



# OVTrack: Open-Vocabulary Multiple Object Tracking

**Conference Paper****Author(s):**

Li, Siyuan; Fischer, Tobias; Ke, Lei; [Ding, Henghui](#) ; [Danelljan, Martin](#) ; Yu, Fisher

**Publication date:**

2023-08-22

**Permanent link:**

<https://doi.org/10.3929/ethz-b-000651808>

**Rights / license:**

[In Copyright - Non-Commercial Use Permitted](#)

**Originally published in:**

<https://doi.org/10.1109/CVPR52729.2023.00539>

# OVTrack: Open-Vocabulary Multiple Object Tracking

Siyuan Li\* Tobias Fischer\* Lei Ke Henghui Ding  
 Martin Danelljan Fisher Yu  
 Computer Vision Lab, ETH Zürich

<https://www.vis.xyz/pub/ovtrack/>

## Abstract

The ability to recognize, localize and track dynamic objects in a scene is fundamental to many real-world applications, such as self-driving and robotic systems. Yet, traditional multiple object tracking (MOT) benchmarks rely only on a few object categories that hardly represent the multitude of possible objects that are encountered in the real world. This leaves contemporary MOT methods limited to a small set of pre-defined object categories. In this paper, we address this limitation by tackling a novel task, open-vocabulary MOT, that aims to evaluate tracking beyond pre-defined training categories. We further develop OVTrack, an open-vocabulary tracker that is capable of tracking arbitrary object classes. Its design is based on two key ingredients: First, leveraging vision-language models for both classification and association via knowledge distillation; second, a data hallucination strategy for robust appearance feature learning from denoising diffusion probabilistic models. The result is an extremely data-efficient open-vocabulary tracker that sets a new state-of-the-art on the large-scale, large-vocabulary TAO benchmark, while being trained solely on static images.

## 1. Introduction

Multiple Object Tracking (MOT) aims to recognize, localize and track objects in a given video sequence. It is a cornerstone of dynamic scene analysis and vital for many real-world applications such as autonomous driving, augmented reality, and video surveillance. Traditionally, MOT benchmarks [9, 11, 19, 64, 71] define a set of semantic categories that constitute the objects to be tracked in the training and testing data distributions. The potential of traditional MOT methods [3, 4, 33, 43] is therefore limited by the taxonomies of those benchmarks. As consequence, contemporary MOT methods struggle with unseen events, leading to a gap between evaluation performance and real-world

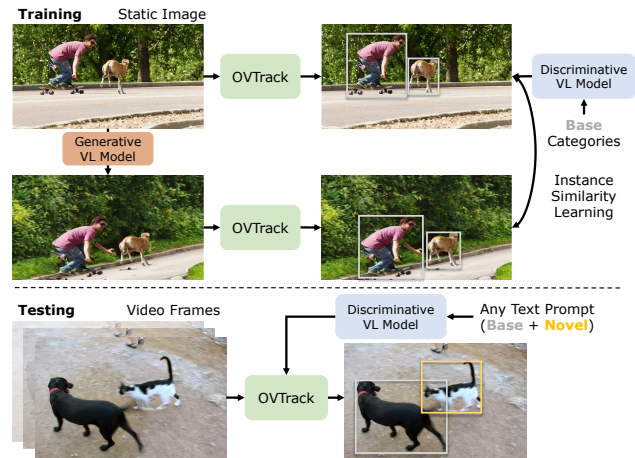


Figure 1. **OVTrack**. We approach the task of open-vocabulary multiple object tracking. During training, we leverage vision-language (VL) models both for generating samples and knowledge distillation. During testing, we track both base and novel classes unseen during training by querying a vision-language model.

deployment.

To bridge this gap, previous works have tackled MOT in an open-world context. In particular, Ošep *et al.* [46, 48] approach generic object tracking by first segmenting the scene and performing tracking before classification. Other works have used class agnostic localizers [10, 49] to perform MOT on arbitrary objects. Recently, Liu *et al.* [37] defined open-world tracking, a task that focuses on the evaluation of previously unseen objects. In particular, it requires any-object tracking as a stage that precedes object classification. This setup comes with two inherent difficulties. First, in an open-world context, densely annotating all objects is prohibitively expensive. Second, without a pre-defined taxonomy of categories, the notion of *what is* an object is ambiguous. As a consequence, Liu *et al.* resort to recall-based evaluation, which is limited in two ways. Penalizing false positives (FP) becomes impossible, *i.e.* we cannot measure the tracker *precision*. Moreover, by evaluating tracking in a class-agnostic manner, we lose the ability to evaluate how

\*Equal contribution.

well a tracker can infer the semantic category of an object.

In this paper, we propose open-vocabulary MOT as an effective solution to these problems. Similar to open-world MOT, open-vocabulary MOT aims to track multiple objects beyond the pre-defined training categories. However, instead of dismissing the classification problem and resorting to recall-based evaluation, we assume that *at test time* we are given the classes of objects we are interested in. This allows us to apply existing closed-set tracking metrics [34, 70] that capture both precision and recall, while still evaluating the tracker’s ability to track arbitrary objects during inference.

We further present the first **Open-Vocabulary Tracker**, OVTrack (see Fig. 1). To this end, we identify and address two fundamental challenges to the design of an open-vocabulary multi-object tracker. The first is that closed-set MOT methods are simply not capable of extending their pre-defined taxonomies. The second is data availability, *i.e.* scaling video data annotation to a large vocabulary of classes is extremely costly. Inspired by existing works in open-vocabulary detection [1, 13, 20, 76], we replace our classifier with an embedding head, which allows us to measure similarities of localized objects to an open vocabulary of semantic categories. In particular, we distill knowledge from CLIP [52] into our model by aligning the image feature representations of object proposals with the corresponding CLIP image and text embeddings.

Beyond detection, association is the core of modern MOT methods. It is driven by two affinity cues: motion and appearance. In an open-vocabulary context, motion cues are brittle since arbitrary scenery contains complex and diverse camera and object motion patterns. In contrast, diverse objects usually exhibit heterogeneous appearance. However, relying on appearance cues requires robust representations that generalize to novel object categories. We find that CLIP feature distillation helps in learning better appearance representations for improved association. This is especially intriguing since object classification and appearance modeling are usually distinct in the MOT pipeline [3, 16, 65].

Learning robust appearance features also requires strong supervision that captures object appearance changes in different viewpoints, background, and lighting. To approach the data availability problem, we utilize the recent success of denoising diffusion probabilistic models (DDPMs) in image synthesis [54, 59] and propose an effective data hallucination strategy tailored to appearance modeling. In particular, from a static image, we generate both simulated positive and negative instances along with random background perturbations.

The main contributions are summarized as follows:

1. We define the task of open-vocabulary MOT and provide a suitable benchmark setting on the large-scale, large-vocabulary MOT benchmark TAO [9].
2. We develop OVTrack, the first open-vocabulary multi-object tracker. It leverages vision-language models to

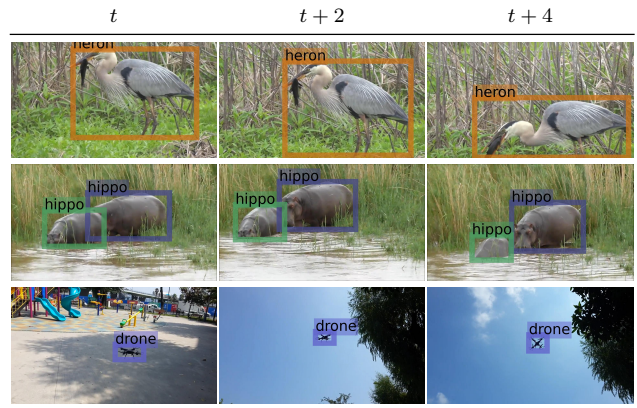


Figure 2. **OVTrack qualitative results.** We condition our tracker on text prompts unseen during training, namely ‘heron’, ‘hippo’ and ‘drone’, and successfully track the corresponding objects in the videos. The box color depicts object identity.

improve both classification and association compared to closed-set trackers.

3. We propose an effective data hallucination strategy that allows us to address the data availability problem in open-vocabulary settings by leveraging DDPMs.

Owing to its thoughtful design, OVTrack sets a new state-of-the-art on the challenging TAO benchmark [9], outperforming existing trackers by a significant margin while being trained *on static images only*. In addition, OVTrack is capable of tracking *arbitrary* object classes (see Fig. 2), overcoming the limitation of closed-set trackers.

## 2. Related work

**Multiple object tracking.** The dominant paradigm in MOT literature is tracking-by-detection [53], where objects are first detected in each frame, and subsequently associated across time. Thus, many works have focused on data association, aiming to exploit similarity cues such as visual appearance [3, 16, 32, 42, 50, 60, 65, 69], 2D object motion [4, 5, 15, 26, 68] or 3D object motion [25, 40, 44, 47, 48, 62] most effectively. Recently, researchers have focused on learning data association with graph neural networks [6, 61] or transformers [41, 63, 73, 78]. However, those works dismiss a more profound problem in the tracking-by-detection pipeline that precedes data association: Contemporary object detectors [23, 36, 55–57] are designed for closed-set scenarios where all objects appear frequently in the training *and* testing data distributions. Hence, Dave *et al.* [9] proposed a new benchmark, TAO, that focuses on studying MOT in the long-tail of the object category distribution. On this benchmark, GTR [78], AOA [12], QDTrack [16] and TET [34] achieve impressive performance. However, those works are still limited to pre-defined object categories and thus do not scale

to the diversity of real-world settings. Our work enables tracking of unseen classes from an open vocabulary.

**Open-world detection and tracking.** Open-world detection methods aim to detect any salient object in a given input image *irrespective* of its category and *beyond* the training data distribution in particular. However, object classification under such a setting is ill-posed since novel classes will be unknown by definition [2, 29]. As such, open-world detection methods utilize class agnostic localizers [10] and treat classification as a clustering problem [29], estimating a similarity between novel instances and grouping them into novel classes via incremental learning.

Instead, open-vocabulary object detection methods aim to detect arbitrary, but *given* classes of objects at test time [72]. For this, Bansal *et al.* [1] connect an object detector with word representations [51]. Recently, models like CLIP [52] learn visual representations from natural language supervision. Their main advantage over word representations is better alignment of visual concepts and language description. Consequently, many works have focused on leveraging image-text representations for open-vocabulary and few-shot object detection [20, 76]. ViLD [20] distills CLIP image features, while Detic [76] leverages classification data for joint training. Other works have focused on learning good language prompts for open-vocabulary object detection [13].

Fewer works have tackled the open-world problem in the MOT domain. Existing works perform scene segmentation and class agnostic tracking before classification [44, 46, 48] or utilize class-agnostic proposal generation [10, 49], similar to open-world detection methods. Liu *et al.* [37] propose an open-world tracking benchmark, TAO-OW, that evaluates class-agnostic tracking as a task that precedes classification. However, this comes with the limitation that the evaluation only captures tracker recall and no classification accuracy. Instead of dismissing classification, we pose the problem in a different way, *i.e. at test time* we know the novel classes we are interested in. This allows us to capture both the precision and recall of novel classes in our evaluation, while our method maintains the ability to track any object.

**Learning tracking from static images.** Since labelled video data is expensive to acquire at scale, recent methods have proposed to use static images to supervise MOT methods [16, 66, 74, 77]. CenterTrack [77] proposes to learn motion offsets from static images by random translation of the input, while FairMOT [74] treats objects in a dataset of static images as unique classes to distinguish. Inspired by recent progress in self-supervised representation learning [7, 22, 45], Fischer *et al.* [16] propose to utilize data augmentation in combination with a contrastive learning objective to learn appearance-based tracking from static images. We go beyond classic data augmentation used in existing works by generating positive and negative examples of objects along with background perturbations via generative

models, offering a more targeted approach to guiding appearance similarity learning from static images.

**Data generation for tracking.** While aforementioned methods alleviate the data availability problem in MOT, there is still room for improvement when it comes to data generation in video tasks. Therefore, a large body of research has focused on data generation strategies that can benefit tracking methods [8, 14, 17, 28, 30, 31, 58]. While early works focused on obtaining synthetic data from computer graphics engines [14, 17, 28, 58], newer approaches combine 3D assets with generative models for improved realism [8, 31]. Fewer works have tackled data generation with generative models only [30]. Recently, DDPMs [27, 54, 59] showed impressive results in image synthesis. We leverage their data generation fidelity to address the data availability problem that is particularly pronounced in open-vocabulary MOT with a novel data hallucination strategy tailored to appearance modeling.

### 3. Open-Vocabulary MOT

In real-world scenarios, object categories follow a long-tailed distribution with a rich vocabulary. The remarkable diversity of the open world cannot be covered by a monolithic dataset. However, existing MOT benchmarks focus on closed-set evaluation, with often only a handful of object classes being evaluated. Furthermore, the task setup requires trackers to only track objects within a small set of training categories. To bridge the gap between existing MOT benchmarks and algorithms and real-world settings, we propose the task of open-vocabulary MOT and define its training and evaluation setup as follows.

At training time, we train a tracker  $M$  on the training data distribution  $\mathcal{D}^{\text{train}} = \{\mathbf{X}^{\text{train}}, \mathcal{A}^{\text{train}}\}$  that contains video sequences  $\mathbf{X}^{\text{train}}$  and their respective annotations  $\mathcal{A}^{\text{train}}$  of objects with semantic categories  $\mathcal{C}^{\text{base}} \subset \mathbb{N}$ . Each annotation  $\alpha \in \mathcal{A}^{\text{train}}$  consists of a set of states  $\{\alpha_t\}_{t \in T}$  for each frame  $t \in T$  that the object is visible in. A state  $\alpha_t = (\mathbf{b}_t, c_t)$  comprises the object class  $c \in \mathbb{N}$  and the 2D bounding box  $\mathbf{b} = [x, y, w, h]$ , where  $(x, y)$  is the center location in pixel coordinates and  $(w, h)$  are width and height, respectively. At test time, we are given video sequences  $\mathbf{X}^{\text{test}}$  and a set of object classes  $\mathcal{C}^{\text{novel}} \subset \mathbb{N} \setminus \mathcal{C}^{\text{base}}$  that we are interested in. We aim to find all tracks  $\mathcal{T}$  of objects in  $\mathbf{X}^{\text{test}}$  belonging to classes  $\mathcal{C}^{\text{base}} \cup \mathcal{C}^{\text{novel}}$ . Each track state  $\tau_t = (\mathbf{b}_t, p_t, c_t) \in \mathcal{T}$  contains predicted object confidence  $p \in [0, 1]$ , class  $c \in \mathbb{N}$  and 2D bounding box  $\mathbf{b} = [x, y, w, h]$ . The important distinctions to closed-set tracking are two-fold: **1)** We evaluate the tracker  $M$  not only on  $\mathcal{C}^{\text{base}}$  but also on  $\mathcal{C}^{\text{novel}}$  with  $\mathcal{C}^{\text{novel}} \cap \mathcal{C}^{\text{base}} = \emptyset$ , and **2)** While  $\mathcal{C}^{\text{novel}}$  is known at test time, our setup requires the tracker  $M$  to track arbitrary object classes  $c \in \mathbb{N}$  since  $\mathcal{C}^{\text{novel}}$  remains unknown *at training time*. In particular, the evaluation of  $\mathcal{C}^{\text{novel}}$  illustrates the ability of tracker  $M$  to track *any* unknown class  $c \in \mathbb{N} \setminus \mathcal{C}^{\text{base}}$ , while the classes in  $\mathcal{C}^{\text{novel}}$  serve as *proxy*.



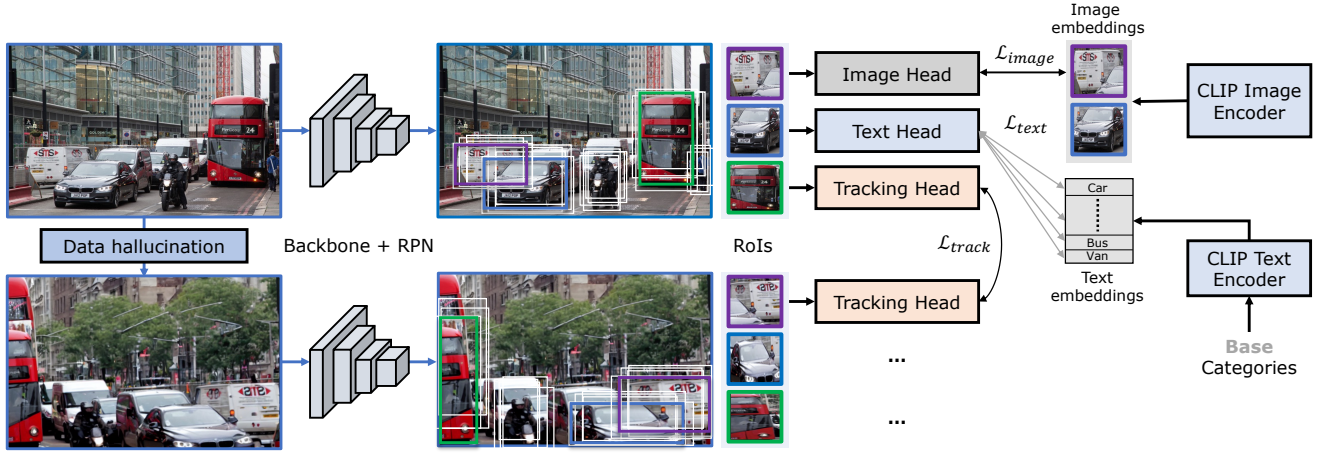


Figure 3. **OVTrack training.** From a single static  $I_{key}$ , we generate  $I_{ref}$  with our data hallucination strategy. We extract RoIs via a RPN [57] and perform knowledge distillation from CLIP [52] via the embeddings of the text and image heads. Note that we train classification only on  $\mathcal{C}^{base}$ . Further, we obtain appearance embeddings from the tracking head and apply our instance similarity loss on the image pair.

### 3.1. Benchmark

We utilize the large-scale, large-vocabulary MOT dataset TAO [9] to establish a suitable benchmark for open-vocabulary MOT. TAO mostly follows the taxonomy of LVIS [21], which divides classes according to their occurrence into frequent, common and rare classes. To obtain our held-out set  $\mathcal{C}^{novel}$ , we follow open-vocabulary detection literature [20] and use the rare classes as defined by LVIS. The intuition behind this is that the occurrence of rare classes is correlated with uncommon scenarios and events that we are particularly interested in evaluating.

With respect to evaluation, the advantage of defining  $\mathcal{C}^{novel}$  is that we can apply closed-set tracking metrics in a straightforward manner, while open-world MOT [37] needs to resort to recall-based evaluation. Further, previous works [34, 37] have shown that the official evaluation metric in TAO, Track mAP [70], is sub-optimal in terms of handling FPs in presence of missing annotations. On the contrary, the recently proposed TETA metric [34] handles this shortcoming via local cluster evaluation. Also, TETA disentangles classification from localization and association performance. Thus, we choose TETA as the evaluation metric for our setup to provide a comprehensive insight into the localization, association, and open-vocabulary classification performance of tracker  $M$ .

## 4. OVTrack

We present our **Open-Vocabulary Tracker**, OVTrack. We address two perspectives of its design: **1) Model perspective:** We show how to handle the open-vocabulary setting in the localization, classification, and association modules of the tracker in Section 4.1; **2) Data perspective:** Collecting and annotating the necessary amount of training videos is impractical for open-vocabulary MOT. Therefore, we contribute a

novel training approach for learning object tracking **without** video data in Section 4.2.

### 4.1. Model design

We decompose OVTrack’s functionality into localization, classification and association and discuss our open-vocabulary design philosophy in tackling difficulties for each of those parts. The model design is illustrated in Fig. 3.

**1) Localization:** To localize objects of arbitrary and possibly unknown classes  $c \in \mathbb{N}$  in a video, we train Faster R-CNN [57] in a class-agnostic manner, *i.e.* we use only the RPN and regression losses defined in [57]. We find that this localization procedure can generalize well to object classes that are unknown at training time, as also validated by previous works [10, 20, 76]. During training, we use RPN proposals as object candidates  $P$  for greater diversity, while during inference, we use the refined RCNN outputs as object candidates. Each candidate  $r \in P$  is defined by confidence  $p_r$  and bounding box  $\mathbf{b}_r$ .

**2) Classification:** Existing closed-set trackers [3, 4, 16, 77] can only track objects of categories in  $\mathcal{C}^{base}$ , *i.e.* objects present and annotated in the training data distribution  $\mathcal{D}^{train}$ . To enable open-vocabulary classification, we need to be able to configure the classes we are interested in without re-training. Inspired by open-vocabulary detection literature [1], we connect our Faster R-CNN with the vision-language model CLIP [52] that has been pre-trained on over 400 million image-text pairs for contrastive learning.

After extracting the RoI feature embeddings  $\mathbf{f}_r = \mathcal{R}(\phi(I), \mathbf{b}_r)$ ,  $\forall r \in P$  from the backbone  $\phi$ , we replace the original classifier in Faster R-CNN with a text head and add an image head generating the embeddings  $\hat{\mathbf{t}}_r$  and  $\hat{\mathbf{i}}_r$ , for each  $\mathbf{f}_r$ . We use the CLIP text and image encoders to supervise the heads following [13, 20]. In particular, we use the class names to generate text prompts  $\mathcal{P}(c) = \{\mathbf{v}_1^c, \dots, \mathbf{v}_L^c, \mathbf{w}_c\}$

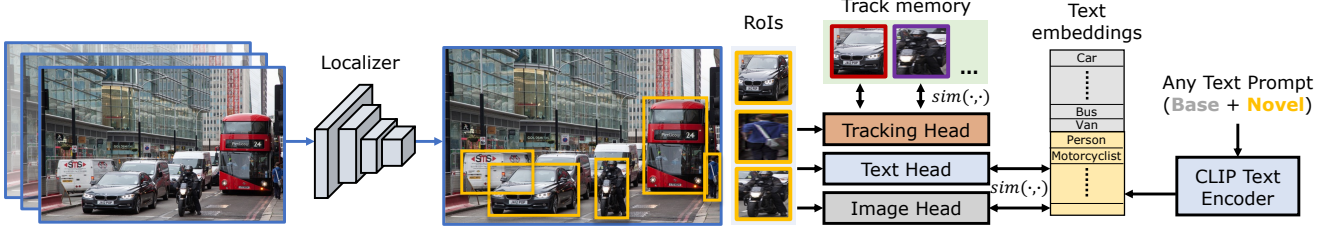


Figure 4. **OVTrack inference.** Given an input video stream, we track objects of arbitrary classes, e.g.  $\mathcal{C}^{\text{base}} \cup \mathcal{C}^{\text{novel}}$ . We first localize objects agnostic of their class, then assign a semantic class label via the text embedding head, and finally associate them to existing tracks by comparing their appearance embedding  $\mathbf{q}$  obtained from the tracking head to the appearance embeddings in the track memory.

that consist of  $L$  context vectors  $\mathbf{v}^c$  and a class name embedding  $\mathbf{w}_c$ . We feed the prompts into the CLIP text encoder  $\mathcal{E}$ , generating text embeddings  $\mathbf{t}_c = \mathcal{E}(\mathcal{P}(c))$ ,  $\forall c \in \mathcal{C}^{\text{base}}$ . We compute the affinity between the predicted embeddings  $\hat{\mathbf{t}}_r$  and their CLIP counterpart  $\mathbf{t}_c$ .

$$\mathbf{z}(r) = [\cos(\hat{\mathbf{t}}_r, \mathbf{t}_{bg}), \cos(\hat{\mathbf{t}}_r, \mathbf{t}_1), \dots, \cos(\hat{\mathbf{t}}_r, \mathbf{t}_{|\mathcal{C}^{\text{base}}|})] \quad (1)$$

$$\mathcal{L}_{\text{text}} = \frac{1}{|P|} \sum_{r \in P} \mathcal{L}_{\text{CE}}(\text{softmax}(\mathbf{z}(r)/\lambda), c_r), \quad (2)$$

where  $\cos(\mathbf{v}, \mathbf{k}) = \frac{\mathbf{v} \cdot \mathbf{k}}{\|\mathbf{v}\| \|\mathbf{k}\|}$ ,  $\mathbf{t}_{bg}$  a learned background prompt,  $\lambda$  a temperature parameter,  $\mathcal{L}_{\text{CE}}$  the cross-entropy loss and  $c_r$  is the class label of  $r$ . Furthermore, we align each  $\hat{\mathbf{i}}_r$  with the CLIP image encoder  $\mathcal{I}$ . For each  $r$ , we crop the input image to  $\mathbf{b}_r$ , and resize it to the required input size to obtain the image embedding  $\mathbf{i}_r = \mathcal{I}(\mathcal{R}(I, \mathbf{b}_r))$ . We minimize the distance between the corresponding  $\hat{\mathbf{i}}_r$  and  $\mathbf{i}_r$ .

$$\mathcal{L}_{\text{image}} = \frac{1}{|P|} \sum_{r \in P} \|\hat{\mathbf{i}}_r - \mathbf{i}_r\|_1. \quad (3)$$

**3) Association:** An open-vocabulary tracker should handle diverse scenarios that comprise complex camera motion and heterogeneous object motion patterns. However, those patterns are difficult to model especially when there are not enough video annotations available [16, 37]. Therefore, we rely on appearance cues to robustly track objects in an open-vocabulary context. Specifically, we employ a contrastive learning approach inspired by [16, 34]. Given an image pair  $(I_{\text{key}}, I_{\text{ref}})$  we extract RoIs from both images and match the RoIs to the annotations using intersection-over-union (IoU). For each matched RoI in  $I_{\text{key}}$  with appearance embedding  $\mathbf{q} \in Q$ , we cluster objects  $Q^+$  with the same identity and divide objects  $Q^-$  with different identity in  $I_{\text{ref}}$ .

$$\text{PosD}(\mathbf{q}) = \frac{1}{|Q^+(\mathbf{q})|} \sum_{\mathbf{q}^+ \in Q^+} \exp(\mathbf{q} \cdot \mathbf{q}^+ / \tau), \quad (4)$$

$$\text{Sim}(\mathbf{q}) = \frac{\exp(\mathbf{q} \cdot \mathbf{q}^+ / \tau)}{\text{PosD}(\mathbf{q}) + \sum_{\mathbf{q}^- \in Q^-} \exp(\mathbf{q} \cdot \mathbf{q}^- / \tau)}, \quad (5)$$

$$\mathcal{L}_{\text{track}} = - \sum_{\mathbf{q} \in Q} \frac{1}{|Q^+(\mathbf{q})|} \sum_{\mathbf{q}^+ \in Q^+} \log(\text{Sim}(\mathbf{q}^+)). \quad (6)$$

We further apply an auxiliary loss  $\mathcal{L}_{\text{aux}}$  to constrain the magnitude of the logits following [16].

During inference, we use straightforward appearance feature similarity for associating existing tracks  $\mathcal{T}$  with objects in  $P$ . In particular, for each track  $\tau \in \mathcal{T}$  and its corresponding appearance embedding  $\mathbf{q}_\tau$ , we compare its similarity with all candidate objects  $r \in P$  using appearance embedding  $\mathbf{q}_r$ . We measure the similarity  $s(\tau, r)$  of existing tracks with the candidate objects using both bi-directional softmax [16] and cosine similarity. We assign  $r$  to the track  $\tau$  with its maximum similarity, if  $s(\tau, r) > \beta$ . If  $r$  does not have a matching track, it starts a new track if its confidence  $p_r > \gamma$  and is discarded otherwise. The inference pipeline is illustrated in Fig. 4.

## 4.2. Learning to track without video data

In this section, we focus on how to train OVTrack in an open-vocabulary context. In particular, open-vocabulary MOT is challenging from a data perspective since we need to localize, classify, and associate possibly unknown objects of extremely diverse appearance *with fixed method components*. Thus, it is of fundamental importance to align our training data distribution  $\mathcal{D}^{\text{train}}$  with the conditions found in evaluation. However, existing video datasets lack the diversity of contemporary image datasets. Hence, it is essential for open-vocabulary trackers to leverage not only video but more importantly static image data during training.

We use the large-scale, diverse image dataset LVIS [21] to train OVTrack. In particular, for each image  $I_{\text{key}}$  we generate a reference image  $I_{\text{ref}}$ . Referring to our instance similarity loss in Eq. 6, the appearance similarity learning is constituted by contrasting positive and negative examples  $\mathbf{q}^+$  and  $\mathbf{q}^-$ . The corollary of this is that learning will be optimal if  $\mathbf{q}^+$  consists of examples with distortions commonly encountered in video data, such as change in object scale, viewpoint or lighting, while  $\mathbf{q}^-$  contains examples with appearance changes associated with object identity, such as different material. While distortions like translation, scaling, and rotation can be simulated via classic data augmentation strategies [16, 74, 77], there are certain phenomena like changes in object viewpoint, lighting or context that cannot

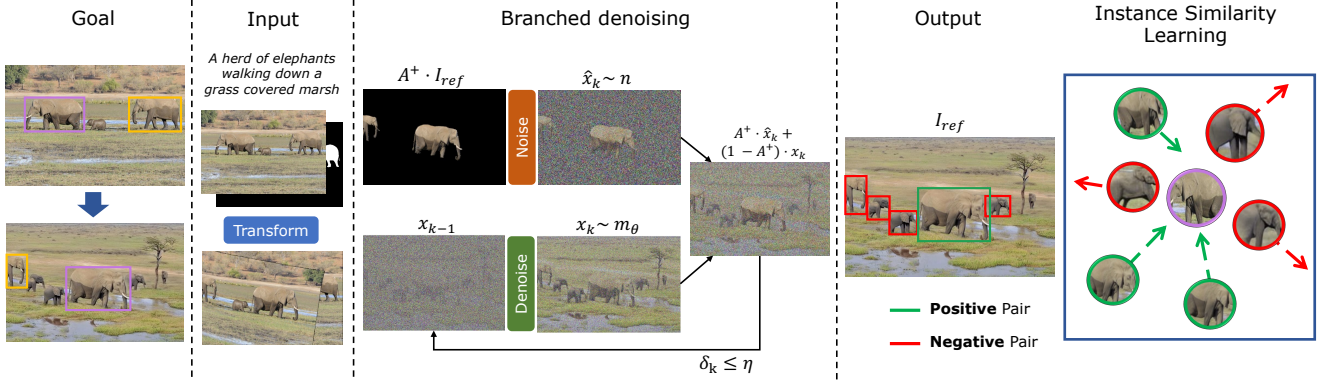


Figure 5. **Data hallucination strategy.** Given an input image, its annotations and its caption, we generate  $x_0 \sim n$  and input it to the diffusion model [59], which progressively denoises  $x$  from  $\delta_0 = 0.75$  to  $\eta$ . At the same time, the foreground regions are kept fixed at each iteration. Specifically, we compose the generated images  $x_k \sim m_\theta$  with foreground regions  $\hat{x}_k \sim n$  at the current noise level, to obtain  $x_k$  as input for the next iteration. Finally, we arrive at  $I_{ref}$  which we use for instance similarity learning.

be simulated by these.

Therefore, to simulate all desired properties of our instance embedding space in  $I_{ref}$ , we combine classic data augmentations with a DDPM-based data hallucination strategy. We use the data generation fidelity of stable diffusion [59] to simulate  $\mathbf{q}^+$  and  $\mathbf{q}^-$  via a specialized denoising process. Generally, the denoising process of a DPPM can be viewed as the inversion of a forward process that maps an input  $x$  to Gaussian white noise  $\mathcal{N}(0, \mathbf{I})$ . In the forward direction, Gaussian noise is added to the input image according to a variance schedule  $\delta_k$  in  $K$  steps.

$$n(x_k|x_{k-1}) = \mathcal{N}(x_k; \sqrt{1 - \delta_k} \cdot x_{k-1}, \delta_k \mathbf{I}). \quad (7)$$

In the backward direction, a neural network with parameters  $\theta$  predicts the parameters  $\mu$  and  $\Sigma$  of a Gaussian distribution that reverses a forward step.

$$m_\theta(x_{k-1}|x_k) = \mathcal{N}(x_{k-1}; \mu_\theta(x_k, k), \Sigma_\theta(x_k, k)). \quad (8)$$

Our specialized denoising process is illustrated in Fig. 5. We initialize  $I_{ref}$  as  $I_{key}$  and apply a random geometric transformation to  $I_{ref}$ . Next, we use the instance mask annotations in LVIS to define the set of positive examples  $A^+$ . We divide each iteration of the denoising process of image  $I_{ref}$  using the union of all object masks in  $A^+$  into two branches following [39]. In addition, we use the conditioning mechanism in [59] to guide the backward process with the corresponding image caption. To initialize the backward process, we set  $x_0$  to  $I_{ref}$  at  $\delta_0 = 0.75$  via the forward process  $n(x_k|x_{k-1})$ . Note that  $x$  corresponds to a latent representation of  $I_{ref}$  obtained via the encoder of stable diffusion. In each iteration, we apply  $m_\theta(x_{k-1}|x_k)$  to obtain a new sample  $x_k \sim m_\theta$ . At the same time, we use the forward process  $n(x_k|x_{k-1})$  on the areas of  $A^+$  to generate  $\hat{x}_k \sim n$ . At the end of each reverse iteration, we compose the two versions via

$x_k = A^+ \cdot \hat{x}_k + (1 - A^+) \cdot x_k$ . We iterate until  $\delta_k \leq \eta$ . Finally, we apply  $m_\theta(x_{k-1}|x_k)$  to the whole image, without branching, as a homogenization step between  $A^+ x_k$  and  $(1 - A^+) x_k$ , while  $\eta > \delta_k > 0$ .

By this process, we achieve three goals. First, we generate random perturbations of the background. Second, we keep the areas of  $A^+$  close to its original content in each denoising step so that positive instances are integrated well into the new background. Third, we generate distractor objects by caption guided hallucination.

## 5. Experiments

### 5.1. Evaluation metrics

**TETA.** The tracking-every-thing accuracy (TETA) [34] is calculated from three independent scores. First, the localization accuracy (LocA) is calculated by matching all annotated boxes  $\alpha$  to the predicted boxes of  $\mathcal{T}$  without taking classification into account:  $\text{LocA} = \frac{|\text{TPL}|}{|\text{TPL}| + |\text{FPL}| + |\text{FNL}|}$ . Next, classification accuracy (ClsA) is computed based on all well-localized TPL, comparing the predicted semantic classes to the matched ground-truths:  $\text{ClsA} = \frac{|\text{TPC}|}{|\text{TPC}| + |\text{FPC}| + |\text{FNC}|}$ . Finally, association accuracy (AssocA) is computed in a similar fashion, comparing the identity of associated ground truths with well-localized predictions:  $\text{AssocA} = \frac{1}{|\text{TPL}|} \sum_{b \in \text{TPL}} \frac{|\text{TPA}(b)|}{|\text{TPA}(b)| + |\text{FPA}(b)| + |\text{FNA}(b)|}$ . The TETA score is computed as the arithmetic mean of the three scores.

**Track mAP.** The Track mAP [70] is calculated using the 3D IoU between the bounding boxes of a predicted track  $\tau$  and an annotated track  $\alpha$  by  $\text{IoU}_{3D}(\tau, \alpha) = \frac{\sum_{t \in T} \tau_t \cap \alpha_t}{\sum_{t \in T} \tau_t \cup \alpha_t}$ . It is used analogous to 2D bounding box IoU to calculate the popular average precision metric per class as in [21]. The Track mAP is the average of the per-class scores across a set of  $\text{IoU}_{3D}$  thresholds.

Table 1. **Open-vocabulary MOT comparison.** We compare our method with existing closed-set trackers and off-the-shelf open-vocabulary baselines on base and novel classes on the validation and test sets of TAO [9]. We indicate the classes and data the methods trained on. Note that methods using TAO data utilize videos for training. All methods use ResNet50 [18] as backbone.

Method	Classes		Data			Base				Novel			
Validation set	Base	Novel	CC3M	LVIS	TAO	TETA	LocA	AssocA	ClsA	TETA	LocA	AssocA	ClsA
QDTrack [16]	✓	✓	-	✓	✓	27.1	45.6	24.7	11.0	22.5	42.7	24.4	0.4
TETer [34]	✓	✓	-	✓	✓	30.3	47.4	31.6	12.1	25.7	45.9	31.1	0.2
DeepSORT (ViLD) [65]	✓	-	-	✓	✓	26.9	47.1	15.8	17.7	21.1	46.4	14.7	<b>2.3</b>
Tracktor++ (ViLD) [3]	✓	-	-	✓	✓	28.3	47.4	20.5	17.0	22.7	46.7	19.3	2.2
<b>OVTrack</b>	✓	-	-	✓	-	<b>35.5</b>	<b>49.3</b>	<b>36.9</b>	<b>20.2</b>	<b>27.8</b>	<b>48.8</b>	<b>33.6</b>	1.5
RegionCLIP [75]													
+ DeepSORT [65]	✓	-	✓	✓	✓	28.4	52.5	15.6	17.0	24.5	49.2	15.3	9.0
+ Tracktor++ [3]	✓	-	✓	✓	✓	29.6	52.4	19.6	16.9	25.7	50.1	18.9	8.1
+ <b>OVTrack</b>	✓	-	✓	✓	-	<b>36.3</b>	<b>53.9</b>	<b>36.3</b>	<b>18.7</b>	<b>32.0</b>	<b>51.4</b>	<b>33.2</b>	<b>11.4</b>
Test set	Base	Novel	CC3M	LVIS	TAO	TETA	LocA	AssocA	ClsA	TETA	LocA	AssocA	ClsA
QDTrack [16]	✓	✓	-	✓	✓	25.8	43.2	23.5	10.6	20.2	39.7	20.9	0.2
TETer [34]	✓	✓	-	✓	✓	29.2	44.0	30.4	10.7	21.7	39.1	25.9	0.0
DeepSORT (ViLD) [65]	✓	-	-	✓	✓	24.5	43.8	14.6	15.2	17.2	38.4	11.6	1.7
Tracktor++ (ViLD) [3]	✓	-	-	✓	✓	26.0	44.1	19.0	14.8	18.0	39.0	13.4	1.7
<b>OVTrack</b>	✓	-	-	✓	-	<b>32.6</b>	<b>45.6</b>	<b>35.4</b>	<b>16.9</b>	<b>24.1</b>	<b>41.8</b>	<b>28.7</b>	<b>1.8</b>
RegionCLIP [75]													
+ DeepSORT [65]	✓	-	✓	✓	✓	27.0	49.8	15.1	16.1	18.7	41.8	9.1	5.2
+ Tracktor++ [3]	✓	-	✓	✓	✓	28.0	49.4	18.8	15.7	20.0	42.4	12.0	5.7
+ <b>OVTrack</b>	✓	-	✓	✓	-	<b>34.8</b>	<b>51.1</b>	<b>36.1</b>	<b>17.3</b>	<b>25.7</b>	<b>44.8</b>	<b>26.2</b>	<b>6.1</b>

Table 2. **Closed-set MOT Track mAP comparison.** We compare to existing trackers on TAO [9] validation. Competing methods use ResNet101 [24], we use ResNet50 as backbone. All methods use Faster R-CNN [57]. We include results with stronger detectors and additional data in gray. † does not use videos for training.

Method	Track mAP50	Track mAP75	Track mAP
SORT-TAO [9]	13.2	-	-
QDTrack [16]	15.9	5.0	10.6
GTR† [78]	20.4	-	-
TAC [67]	17.7	5.80	7.30
BIV [66]	19.6	7.30	13.6
<b>OVTrack†</b>	<b>21.2</b>	<b>10.6</b>	<b>15.9</b>
GTR + CenterNet2† [78]	22.5	-	-
AOA [12]	25.8	-	-

## 5.2. Implementation details

We use ResNet50 [24] with FPN [35]. We filter object candidates  $P$  by non-maximum suppression (NMS) with an IoU threshold of 0.7 and randomly select  $|P| = 256$  candidates per image. We set  $\lambda = 0.07$  in  $\mathcal{L}_{\text{text}}$ . We use a two-stage training process, first training the detection components following [13, 20], second fine-tuning the model for tracking with loss weights 0.25 for  $\mathcal{L}_{\text{track}}$  and 1.0 for  $\mathcal{L}_{\text{aux}}$  following [16]. We train on the LVIS dataset, with one hallucinated counterpart per image in the dataset. While we use the full dataset for the state-of-the-art comparison, we use a subset of 10,000 images for the ablation studies due to resource constraints. Unless otherwise noted, we use the following data augmentations in training: resizing, random horizontal flipping, color jittering, random affine transformation, and mosaic composition with varying parameters between  $I_{\text{key}}$  and  $I_{\text{ref}}$ . We use  $\eta = 0.02$  for data generation. For inference, we select object candidates  $P$  by NMS with

Table 3. **Closed-set MOT TETA comparison.** We compare to existing trackers on the TAO [9] validation. Benchmark results are taken from [34]. All competing methods use ResNet101 [24] except AOA [12], we use ResNet50 as backbone. All methods use Faster R-CNN [57]. † does not use videos for training.

Method	TETA	LocA	AssocA	ClsA
SORT-TAO [9]	24.8	48.1	14.3	12.1
Tracktor [3]	24.2	47.4	13.0	12.1
DeepSORT [65]	26.0	48.4	17.5	12.1
AOA [12]	25.3	23.4	30.6	<b>21.9</b>
Tracktor++ [9]	28.0	49.0	22.8	12.1
QDTrack [16]	30.0	50.5	27.4	12.1
TETer [34]	33.3	<b>51.6</b>	35.0	13.2
<b>OVTrack†</b>	<b>34.7</b>	49.3	<b>36.7</b>	18.1

an IoU threshold of 0.5. We keep a track memory of 10 frames to re-identify objects after occlusion and set  $\beta = 0.5$  and  $\gamma = 0.0001$  (see Sec. 4.1).

## 5.3. Comparison to state-of-the-art

**Open-vocabulary MOT.** In Tab. 1, we show the open-vocabulary MOT evaluation on the TAO validation and test sets, divided into base classes  $\mathcal{C}^{\text{base}}$  and novel classes  $\mathcal{C}^{\text{novel}}$ . For details on the setup, please refer to the supplemental material. The baselines we establish are composed of both closed-set and open-vocabulary trackers. We choose the two state-of-the-art closed-set trackers, TETer [34] and QDTrack [16], trained on  $\mathcal{C}^{\text{base}} \cup \mathcal{C}^{\text{novel}}$ . In addition, we combine off-the-shelf trackers DeepSORT [65] and Tracktor++ [3] with the open-vocabulary detector ViLD [20] as baseline open-vocabulary trackers. These are, like OVTrack, trained on  $\mathcal{C}^{\text{base}}$  only. Note that all baselines use video data for training, while we use only static images.

Our approach substantially outperforms all closed-set and



open-vocabulary baselines. We achieve consistent improvement across LocA, AssocA, and ClsA on both base and novel classes. The baselines trained on  $\mathcal{C}^{\text{base}} \cup \mathcal{C}^{\text{novel}}$  can, in some cases, correctly classify objects in  $\mathcal{C}^{\text{novel}}$  but achieve poor results. On the contrary, both the open-vocabulary baselines and our tracker achieve significantly higher ClsA on novel classes. However, we note that classification on the TAO dataset remains a very challenging task. The absolute ClsA scores on novel classes are low. This is partially due to the nature of the ClsA metric, which only considers top-1 classification accuracy, while classes on the TAO dataset are diverse and fine-grained.

Therefore, we investigate the use of stronger, recently proposed open-vocabulary detectors. We combine RegionCLIP [75] with our off-the-shelf baselines and OVTrack. We replace the localization and classification parts of OVTrack with RegionCLIP while keeping the association fixed. We observe that ClsA increases substantially for all trackers on novel classes. Our method achieves the best performance by a wide margin and achieves the best ClsA scores with 11.4 and 6.1 on the validation and test sets, respectively. Note however that RegionCLIP makes use of additional data.

**Closed-set MOT.** In Tab. 2 and Tab. 3 we compare to existing works on the validation split of TAO using Track mAP and TETA metrics, respectively. Note that our method neither uses video data for training, nor is it trained on rare classes as defined in Sec. 3.1, while all of the compared closed-set trackers train on video data and use the held-out rare classes for training as they are part of the closed-set evaluation in TAO. We outperform all previous works by a sizable margin on both metrics. By examining the TETA scores in Tab. 3, we observe that our tracker obtains 2.3 points less in LocA compared to TETer [34]. However, our approach beats TETer in terms of AssocA by 1.7 points and greatly improves in ClsA by 4.9 points, illustrating the positive effect of CLIP distillation on both classification and associated compared to closed-set trackers. This validates our design in Sec. 4.1. Note that while AOA [12] has a better ClsA, it ensembles multiple few-shot detection and re-identification models trained on additional datasets as reported by previous works [34, 78]. Overall, our approach surpasses the previous state-of-the-art by 1.4 points in TETA and 2.3 points in Track mAP while using a weaker backbone and the same detector.

#### 5.4. Ablation studies

**CLIP knowledge distillation.** In Tab. 4 we analyze the effect of the knowledge distillation described in Sec. 4.1. In particular, we observe that using both  $\mathcal{L}_{\text{text}}$  and  $\mathcal{L}_{\text{image}}$  is more effective for classification than using only  $\mathcal{L}_{\text{text}}$ , improving ClsA significantly from 15.6 to 18.1, while LocA and AssocA stay at the same level of performance.

**Data hallucination strategy.** We validate the effective-

Table 4. **Ablation study on CLIP knowledge distillation.** We show that using both  $\mathcal{L}_{\text{text}}$  and  $\mathcal{L}_{\text{image}}$  is important to classification performance when doing CLIP knowledge distillation (Sec. 4.1).

$\mathcal{L}_{\text{text}}$	$\mathcal{L}_{\text{image}}$	TETA	LocA	AssocA	ClsA
✓	-	34.0	<b>50.5</b>	<b>35.7</b>	15.6
✓	✓	<b>34.3</b>	49.3	35.4	<b>18.1</b>

Table 5. **Ablation study on data hallucination strategy.** We show that our data hallucination strategy (‘DDPM’, Sec. 4.2) improves the association of a closed-set tracker [34] and our OVTrack on TAO [9] validation. We ensemble it with data augmentations, where ‘Standard’ refers to random resize and horizontal flip, ‘Heavy’ to color jitter, random affine transformation and mosaic.

Standard	DDPM	Heavy	TETA	LocA	AssocA	ClsA
<b>TEter-SwinT</b>						
✓	-	-	32.3	50.7	30.6	15.5
✓	✓	-	33.2	51.2	33.0	15.4
✓	✓	✓	<b>34.3</b>	<b>51.4</b>	<b>35.5</b>	<b>15.8</b>
<b>OVTrack</b>						
✓	-	-	32.5	48.9	31.1	17.6
✓	✓	-	33.3	48.9	32.9	18.0
✓	✓	✓	<b>34.4</b>	<b>49.1</b>	<b>35.7</b>	<b>18.3</b>

ness of our data hallucination strategy described in Sec. 4.2 by training both the closed-set tracker TETer [34] and our OVTrack with it. Note that we choose to use SwinT [38] with TETer and ResNet50 [24] with our OVTrack to achieve similar performance, in order to fairly compare the performance difference on both trackers. Tab. 5 shows the TETA results on the TAO validation set. We observe that our data hallucination strategy improves the AssocA significantly for both trackers, while LocA and ClsA are comparable. In particular, we improve 2.4 and 1.8 points in AssocA for TETer and OVTrack, respectively. Further, ensembling our data generation strategy with heavy data augmentations yields another 2.5 and 1.8 points improvement. Overall, we show that our data generation strategy improves instance similarity learning across both closed-set and open-vocabulary trackers while being complementary to classic data augmentation.

## 6. Conclusion

This work introduced open-vocabulary MOT as an effective solution to evaluating multi-object trackers beyond pre-defined training categories. We defined a suitable benchmark setting and presented OVTrack, a data-efficient open-vocabulary tracker. By using knowledge distillation from vision-language models, we improve tracking while going beyond limited dataset taxonomies. In addition, we put forth a data hallucination strategy tailored to instance similarity learning that addresses the data availability problem in open-vocabulary MOT. As a result, OVTrack learns tracking from static images and is able to track arbitrary objects in videos while outperforming existing trackers by a sizable margin on the large-scale, large-vocabulary TAO [9] benchmark.

## References

- [1] Ankan Bansal, Karan Sikka, Gaurav Sharma, Rama Chellappa, and Ajay Divakaran. Zero-shot object detection. In *ECCV*, 2018. 2, 3, 4
- [2] Abhijit Bendale and Terrance Boulton. Towards open world recognition. In *CVPR*, 2015. 3
- [3] Philipp Bergmann, Tim Meinhardt, and Laura Leal-Taixé. Tracking without bells and whistles. In *ICCV*, 2019. 1, 2, 4, 7
- [4] Alex Bewley, Zongyuan Ge, Lionel Ott, Fabio Ramos, and Ben Uppcroft. Simple online and realtime tracking. In *ICIP*, 2016. 1, 2, 4
- [5] Erik Bochinski, Volker Eiselein, and Thomas Sikora. High-speed tracking-by-detection without using image information. In *AVSS*, 2017. 2
- [6] Guillem Brasó and Laura Leal-Taixé. Learning a neural solver for multiple object tracking. In *CVPR*, 2020. 2
- [7] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, 2020. 3
- [8] Yun Chen, Frieda Rong, Shivam Duggal, Shenlong Wang, Xinchun Yan, Sivabalan Manivasagam, Shangjie Xue, Ersin Yumer, and Raquel Urtasun. Geosim: Realistic video simulation via geometry-aware composition for self-driving. In *CVPR*, 2021. 3
- [9] Achal Dave, Tarasha Khurana, Pavel Tokmakov, Cordelia Schmid, and Deva Ramanan. Tao: A large-scale benchmark for tracking any object. In *ECCV*, 2020. 1, 2, 4, 7, 8
- [10] Achal Dave, Pavel Tokmakov, and Deva Ramanan. Towards segmenting anything that moves. In *CVPRW*, 2019. 1, 3, 4
- [11] 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(4):845–881, 2021. 1
- [12] 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. 2, 7, 8
- [13] Yu Du, Fangyun Wei, Zihe Zhang, Miaoqing Shi, Yue Gao, and Guoqi Li. Learning to prompt for open-vocabulary object detection with vision-language model. In *CVPR*, 2022. 2, 3, 4, 7
- [14] Matteo Fabbri, Guillem Brasó, Gianluca Maugeri, Orcun Cetintas, Riccardo Gasparini, Aljoša Ošep, Simone Calderara, Laura Leal-Taixé, and Rita Cucchiara. Motosynth: How can synthetic data help pedestrian detection and tracking? In *CVPR*, 2021. 3
- [15] Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman. Detect to track and track to detect. In *ICCV*, 2017. 2
- [16] Tobias Fischer, Jiangmiao Pang, Thomas E Huang, Linlu Qiu, Haofeng Chen, Trevor Darrell, and Fisher Yu. Qdtrack: Quasi-dense similarity learning for appearance-only multiple object tracking. *arXiv preprint arXiv:2210.06984*, 2022. 2, 3, 4, 5, 7
- [17] Adrien Gaidon, Qiao Wang, Yohann Cabon, and Eleonora Vig. Virtual worlds as proxy for multi-object tracking analysis. In *CVPR*, 2016. 3
- [18] Shang-Hua Gao, Ming-Ming Cheng, Kai Zhao, Xin-Yu Zhang, Ming-Hsuan Yang, and Philip Torr. Res2net: A new multi-scale backbone architecture. *IEEE TPAMI*, 43(2):652–662, 2019. 7
- [19] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *CVPR*, 2012. 1
- [20] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. *arXiv preprint arXiv:2104.13921*, 2021. 2, 3, 4, 7
- [21] Agrim Gupta, Piotr Dollar, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation. In *CVPR*, 2019. 4, 5, 6
- [22] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 2020. 3
- [23] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, 2017. 2
- [24] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 7, 8
- [25] David Held, Jesse Levinson, and Sebastian Thrun. Precision tracking with sparse 3d and dense color 2d data. In *ICRA*, 2013. 2
- [26] David Held, Sebastian Thrun, and Silvio Savarese. Learning to track at 100 FPS with deep regression networks. In *ECCV*, 2016. 2
- [27] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *NeurIPS*, 2020. 3
- [28] Hou-Ning Hu, Yung-Hsu Yang, Tobias Fischer, Trevor Darrell, Fisher Yu, and Min Sun. Monocular quasi-dense 3d object tracking. *IEEE TPAMI*, 2022. 3
- [29] KJ Joseph, Salman Khan, Fahad Shahbaz Khan, and Vineeth N Balasubramanian. Towards open world object detection. In *CVPR*, 2021. 3
- [30] Anna Khoreva, Rodrigo Benenson, Eddy Ilg, Thomas Brox, and Bernt Schiele. Lucid data dreaming for video object segmentation. *IJCV*, 127(9):1175–1197, 2019. 3
- [31] Seung Wook Kim, Jonah Philion, Antonio Torralba, and Sanja Fidler. Drivegan: Towards a controllable high-quality neural simulation. In *CVPR*, 2021. 3
- [32] Laura Leal-Taixé, Cristian Canton-Ferrer, and Konrad Schindler. Learning by tracking: Siamese CNN for robust target association. In *CVPRW*, 2016. 2
- [33] Bastian Leibe, Aleš Leonardis, and Bernt Schiele. Robust object detection with interleaved categorization and segmentation. *IJCV*, 77(1):259–289, 2008. 1
- [34] Siyuan Li, Martin Danelljan, Henghui Ding, Thomas E Huang, and Fisher Yu. Tracking every thing in the wild. In *ECCV*, 2022. 2, 4, 5, 6, 7, 8
- [35] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *CVPR*, 2017. 7
- [36] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C Berg. Ssd: Single shot multibox detector. In *ECCV*, 2016. 2

- [37] 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*, 2022. 1, 3, 4, 5
- [38] 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*, 2021. 8
- [39] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *CVPR*, 2022. 6
- [40] Jonathon Luiten, Tobias Fischer, and Bastian Leibe. Track to reconstruct and reconstruct to track. *RA-L*, 5(2):1803–1810, 2020. 2
- [41] Tim Meinhardt, Alexander Kirillov, Laura Leal-Taixe, and Christoph Feichtenhofer. Trackformer: Multi-object tracking with transformers. In *CVPR*, 2022. 2
- [42] Anton Milan, Seyed Hamid Rezatofighi, Anthony R. Dick, Ian D. Reid, and Konrad Schindler. Online multi-target tracking using recurrent neural networks. In *AAAI*, 2017. 2
- [43] Anton Milan, Stefan Roth, and Konrad Schindler. Continuous energy minimization for multitarget tracking. *IEEE TPAMI*, 36(1):58–72, 2013. 1
- [44] Dennis Mitzel and Bastian Leibe. Taking mobile multi-object tracking to the next level: People, unknown objects, and carried items. In *ECCV*, 2012. 2, 3
- [45] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 3
- [46] Aljoša Ošep, Alexander Hermans, Francis Engelmann, Dirk Klosternann, Markus Mathias, and Bastian Leibe. Multi-scale object candidates for generic object tracking in street scenes. In *ICRA*, 2016. 1, 3
- [47] Aljoša Ošep, Wolfgang Mehner, Markus Mathias, and Bastian Leibe. Combined image- and world-space tracking in traffic scenes. In *ICRA*, 2017. 2
- [48] Aljoša Ošep, Wolfgang Mehner, Paul Voigtlaender, and Bastian Leibe. Track, then decide: Category-agnostic vision-based multi-object tracking. In *ICRA*, 2018. 1, 2, 3
- [49] Aljoša Ošep, Paul Voigtlaender, Mark Weber, Jonathon Luiten, and Bastian Leibe. 4d generic video object proposals. In *ICRA*, 2020. 1, 3
- [50] 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*, 2021. 2
- [51] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *EMNLP*, 2014. 3
- [52] 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*, 2021. 2, 3, 4
- [53] Deva Ramanan and David A Forsyth. Finding and tracking people from the bottom up. In *CVPR*, 2003. 2
- [54] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *ICML*, 2021. 2, 3
- [55] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *CVPR*, 2016. 2
- [56] Joseph Redmon and Ali Farhadi. Yolo9000: better, faster, stronger. In *CVPR*, 2017. 2
- [57] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NeurIPS*, 2015. 2, 4, 7
- [58] Stephan R Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. Playing for data: Ground truth from computer games. In *ECCV*, 2016. 3
- [59] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 2, 3, 6
- [60] Amir Sadeghian, Alexandre Alahi, and Silvio Savarese. Tracking the untrackable: Learning to track multiple cues with long-term dependencies. In *ICCV*, 2017. 2
- [61] Samuel Schuster, Paul Vernaza, Wongun Choi, and Manmohan Chandraker. Deep network flow for multi-object tracking. In *CVPR*, 2017. 2
- [62] Sarthak Sharma, Junaid Ahmed Ansari, J. Krishna Murthy, and K. Madhava Krishna. Beyond pixels: Leveraging geometry and shape cues for online multi-object tracking. In *ICRA*, 2018. 2
- [63] Peize Sun, Jinkun Cao, Yi Jiang, Rufeng Zhang, Enze Xie, Zehuan Yuan, Changhu Wang, and Ping Luo. Transtrack: Multiple object tracking with transformer. *arXiv preprint arXiv:2012.15460*, 2020. 2
- [64] 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*, 2020. 1
- [65] Nicolai Wojke, Alex Bewley, and Dietrich Paulus. Simple online and realtime tracking with a deep association metric. In *ICIP*, 2017. 2, 7
- [66] Sanghyun Woo, Kwanyong Park, Seoung Wug Oh, In So Kweon, and Joon-Young Lee. Bridging images and videos: A simple learning framework for large vocabulary video object detection. In *ECCV*, 2022. 3, 7
- [67] Sanghyun Woo, Kwanyong Park, Seoung Wug Oh, In So Kweon, and Joon-Young Lee. Tracking by associating clips. In *ECCV*, 2022. 7
- [68] Bin Xiao, Haiping Wu, and Yichen Wei. Simple baselines for human pose estimation and tracking. In *ECCV*, 2018. 2
- [69] Bo Yang and Ram Nevatia. An online learned CRF model for multi-target tracking. In *CVPR*, 2012. 2
- [70] Linjie Yang, Yuchen Fan, and Ning Xu. Video instance segmentation. In *ICCV*, 2019. 2, 4, 6
- [71] 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*, 2020. 1

- [72] Alireza Zareian, Kevin Dela Rosa, Derek Hao Hu, and Shih-Fu Chang. Open-vocabulary object detection using captions. In *CVPR*, 2021. 3
- [73] Fangao Zeng, Bin Dong, Tiancai Wang, Xiangyu Zhang, and Yichen Wei. Motr: End-to-end multiple-object tracking with transformer. *arXiv preprint arXiv:2105.03247*, 2021. 2
- [74] 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(11):3069–3087, 2021. 3, 5
- [75] 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*, 2022. 7, 8
- [76] Xingyi Zhou, Rohit Girdhar, Armand Joulin, Phillip Krähenbühl, and Ishan Misra. Detecting twenty-thousand classes using image-level supervision. In *ECCV*, 2022. 2, 3, 4
- [77] Xingyi Zhou, Vladlen Koltun, and Philipp Krähenbühl. Tracking objects as points. In *ECCV*, 2020. 3, 4, 5
- [78] Xingyi Zhou, Tianwei Yin, Vladlen Koltun, and Philipp Krähenbühl. Global tracking transformers. In *CVPR*, 2022. 2, 7, 8