountain View, CA USA (October 2017)
59. Soleimanitaleb, Z., Keyvanrad, M.A.: Single object tracking: A survey of methods, datasets, and evaluation metrics. arXiv preprint arXiv:2201.13066 (2022)
60. Sun, P., Kretzschmar, H., Dotiwalla, X., Chouard, A., Patnaik, V., Tsui, P., Guo, J., Zhou, Y., Chai, Y., Caine, B., et al.: Scalability in perception for autonomous driving: Waymo open dataset. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2446–2454 (2020)
61. Sun, P., Cao, J., Jiang, Y., Yuan, Z., Bai, S., Kitani, K., Luo, P.: Dancetrack: Multi-object tracking in uniform appearance and diverse motion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20993–21002 (2022)
62. Sun, P., Cao, J., Jiang, Y., Zhang, R., Xie, E., Yuan, Z., Wang, C., Luo, P.: Transtrack: Multiple object tracking with transformer. arXiv preprint arXiv:2012.15460 (2020)
63. Thomas, G., Gade, R., Moeslund, T.B., Carr, P., Hilton, A.: Computer vision for sports: Current applications and research topics. Computer Vision and Image Understanding **159**, 3–18 (2017)
64. Thomee, B., Shamma, D.A., Friedland, G., Elizalde, B., Ni, K., Poland, D., Borth, D., Li, L.J.: Yfcc100m: The new data in multimedia research. Commun. ACM **59**(2), 64–73 (jan 2016). https://doi.org/10.1145/2812802, https://doi.org/10.1145/2812802
65. Valmadre, J., Bertinetto, L., Henriques, J.F., Tao, R., Vedaldi, A., Smeulders, A.W., Torr, P.H., Gavves, E.: Long-term tracking in the wild: A benchmark. In: Proceedings of the European conference on computer vision (ECCV). pp. 670–685 (2018)
66. Wang, X., Li, C., Yang, R., Zhang, T., Tang, J., Luo, B.: Describe and attend to track: Learning natural language guided structural representation and visual attention for object tracking. arXiv preprint arXiv:1811.10014 (2018)

67. Wang, X., Shu, X., Zhang, Z., Jiang, B., Wang, Y., Tian, Y., Wu, F.: Towards more flexible and accurate object tracking with natural language: Algorithms and benchmark. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 13763–13773 (June 2021)
68. Wang, Z., Zheng, L., Liu, Y., Li, Y., Wang, S.: Towards real-time multi-object tracking. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16. pp. 107–122. Springer (2020)
69. Wen, L., Du, D., Cai, Z., Lei, Z., Chang, M.C., Qi, H., Lim, J., Yang, M.H., Lyu, S.: Ua-detrac: A new benchmark and protocol for multi-object detection and tracking. Computer Vision and Image Understanding **193**, 102907 (2020)
70. Wojke, N., Bewley, A.: Deep cosine metric learning for person re-identification. In: 2018 IEEE Winter Conference on Applications of Computer Vision (WACV). pp. 748–756. IEEE (2018). https://doi.org/10.1109/WACV.2018.00087
71. Wu, Y., Lim, J., Yang, M.H.: Object tracking benchmark. IEEE Transactions on Pattern Analysis and Machine Intelligence **37**(9), 1834–1848 (2015)
72. Xu, L., Qu, H., Kuen, J., Gu, J., Liu, J.: Meta spatio-temporal debiasing for video scene graph generation. In: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T. (eds.) Computer Vision – ECCV 2022. pp. 374–390. Springer Nature Switzerland, Cham (2022)
73. Yu, F., Chen, H., Wang, X., Xian, W., Chen, Y., Liu, F., Madhavan, V., Darrell, T.: Bdd100k: A diverse driving dataset for heterogeneous multitask learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2636–2645 (2020)
74. Zeng, F., Dong, B., Zhang, Y., Wang, T., Zhang, X., Wei, Y.: Motr: End-to-end multiple-object tracking with transformer. In: Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXVII. pp. 659–675. Springer (2022)
75. Zhang, C., Huang, G., Liu, L., Huang, S., Yang, Y., Wan, X., Ge, S., Tao, D.: Webuav-3m: A benchmark for unveiling the power of million-scale deep uav tracking. IEEE Transactions on Pattern Analysis and Machine Intelligence (2022). https://doi.org/10.1109/TPAMI.2022.3232854
76. Zhang, L., Gao, J., Xiao, Z., Fan, H.: Animaltrack: A benchmark for multi-animal tracking in the wild. International Journal of Computer Vision pp. 1–18 (2022)
77. Zhang, X., Zhao, Z., Tsiligkaridis, T., Zitnik, M.: Self-supervised contrastive pre-training for time series via time-frequency consistency. In: Proceedings of Neural Information Processing Systems, NeurIPS (2022)
78. Zhang, Y., Sun, P., Jiang, Y., Yu, D., Weng, F., Yuan, Z., Luo, P., Liu, W., Wang, X.: Bytetrack: Multi-object tracking by associating every detection box (2022)
79. Zhang, Y., Wang, C., Wang, X., Zeng, W., Liu, W.: Fairmot: On the fairness of detection and re-identification in multiple object tracking. International Journal of Computer Vision **129**, 3069–3087 (2021)
80. Zhang, Y., Wang, T., Liu, K., Zhang, B., Chen, L.: Recent advances of single-object tracking methods: A brief survey. Neurocomputing **455**, 1–11 (2021)
81. Zhao, H., Torralba, A., Torresani, L., Yan, Z.: Hacs: Human action clips and segments dataset for recognition and temporal localization. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 8668–8678 (2019)