TAO: A Large-Scale Benchmark for Tracking Any Object

Achal Dave[1], Tarasha Khurana[1], Pavel Tokmakov[1]
Cordelia Schmid[2], and Deva Ramanan[1,3]

[1] Carnegie Mellon University
[2] Inria [3] Argo AI

Abstract. For many years, multi-object tracking benchmarks have focused on a handful of categories. Motivated primarily by surveillance and self-driving applications, these datasets provide tracks for people, vehicles, and animals, ignoring the vast majority of objects in the world. By contrast, in the related field of object detection, the introduction of large-scale, diverse datasets (e.g., COCO) have fostered significant progress in developing highly robust solutions. To bridge this gap, we introduce a similarly diverse dataset for Tracking Any Object (TAO)[4]. It consists of 2,907 high resolution videos, captured in diverse environments, which are half a minute long on average. Importantly, we adopt a bottom-up approach for discovering a large vocabulary of 833 categories, an order of magnitude more than prior tracking benchmarks. To this end, we ask annotators to label objects that move at any point in the video, and give names to them post factum. Our vocabulary is both significantly larger and qualitatively different from existing tracking datasets. To ensure scalability of annotation, we employ a federated approach that focuses manual effort on labeling tracks for those relevant objects in a video (e.g., those that move). We perform an extensive evaluation of state-of-the-art trackers and make a number of important discoveries regarding large-vocabulary tracking in an open-world. In particular, we show that existing single- and multi-object trackers struggle when applied to this scenario in the wild, and that detection-based, multi-object trackers are in fact competitive with user-initialized ones. We hope that our dataset and analysis will boost further progress in the tracking community.

Keywords: datasets, video object detection, tracking

1 Introduction

A key component in the success of modern object detection methods was the introduction of large-scale, diverse benchmarks, such as MS COCO [38] and LVIS [27]. By contrast, multi-object tracking datasets tend to be small [40,56], biased towards short videos [65], and, most importantly, focused on a very small vocabulary of categories [40,56,60] (see Table 1). As can be seen from

[4] http://taodataset.org/

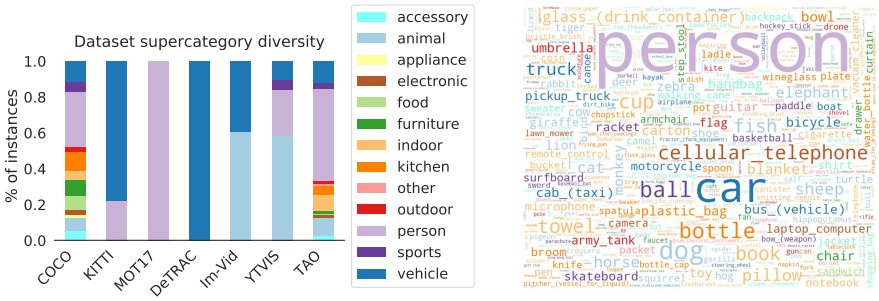

Fig. 1. (left) Super-category distribution in existing multi-object tracking datasets compared to TAO and COCO [38]. Previous work focused on people, vehicles and animals. By contrast, our bottom-up category discovery results in a more diverse distribution, covering many small, hand-held objects that are especially challenging from the tracking perspective. (right) Wordcloud of TAO categories, weighted by number of instances, and colored according to their supercategory.

Figure 1, they predominantly target people and vehicles. Due to the lack of proper benchmarks, the community has shifted towards solutions tailored to the few videos used for evaluation. Indeed, Bergmann et al. [5] have recently and convincingly demonstrated that simple baselines perform on par with state-of-the-art (SOTA) multi-object trackers.

In this work we introduce a large-scale benchmark for Tracking Any Object (TAO). Our dataset features 2,907 high resolution videos captured in diverse environments, which are 30 seconds long on average, and has tracks labeled for 833 object categories. We compare the statistics of TAO to existing multi-object tracking benchmarks in Table 1 and Figure 1, and demonstrate that it improves upon them both in terms of complexity and in terms of diversity (see Figure 2 for representative frames from TAO). Collecting such a dataset presents three main challenges: (1) how to select a large number of diverse, long, high-quality videos; (2) how to define a set of categories covering all the objects that might be of interest for tracking; and (3) how to label tracks for these categories at a realistic cost. Below we summarize our approach for addressing these challenges. A detailed description of dataset collection is provided in Section 4.

Existing datasets tend to focus on one or just a few domains when selecting the videos, such as outdoor scenes in MOT [40], or road scenes in KITTI [24]. This results in methods that fail when applied in the wild. To avoid this bias, we construct TAO with videos from as many environments as possible. We include indoor videos from Charades [52], movie scenes from AVA [26], outdoor videos from LaSOT [21], road-scenes from ArgoVerse [14], and a diverse sample of videos from HACS [68] and YFCC100M [54]. We ensure all videos are of high quality, with the smallest dimension larger or equal to 480px, and contain at least 2 moving objects. Table 1 reports the full statistics of the collected videos, showing that TAO provides an evaluation suite that is significantly larger, longer, and

Table 1. Statistics of major multi-object tracking datasets. TAO is by far the largest dataset in terms of the number of classes and total duration of evaluation videos. In addition, we ensure that each video is challenging (long, containing several moving objects) and high quality.

Dataset	Classes	Videos		Avg	Tracks	Min	Ann.	Total Eval
		Eval.	Train	length (s)	/ video	resolution	fps	length (s)
MOT17 [40]	1	7	7	35.4	112	640x480	30	248
KITTI [24]	2	29	21	12.6	52	1242x375	10	365
UA-DETRAC [60]	4	40	60	56	57.6	960x540	5	2,240
ImageNet-Vid [48]	30	1,314	4,000	10.6	2.4	480x270	~25	13,928
YTVIS [65]	40	645	2,238	4.6	1.7	320x240	5	2,967
TAO (Ours)	833	2,407	500	36.8	5.9	640x480	1	88,605

more diverse than prior work. Note that TAO contains fewer training videos than recent tracking datasets, as we intentionally dedicate the majority of videos for in-the-wild *benchmark* evaluation, the focus of our effort.

Given the selected videos, we must choose *what* to annotate. Most datasets are constructed with a top-down approach, where categories of interest are pre-defined by benchmark curators. That is, curators first select the subset of categories deemed relevant for the task, and then collect images or videos expressly for these categories [19,38,55]. This approach naturally introduces curator bias. An alternative strategy is bottom-up, open-world *discovery* of what objects are present in the data. Here, the vocabulary emerges post factum [26,27,69], an approach that dates back to LabelMe [49]. Inspired by this line of work, we devise the following strategy to discover an ontology of objects relevant for tracking: first annotators are asked to label *all* objects that either move by themselves or are moved by people. They then give names to the labeled objects, resulting in a vocabulary that is not only significantly larger, but is also qualitatively different from that of any existing tracking dataset (see Figure 1). To facilitate training of object detectors, that can be later used by multi-object trackers on our dataset, we encourage annotators to choose categories that exists in the LVIS dataset [27]. If no appropriate category can be found in the LVIS vocabulary, annotators can provide free-form names (see Section 4.2 for details).

Exhaustively labeling tracks for such a large collection of objects in 2,907 long videos is prohibitively expensive. Instead, we extend the federated annotation approach proposed in [27] to the tracking domain. In particular, we ask the annotators to label tracks for up to 10 objects in every video. We then separately collect exhaustive labels for every category for a subset of videos, indicating whether all the instances of the category have been labeled in the video. During evaluation of a particular category, we use only videos with exhaustive labels for computing precision and all videos for computing recall. This allows us to reliably measure methods' performance at a fraction of the cost of exhaustively annotating the videos. We use the LVIS federated mAP metric [27] for evaluation, replacing 2D IoU with 3D IoU [65]. For detailed comparisons, we further report the standard MOT challenge [40] metrics in supplementary.

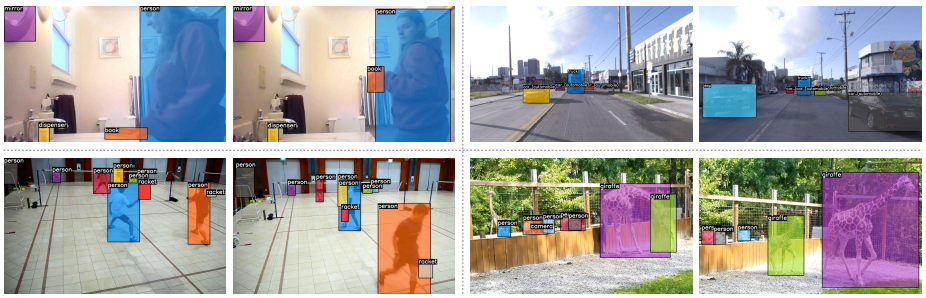

Fig. 2. Representative frames from TAO, showing videos sourced from multiple domains with annotations at two different timesteps.

Equipped with TAO, we set out to answer several questions about the state of the tracking community. In particular, in Section 5 we report the following discoveries: (1) SOTA trackers struggle to generalize to a large vocabulary of objects, particularly for infrequent object categories in the tail; (2) while trackers work significantly better for the most-explored category of people, tracking people in diverse scenarios (e.g., frequent occlusions or camera motion) remains challenging; (3) when scaled to a large object vocabulary, multi-object trackers become competitive with user-initialized trackers, despite the latter being provided with a ground truth initializations. We hope that these insights will help to define the most promising directions for future research.

2 Related work

The domain of object tracking is subdivided based on how tracks are initialized. Our work falls into the multi-object tracking category, where all objects out of a fixed vocabulary of classes must be detected and tracked. Other formulations include user-initialized and saliency-based tracking. In this section, we first review the most relevant benchmarks datasets in each of these areas, and then discuss SOTA methods for multi-object and user-initialized tracking.

2.1 Benchmarks

Multi-object tracking (MOT) is the task of tracking an unknown number of objects from a known set of categories. Most MOT benchmarks [23,24,40,60] focus on either people or vehicles (see Figure 1), motivated by surveillance and self-driving applications. Moreover, they tend to include only a few dozen videos, captured in outdoor or road environments, encouraging methods that are overly adapted to the benchmark and do not generalize to different scenarios (see Table 1). In contrast, TAO focuses on diversity both in the category and visual domain distribution, resulting in a realistic benchmark for tracking *any* object.

Several works have attempted to extend the MOT task to a wider vocabulary of categories. In particular, the ImageNet-Vid [48] benchmark provides exhaustive trajectories annotations for objects of 30 categories in 1314 videos. While this dataset is both larger and more diverse than standard MOT benchmarks, videos

tend to be relatively short and the categories cover only animals and vehicles. The recent YTVIS dataset [65] has the most broad vocabulary to date, covering 40 classes, but the majority of the categories still correspond to people, vehicles and animals. Moreover, the videos are 5 seconds long on average, making the tracking problem considerably easier in many cases. Unlike previous work, we take a bottom-up approach for defining the vocabulary. This results in not only the largest set of categories among MOT datasets to date, but also in a qualitatively different category distribution. In addition, our dataset is over 7 times larger than YTVIS in the number of frames. The recent VidOR dataset [51] explores Video Object Relations, including tracks for a large vocabulary of objects. But, since ViDOR focuses on relations rather than tracks, object trajectories tend to be missing or incomplete, making it hard to repurpose for tracker benchmarking. In contrast, we ensure TAO maintains high quality for both accuracy and completeness of labels (see supplementary for a quantitative analysis).

Finally, several recent works have proposed to label masks instead of bounding boxes for benchmarking multi-object tracking [56,65]. In collecting TAO we made a conscious choice to prioritize scale and diversity of the benchmark over pixel-accurate labeling. Instance mask annotations are significantly more expensive to collect than bounding boxes, and we show empirically that tracking at the box level is already a challenging task that current methods fail to solve.

User-initialized tracking forgoes a fixed vocabulary of categories and instead relies on the user to provide bounding box annotations for objects at need to be tracked at test time [21,30,34,55,61] (in particular, the VOT challenge [34] has driven the progress in this field for many years). The benchmarks in this category tend to be larger and more diverse than their MOT counterparts, but most still offer a tradeoff between the number of videos and the average length of the videos (see supplementary). Moreover, even if the task itself is category-agnostic, empirical distribution of categories in the benchmarks tends to be heavily skewed towards a few common objects. We study whether this bias in category selection results in methods failing to generalize to more challenging objects by evaluating state-of-the-art user-initialized trackers on TAO in Section 5.2.

Semi-supervised video object segmentation differs from user-initialized tracking in that both the input to the tracker and the output are object masks, not boxes [43,64]. As a result, such datasets are a lot more expensive to collect, and videos tend to be extremely short. The main focus of the works in this domain [12,33,57] is on accurate mask propagation, not solving challenging identity association problems, thus their effort is complementary to ours.

Saliency-based tracking is an intriguing direction towards open-world tracking, where the objects of interest are defined not with a fixed vocabulary of categories, or manual annotations, but with bottom-up, motion- [42,43] or appearance-based [13,59] saliency cues. Our work similarly uses motion-based saliency to define a comprehensive vocabulary of categories, but presents a significantly larger benchmark with class labels for each object, enabling the use and evaluation of large-vocabulary object recognition approaches.

2.2 Algorithms

Multi-object trackers for people and other categories have historically been studied by separate communities. The former have been mainly developed on the MOT benchmark [40] and follow the tracking-by-detection paradigm, linking outputs of person detectors in an offline, graph-based framework [3,4,10,20]. These methods mainly differ in the way they define the edge cost in the graph. Classical approaches use overlap between detections in consecutive frames [31,44,67]. More recent methods define edge costs based on appearance similarity [41,47], or motion-based models [1,15,16,35,45,50]. Very recently, Bergmann et al. [5] proposed a simple baseline approach for tracking people that performs on par with SOTA by repurposing an object detector's bounding box regression capability to predict the position of an object in the next frame. All these methods have been developed and evaluated on the relatively small MOT dataset, containing 14 videos captured in very similar environments. By contrast, TAO provides a much richer, more diverse set of videos, encouraging trackers more robust to tracking challenges such as occlusion and camera motion.

The more general multi-object tracking scenario is usually studied using ImageNet-Vid [48]. Methods in this group also use offline, graph-based optimization to link frame-level detections into tracks. To define the edge potentials, in addition to box overlap, Feichtenhofer et al. [22] propose a similarity embedding, which is learned jointly with the detector. Kang et al. [32] directly predict short tubelets, and Xiao et al. [63] incorporate a spatio-temporal memory module inside a detector. Inspired by [5], we show that a simple baseline relying on the Viterbi algorithm for linking detections [22,25] performs on par with the aforementioned methods on ImageNet-Vid. We then use this baseline for evaluating generic multi-object tracking on TAO in Section 5.2, and demonstrate that it struggles when faced with a large vocabulary and a diverse data distribution.

User-initialized trackers tend to rely on a Siamese network architecture that was first introduced for signature verification [11], and later adapted for tracking [7,18,29,53]. They learn a patch-level distance embedding and find the closest patch to the one annotated in the first frame in the following frames. To simplify the matching problem, state-of-the-art approaches limit the search space to the region in which the object was localized in the previous frame. Recently there have been several attempts to introduce some ideas from CNN architectures for object detection into Siamese trackers. In particular, Li et al. [37] use the similarity map obtained by matching the object template to the test frame as input to an RPN-like module adapted from Faster-RCNN [46]. Later this architecture was extended by introducing hard negative mining and template updating [71], as well as mask prediction [58]. In another line of work, Siamese-based trackers have been augmented with a target discrimination module to improve their robustness to distractors [9,17]. We evaluate several state-of-the-art methods in this paradigm for which public implementation is available [9,17,18,36,58] on TAO, and demonstrate that they achieve only a moderate improvement over our multi-object tracking baseline, despite being provided with a ground truth initialization for each track (see Section 5.2 for details).

3 Dataset design

Our primary goal is a large-scale video dataset with a diverse vocabulary of labeled objects to evaluate trackers in the wild. This requires designing a strategy for (1) video collection, (2) vocabulary discovery, (3) scalable annotation, and (4) evaluation. We detail our strategies for (2-4) below, and defer (1) to Section 4.1.

Category discovery. Rather than manually defining a set of categories, we discover an object vocabulary from unlabeled videos which span diverse operating domains. Our focus is on *dynamic* objects in the world. Towards this end, we ask annotators to mark all objects that *move* in our collection of videos, without any object vocabulary in mind. We then construct a vocabulary by giving names for all the discovered objects, following the recent trend for open-world dataset collection [27,69]. In particular, annotators are asked to provide a free-form name for every object, but are encouraged to select a category from the LVIS [27] vocabulary whenever possible. We detail this process further in Section 4.2.

Federation. Given this vocabulary, one option might be to exhaustively label all instances of each category in all videos. Unfortunately, exhaustive annotation of a large vocabulary is expensive, even for images [27]. We choose to use our labeling budget instead on collecting a large-scale, diverse dataset, by extending the federated annotation protocol [27] from image datasets to videos. Rather than labeling every video v with every category c, we define three subsets of our dataset for each category: P_c, containing videos where all instances of c are labeled, N_c, videos with no instance of c present in the video, and U_c, videos where *some* instances of c are annotated. Videos not belonging to any of these subsets are ignored when evaluating category c. For each category c, we only use videos in P_c and N_c to measure the *precision* of trackers, and videos in P_c and U_c to measure recall. We describe how to define P_c, N_c, and U_c in Section 4.2.

Granularity of annotations. To collect TAO, we choose to prioritize scale and diversity of the data at the cost of annotation granularity. In particular, we label tracks at 1 frame per second with bounding box labels but don't annotate segmentation masks. This allows us to label 833 categories in 2,907 videos at a relatively modest cost. Our decision is motivated by the observation of [55] that dense frame labeling does not change the relative performance of the methods.

Evaluation and metric. Traditionally, multi-object tracking datasets use either the CLEAR MOT metrics [6,24,40] or a 3D intersection-over-union (IoU) based metric [48,65]. We report the former in supplementary (with modifications for large-vocabularies of classes, including multi-class aggregation and federation), but focus our experiments on the latter. To formally define 3D IoU, let $G = \{g_1, \ldots, g_T\}$ and $D = \{d_1, \ldots, d_T\}$ be a groundtruth and predicted track for a video with T frames. 3D IoU is defined as: $\text{IoU}_{3\text{d}}(D, G) = \frac{\sum_{t=1}^{T} g_t \cap d_t}{\sum_{t=1}^{T} g_t \cup d_t}$. If an object is not present at time t, we assign g_t to an empty bounding box, and similarly for a missing detection. We choose 3D IoU (with a threshold of 0.5) as the default metric for TAO, and provide further analysis in supplementary.

Similar to standard object detection metrics, (3D) IoU together with (track) confidence can be used to compute mean average precision across categories. For

methods that provide a score for each frame in a track, we use the average frame score as the track score. Following [27], we measure precision for a category c in video v only if all instances of the category are verified to be labeled in it.

4 Dataset collection

4.1 Video selection

Most video datasets focus on one or a few domains. For instance, MOT benchmarks [40] correspond to urban, outdoor scenes featuring crowds, while AVA [26] contains produced films, typically capturing actors with close shots in carefully staged scenes. As a result, methods developed on any single dataset (and hence domain) fail to generalize in the wild. To avoid this bias, we constructed TAO by selecting videos from a variety of sources to ensure scene and object diversity.

Diversity. In particular, we used datasets for action recognition, self-driving cars, user-initialized tracking, and in-the-wild Flickr videos. In the action recognition domain we selected 3 datasets: Charades [52], AVA [26], and HACS [68]. Charades features complex human-human and human-object interactions, but all videos are indoor with limited camera motion. By contrast, AVA has a much wider variety of scenes and cinematographic styles but is scripted. HACS provides unscripted, in-the-wild videos. These action datasets are naturally focused on people and objects with which people interact. To include other animals and vehicles, we source clips from LaSOT [21] (a benchmark for user-initialized tracking), BDD [66] and ArgoVerse [14] (benchmarks for self-driving cars). LaSOT is a diverse collection whereas BDD and ArgoVerse consist entirely of outdoor, urban scenes. Finally we sample in-the-wild videos from the YFCC100M [54] Flickr collection.

Quality. The videos are automatically filtered to remove short videos and videos with a resolution below 480p. For longer videos, as in AVA, we use [39] to extract scenes without shot changes. In addition, we manually reviewed each sampled video to ensure it is high quality: i.e., we removed grainy videos as well as videos with excessive camera motion or shot changes. Finally, to focus on the most challenging tracking scenarios, we only kept videos that contain at least 2 moving objects. The full statistics of the collected videos are provided in Table 1. We point out that many prior video datasets tend to limit one or more quality dimensions (in terms of resolution, length, or number of videos) in order to keep evaluation and processing times manageable. In contrast, we believe that in order to truly enable tracking in the open-world, we need to appropriately scale benchmarks.

4.2 Annotation pipeline

Our annotation pipeline is illustrated in Figure 3. We designed it to separate low-level tracking from high-level semantic labeling. As pointed out by others [2], semantic labeling can be subtle and error-prone because of ambiguities and corner-cases that arise in category boundaries. By separating tasks into low vs high-level, we are able to take advantage of unskilled annotators for the former and highly-vetted workers for the latter.

Object mining and tracking. We combine object mining and track labeling into a single stage. Given the videos described above, we ask annotators to

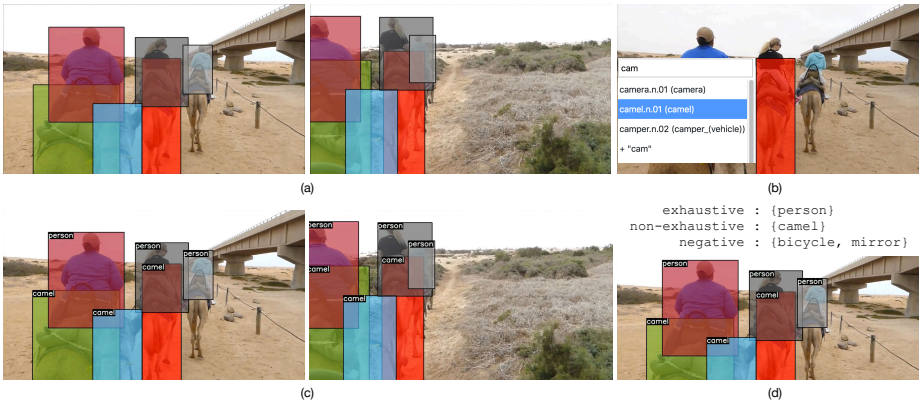

Fig. 3. Our federated video annotation pipeline. First (a), annotators mine and track moving objects. Second (b), annotators categorize tracks using the LVIS vocabulary or free-form text, producing the labeled tracks (c). Finally, annotators identify categories that are exhaustively annotated or verified to be absent. In (d), 'person's are identified as being exhaustively annotated, 'camel's are present but not exhaustively annotated and 'bicycle's and 'mirror's are verified as absent. These labels allow accurately penalizing false-positives and missed detections for exhaustively annotated and verified categories.

mark *objects that move at any point in the video*. To avoid overspending our annotation budget on a few crowded videos, we limited the number of labeled objects per video to 10. Note that this stage is *category-agnostic*: annotators are not instructed to look for objects from any specific vocabulary, but instead to use motion as a *saliency* cue for mining relevant objects. They are then asked to label these objects throughout the video with bounding boxes at 1 frame-per-second. Finally, the tracks are verified by one independent annotator. This process is illustrated in Figure 3, where we can see that 6 objects are discovered and tracked.

Object categorization. Next, we collected category labels for objects discovered in the previous stage and simultaneously constructed the dataset vocabulary. We focus on the large vocabulary from the LVIS [27] object detection dataset, which contains 1,230 synsets discovered in a bottom-up manner similar to ours. Doing so also allows us to make use of LVIS as a training set of relevant object detectors (which we later use within a tracking pipeline to produce strong baselines - Section 5.1). Because maintaining a mental list of 1,230 categories is challenging even for expert annotators, we use an auto-complete annotation interface to suggest categories from the LVIS vocabulary (Fig. 3 (b)). The autocomplete interface displays classes with a matching synset (e.g., "person.n.01"), name, synonym, and finally those with a matching definition. Interestingly, we find that some objects discovered in TAO, such as "door" or "marker cap", do not exist in LVIS. To accommodate such important exceptions, we allow annotators to label objects with free-form text if they do not fit in the LVIS vocabulary. Overall, annotators labeled 16,144 objects (95%) with 488 LVIS categories, and 894 objects (5%) with 345 free-form categories. We use the 488 LVIS categories

for MOT experiments (because detectors can be trained on LVIS), but use all categories for user-initialized tracking experiments in supplementary.

Federated "exhaustive" labeling. Finally, we ask annotators to verify which categories are exhaustively labeled for each video. For each category c labeled in video v, we ask annotators whether all instances of c are labeled. In Fig. 3, after this stage, annotators marked that 'person' is exhaustively labeled, while 'camel' is not. Next, we show annotators a sampled subset of categories that are not labeled in the video, and ask them to indicate which categories are absent in the video. In Fig. 3, annotators indicated that 'bicycle' and 'mirror' are absent.

4.3 Dataset splits

We split TAO into three subsets: train, validation and test, containing 500, 988 and 1,419 videos respectively. Typically, train splits tend to be larger than val and test. We choose to make TAO train small for several reasons. First, our primary goal is to reliably benchmark trackers in-the-wild. Second, most MOT systems are modularly trained using image-based detectors with hyper-parameter tuning of the overall tracking system. We ensure TAO train is sufficiently large for tuning, and that our large-vocabulary is aligned with the LVIS image dataset. This allows devoting most of our annotation budget for large-scale val and held-out test sets. We ensure that the videos in train, val and test are well-separated (e.g., each Charades subject appears in only one split); see supp. for details.

5 Analysis of state-of-the-art trackers

We now use TAO to analyze how well existing multi- and single-object trackers perform in the wild and when they fail. We tune the hyperparameters of each tracking approach on the 'train' set, and report results on the 'val' set. To capitalize on existing object detectors, we evaluate using the 488 LVIS categories in TAO. We begin by shortly describing the methods used in our analysis.

5.1 Methods

Detection. We analyze how well state-of-the-art object detectors perform on our dataset. To this end, we present results using a standard Mask R-CNN [46] detector trained using [62] in Section 5.2.

Multi-Object Tracking. We analyze SOTA multi-object tracking methods on ImageNet-Vid. We first clarify whether such approaches improve detection or tracking. Table 2 reports the standard ImageNet-Vid Detection and Track mAP. The 'Detection' row corresponds to a detection-only baseline widely reported by prior work [63,22,70]. D&T [22] and STMN [63] are spatiotemporal architectures that produce 6-7% detection mAP improvements over a per-frame detector. However, both D&T and STMN post-process their per-frame outputs using the Viterbi

Table 2. ImageNet-Vid detection and track mAP; see text (left) for details.

	Viterbi	Det mAP	Track mAP
Detection		73.4 [63]	-
D&T [22]	✓	79.8	-
STMN [63]	✓	79.0	60.4
Detection	✓	79.2	60.3

algorithm, which iteratively links and re-weights the confidences of per-frame detections (see [25]). *When the same post-processing is applied to a single-frame detector, one achieves nearly the same performance gain (Table 2, last row).*

Our analysis reinforces the bleak view of multi-object tracking progress suggested by [5]: while ever-more complex approaches have been proposed for the task, their improvements are often attributable to simple, baseline strategies. To foster meaningful progress on TAO, we evaluate a number of strong baselines. We evaluate a per-frame detector trained on LVIS [27] and COCO [38], followed by two linking methods: SORT [8], a simple, online linker initially proposed for tracking people, and the Viterbi post-processing step used by [22,63], in Section 5.2.

Person detection and tracking. Detecting and tracking people has been a distinct focus in the multi-object tracking community. Section 5.2 compares the above baselines to a recent SOTA people-tracker [5].

User-initialized tracking. We evaluate several recent user-initialized trackers for which public implementation is available [9,17,18,36,58]. Unfortunately, these trackers do not classify tracked objects, and cannot directly be compared to multi-object trackers which simultaneously detect and track objects. However, these trackers *can* be evaluated with an oracle classifier, enabling direct comparisons.

Oracles. Finally, to disentangle the complexity of classification and tracking, we use two oracles. The first, a class oracle, computes the best matching between predicted and groundtruth tracks. Predicted tracks that match to a groundtruth track with 3D IoU > 0.5 are assigned the corresponding groundtruth category. Tracks that do not match to a groundtruth track are not modified, and count as false positives. This allows us to evaluate the performance of trackers assuming the semantic *classification* task is solved. The second oracle computes the best possible assignment of per-frame detections to tracks, by comparing them with groundtruth. When doing so, class predictions for each detection are held constant. Any detections that are not matched are discarded. This oracle allows us to analyze the best performance we could expect given a fixed set of detections.

5.2 Results

How hard is object detection on TAO? We start by assessing the difficulty of detection on TAO by evaluating the SOTA object detector [28] using detection mAP. We train this model on LVIS and COCO, as training on LVIS alone led to low accuracy in detecting people. The final model achieves 27.1 mAP on TAO val at IoU 0.5, suggesting that single-frame detection is challenging on TAO.

Do multi-object trackers generalize to TAO? Table 3 reports results using tracking mAP on TAO. As a sanity check, we first evaluate a per-frame detector by assigning each detection to its own track. As expected, this achieves an mAP of nearly 0 (which isn't quite 0 due to the presence of short tracks).

Next, we evaluate two multi-object tracking approaches. We compare the Viterbi linking method to an online SORT tracker [8]. We tune SORT on our diverse 'train' set, which is key for good accuracy. As the offline Viterbi algorithm takes over a month to run on TAO train, we only tune the post-processing score threshold for reporting a detection at each frame. See supplementary for tuning

Method	Oracle Class	Oracle Track	Track mAP
Detection			0.6
Viterbi [22,25]			6.3
SORT [8]			13.2
Detection		✓	31.5
Viterbi [22,25]	✓		15.7
SORT [8]	✓		30.2
Detection	✓	✓	83.6

Table 3. SORT and Viterbi linking provide strong baselines on TAO, but detection and tracking remain challenging. Relabeling and linking detections from current detectors using the class and track oracles leads to high performance, suggesting a pathway for progress on TAO.

Fig. 4. SORT qualitative results, showing (left) a successful tracking result, and (right) a failure case due to semantic flicker between similar classes, suggesting that large-vocabulary tracking on TAO requires additional machinery.

details. Surprisingly, we find that the simpler, online SORT approach outperforms Viterbi, perhaps because the latter has been heavily tuned for ImageNet-Vid. Because of its scalablity (to many categories and long videos) and relatively better performance, we focus on SORT for the majority of our experiments. However, the performance of both methods remains low, suggesting TAO presents a major challenge for the tracking community, requiring principled novel approaches.

To better understand the nature of the complexity of TAO, we separately measure the challenges of tracking and classification. To this end, we first evaluate the "track" oracle that perfectly links per-frame detections. It achieves a stronger mAP of 31.5, compared to 13.2 for SORT. Interestingly, providing SORT tracks with an oracle class label provides a similar improvement, boosting mAP to 30.2. We posit that these improvements are orthogonal, and verify this by combining them; we link detections with oracle tracks and assign these tracks oracle class labels. This provides the largest delta, dramatically improving mAP to 83.6%. This suggests that *large-vocabulary tracking requires jointly improving tracking and classification accuracy (e.g., reducing semantic flicker as shown in Fig. 4).*

How well can we track people? We now evaluate tracking on one particularly important category: people. Measuring AP for individual categories in a federated dataset can be noisy [27], so we emphasize *relative* performance of trackers rather than their absolute AP. We evaluate Tracktor++ [5], the state-of-the-art method designed specifically for people tracking, and compare it to the SORT and Viterbi baselines in Table 4. We update Tracktor++ to use the same detector used by SORT and Viterbi, using only the 'person' predictions. We tune the score threshold on TAO 'train', but find the method is largely robust to this parameter (see supp.). Track-

Table 4. Person-tracking results on TAO. See text (left).

Method	Person AP
Viterbi [22,25]	16.5
SORT [8]	18.5
Tracktor++ [5]	36.7

tor++ strongly outperforms other approaches (36.7 AP), while SORT modestly outperforms Viterbi (18.6 vs 16.5 AP). It is interesting to note that SORT, which can scale to all object categories, performs noticeably worse on all categories on average (13.2 mAP). This delta between 'person' and overall is even more dramatic using the MOTA metric (6.7 overall vs 54.8 for 'person'; see supp.). We attribute the higher accuracy for 'person' to two factors: (1) a rich history of focused research on this one category, which has led to more accurate detectors and trackers, and (2) more complex categories present significant challenges, such as hand-held objects which exhibit frequent occlusions during interactions.

To further investigate Tracktor++'s performance, we evaluate a simpler variant of the method from [5], which does not use appearance-based re-identification nor pixel-level frame alignment. We find that removing these components reduces AP from 36.7 to 25.9, suggesting these components contribute to a majority of improvements over the baselines. Our results contrast [5], which suggests that re-id and frame alignment are not particularly helpful. *Compared to prior benchmarks, the diversity of TAO results in a challenging testbed for person tracking which encourages trackers robust to occlusion and camera jitter.*

Do user-initialized trackers generalize better? Next, we evaluate recent user-initialized trackers in Table 5. We provide the tracker with the groundtruth box for each object from its first visible frame. As these trackers do not report when an object is *absent* [55], we modify them to report an object as absent when the confidence drops below a threshold tuned on TAO 'train' (see supp).

Table 5. User-initialized trackers on 'val'. We re-train some trackers on their train set with TAO videos removed, denoted *.

Method	Oracle Init	Oracle Class	Track mAP
SORT		✓	30.2
ECO [18]	✓	✓	23.7
SiamMask [58]	✓	✓	30.8
SiamRPN++ LT [36]	✓	✓	27.2
SiamRPN++ [36]	✓	✓	29.7
ATOM* [17]	✓	✓	30.9
DIMP* [9]	✓	✓	**33.2**

We compare these trackers to SORT, supplying both with a class oracle. As expected, the use of a ground-truth initialization allows the best user-initialized methods to outperform the multi-object tracker. However, even with this oracle box initialization and an oracle classifier, tracking remains challenging on TAO. Indeed, most user-initialized trackers provide at most modest improvements over SORT. We provide further analysis in supplementary, showing that (1) while a more informative initialization frame improves user-initialized tracker accuracy, SORT remains competitive, and (2) user-initialized trackers accurately track for a few frames after init, leading to improvements in MOTA, but provide little benefits in long-term tracking. We hypothesize that the small improvement of user-initialized trackers over SORT is due to the fact that the former are trained on a small vocabulary of objects with limited occlusions, leading to methods that do not generalize to the most challenging cases in TAO. One goal of user-initialized trackers is open-world tracking of objects without good detectors. TAO's large vocabulary allows us to analyze progress towards this goal, indicating

that *large-vocabulary multi-object trackers may now address the open-world of objects as well as category-agnostic, user-initialized trackers.*

6 Discussion

Developing tracking approaches that can be deployed in-the-wild requires being able to reliably measure their performance. With nearly 3,000 videos, TAO provides a robust evaluation benchmark. Our analysis provides new conclusions about the state of tracking, while raising important questions for future work.

The role of user-initialized tracking. User-initialized trackers aim to track *any* object, without requiring category-specific detectors. In this work, we raise a provocative question: with the advent of large vocabulary object detectors [27], to what extent can (detection-based) multi-object trackers perform generic tracking *without* user initialization? Table 5, for example, shows that large-vocabulary datasets (such as TAO and LVIS) now allow multi-object trackers to match or outperform user-initialization for a number of categories.

Specialized tracking approaches. TAO aims to measure progress in tracking in-the-wild. A valid question is whether progress may be better achieved by building trackers for *application-specific* scenarios. An indoor robot, for example, has little need for tracking elephants. However, success in many computer vision fields has been driven by the pursuit of *generic* approaches, that can then be tailored for specific applications. We do not build one class of object detectors for indoor scenes, and another for outdoor scenes, and yet another for surveillance videos. We believe that tracking will similarly benefit from targeting diverse scenarios. Of course, due to its size, TAO also lends itself to use for evaluating trackers for specific scenarios or categories, as in Section 5.2 for 'person.'

Video object detection. Although image-based object detectors have significantly improved in recent years, our analysis in Section 5.1 suggests that simple post-processing of detection outputs remains a strong baseline for detection in videos. While we do not emphasize it in this work, TAO can also be used to measure progress in video object *detection*, where the goal is not to maintain the identity of objects, but only to reliably detect them in every video frame. TAO's large vocabulary particularly provides avenues for incorporating temporal information to resolve classification errors, which remain challenging (see Fig. 4).

Acknowledgements. We thank Jonathon Luiten and Ross Girshick for detailed feedback, and Nadine Chang and Kenneth Marino for reviewing early drafts. Annotations for this dataset were provided by Scale.ai. This work was supported in part by the CMU Argo AI Center for Autonomous Vehicle Research, the Inria associate team GAYA, and by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior/Interior Business Center (DOI/IBC) contract number D17PC00345. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes not withstanding any copyright annotation theron. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied of IARPA, DOI/IBC or the U.S. Government.

References

1. Alahi, A., Goel, K., Ramanathan, V., Robicquet, A., Fei-Fei, L., Savarese, S.: Social LSTM: Human trajectory prediction in crowded spaces. In: CVPR (2016) 6
2. Barriuso, A., Torralba, A.: Notes on image annotation. arXiv preprint arXiv:1210.3448 (2012) 8
3. Berclaz, J., Fleuret, F., Fua, P.: Robust people tracking with global trajectory optimization. In: CVPR (2006) 6
4. Berclaz, J., Fleuret, F., Turetken, E., Fua, P.: Multiple object tracking using k-shortest paths optimization. IEEE Transactions on Pattern Analysis and Machine Intelligence **33**(9), 1806–1819 (2011) 6
5. Bergmann, P., Meinhardt, T., Leal-Taixe, L.: Tracking without bells and whistles. In: ICCV (2019) 2, 6, 11, 12, 13
6. Bernardin, K., Stiefelhagen, R.: Evaluating multiple object tracking performance: the clear mot metrics. Journal on Image and Video Processing **2008**, 1 (2008) 7
7. Bertinetto, L., Valmadre, J., Henriques, J.F., Vedaldi, A., Torr, P.H.: Fully-convolutional Siamese networks for object tracking. In: ECCV (2016) 6
8. Bewley, A., Ge, Z., Ott, L., Ramos, F., Upcroft, B.: Simple online and realtime tracking. In: ICIP (2016) 11, 12
9. Bhat, G., Danelljan, M., Gool, L.V., Timofte, R.: Learning discriminative model prediction for tracking. In: CVPR (2019) 6, 11, 13
10. Breitenstein, M.D., Reichlin, F., Leibe, B., Koller-Meier, E., Van Gool, L.: Robust tracking-by-detection using a detector confidence particle filter. In: ICCV (2009) 6
11. Bromley, J., Guyon, I., LeCun, Y., Säckinger, E., Shah, R.: Signature verification using a "Siamese" time delay neural network. In: NIPS (1994) 6
12. Caelles, S., Maninis, K.K., Pont-Tuset, J., Leal-Taixé, L., Cremers, D., Van Gool, L.: One-shot video object segmentation. In: CVPR (2017) 5
13. Caelles, S., Pont-Tuset, J., Perazzi, F., Montes, A., Maninis, K.K., Van Gool, L.: The 2019 DAVIS challenge on VOS: Unsupervised multi-object segmentation. arXiv preprint arXiv:1905.00737 (2019) 5
14. Chang, M.F., Lambert, J., Sangkloy, P., Singh, J., Bak, S., Hartnett, A., Wang, D., Carr, P., Lucey, S., Ramanan, D., et al.: Argoverse: 3d tracking and forecasting with rich maps. In: CVPR (2019) 2, 8
15. Chen, B., Wang, D., Li, P., Wang, S., Lu, H.: Real-time 'Actor-Critic' tracking. In: ECCV (2018) 6
16. Choi, W., Savarese, S.: Multiple target tracking in world coordinate with single, minimally calibrated camera. In: ECCV (2010) 6
17. Danelljan, M., Bhat, G., Khan, F.S., Felsberg, M.: Atom: Accurate tracking by overlap maximization. In: CVPR (2019) 6, 11, 13
18. Danelljan, M., Bhat, G., Shahbaz Khan, F., Felsberg, M.: Eco: Efficient convolution operators for tracking. In: CVPR (2017) 6, 11, 13
19. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: CVPR (2009) 3
20. Ess, A., Leibe, B., Schindler, K., Van Gool, L.: A mobile vision system for robust multi-person tracking. In: CVPR (2008) 6
21. 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: CVPR (2019) 2, 5, 8
22. Feichtenhofer, C., Pinz, A., Zisserman, A.: Detect to track and track to detect. In: ICCV (2017) 6, 10, 11, 12

23. Fisher, R., Santos-Victor, J., Crowley, J.: Context aware vision using image-based active recognition. EC's Information Society Technology's Programme Project IST2001-3754 (2001) 4
24. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? The KITTI vision benchmark suite. In: CVPR (2012) 2, 3, 4, 7
25. Gkioxari, G., Malik, J.: Finding action tubes. In: CVPR (2015) 6, 11, 12
26. Gu, C., Sun, C., Ross, D.A., Vondrick, C., Pantofaru, C., Li, Y., Vijayanarasimhan, S., Toderici, G., Ricco, S., Sukthankar, R., et al.: AVA: A video dataset of spatio-temporally localized atomic visual actions. In: CVPR (2018) 2, 3, 8
27. Gupta, A., Dollar, P., Girshick, R.: LVIS: A dataset for large vocabulary instance segmentation. In: CVPR (2019) 1, 3, 7, 8, 9, 11, 12, 14
28. He, K., Gkioxari, G., Dollár, P., Girshick, R.: Mask R-CNN. In: ICCV (2017) 11
29. Held, D., Thrun, S., Savarese, S.: Learning to track at 100 FPS with deep regression networks. In: ECCV (2016) 6
30. Huang, L., Zhao, X., Huang, K.: GOT-10k: A large high-diversity benchmark for generic object tracking in the wild. arXiv preprint arXiv:1810.11981 (2018) 5
31. Jiang, H., Fels, S., Little, J.J.: A linear programming approach for multiple object tracking. In: CVPR (2007) 6
32. Kang, K., Li, H., Xiao, T., Ouyang, W., Yan, J., Liu, X., Wang, X.: Object detection in videos with tubelet proposal networks. In: CVPR (2017) 6
33. Khoreva, A., Benenson, R., Ilg, E., Brox, T., Schiele, B.: Lucid data dreaming for video object segmentation. International Journal of Computer Vision **127**(9), 1175–1197 (2019) 5
34. Kristan, M., Matas, J., Leonardis, A., Vojíř, T., Pflugfelder, R., Fernandez, G., Nebehay, G., Porikli, F., Čehovin, L.: A novel performance evaluation methodology for single-target trackers. TPAMI **38**(11), 2137–2155 (2016) 5
35. Leal-Taixé, L., Fenzi, M., Kuznetsova, A., Rosenhahn, B., Savarese, S.: Learning an image-based motion context for multiple people tracking. In: CVPR (2014) 6
36. Li, B., Wu, W., Wang, Q., Zhang, F., Xing, J., Yan, J.: Siamrpn++: Evolution of siamese visual tracking with very deep networks. In: CVPR (2019) 6, 11, 13
37. Li, B., Yan, J., Wu, W., Zhu, Z., Hu, X.: High performance visual tracking with siamese region proposal network. In: CVPR (2018) 6
38. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: Common objects in context. In: ECCV (2014) 1, 2, 3, 11
39. Lokoč, J., Kovalčík, G., Souček, T., Moravec, J., Čech, P.: A framework for effective known-item search in video. In: ACMM (2019). https://doi.org/10.1145/3343031.3351046, https://doi.org/10.1145/3343031.3351046 8
40. Milan, A., Leal-Taixé, L., Reid, I., Roth, S., Schindler, K.: MOT16: A benchmark for multi-object tracking. arXiv preprint arXiv:1603.00831 (2016) 1, 2, 3, 4, 6, 7, 8
41. Milan, A., Rezatofighi, S.H., Dick, A., Reid, I., Schindler, K.: Online multi-target tracking using recurrent neural networks. In: Thirty-First AAAI Conference on Artificial Intelligence (2017) 6
42. Ochs, P., Malik, J., Brox, T.: Segmentation of moving objects by long term video analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence **36**(6), 1187–1200 (2013) 5
43. Perazzi, F., Pont-Tuset, J., McWilliams, B., Van Gool, L., Gross, M., Sorkine-Hornung, A.: A benchmark dataset and evaluation methodology for video object segmentation. In: CVPR (2016) 5

44. Pirsiavash, H., Ramanan, D., Fowlkes, C.C.: Globally-optimal greedy algorithms for tracking a variable number of objects. In: CVPR (2011) 6
45. Ren, L., Lu, J., Wang, Z., Tian, Q., Zhou, J.: Collaborative deep reinforcement learning for multi-object tracking. In: ECCV (2018) 6
46. Ren, S., He, K., Girshick, R., Sun, J.: Faster R-CNN: Towards real-time object detection with region proposal networks. In: NIPS (2015) 6, 10
47. Ristani, E., Tomasi, C.: Features for multi-target multi-camera tracking and re-identification. In: CVPR (2018) 6
48. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: ImageNet large scale visual recognition challenge. International Journal of Computer Vision **115**(3), 211–252 (2015) 3, 4, 6, 7
49. Russell, B.C., Torralba, A., Murphy, K.P., Freeman, W.T.: LabelMe: a database and web-based tool for image annotation. International journal of computer vision **77**(1-3), 157–173 (2008) 3
50. Scovanner, P., Tappen, M.F.: Learning pedestrian dynamics from the real world. In: ICCV (2009) 6
51. Shang, X., Di, D., Xiao, J., Cao, Y., Yang, X., Chua, T.S.: Annotating objects and relations in user-generated videos. In: ICMR (2019) 5
52. Sigurdsson, G.A., Varol, G., Wang, X., Farhadi, A., Laptev, I., Gupta, A.: Hollywood in homes: Crowdsourcing data collection for activity understanding. In: ECCV (2016) 2, 8
53. Tao, R., Gavves, E., Smeulders, A.W.: Siamese instance search for tracking. In: CVPR (2016) 6
54. 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. arXiv preprint arXiv:1503.01817 (2015) 2, 8
55. 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: ECCV (2018) 3, 5, 7, 13
56. Voigtlaender, P., Krause, M., Osep, A., Luiten, J., Sekar, B.B.G., Geiger, A., Leibe, B.: MOTS: Multi-object tracking and segmentation. In: CPVR (2019) 1, 5
57. Voigtlaender, P., Leibe, B.: Online adaptation of convolutional neural networks for video object segmentation. In: BMVC (2017) 5
58. Wang, Q., Zhang, L., Bertinetto, L., Hu, W., Torr, P.H.: Fast online object tracking and segmentation: A unifying approach. In: CVPR (2019) 6, 11, 13
59. Wang, W., Song, H., Zhao, S., Shen, J., Zhao, S., Hoi, S.C., Ling, H.: Learning unsupervised video object segmentation through visual attention. In: CVPR (2019) 5
60. 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. arXiv preprint arXiv:1511.04136 (2015) 1, 3, 4
61. Wu, Y., Lim, J., Yang, M.H.: Online object tracking: A benchmark. In: CVPR (2013) 5
62. Wu, Y., Kirillov, A., Massa, F., Lo, W.Y., Girshick, R.: Detectron2. https://github.com/facebookresearch/detectron2 (2019) 10
63. Xiao, F., Jae Lee, Y.: Video object detection with an aligned spatial-temporal memory. In: ECCV (2018) 6, 10, 11
64. Xu, N., Yang, L., Fan, Y., Yue, D., Liang, Y., Yang, J., Huang, T.: Youtube-VOS: A large-scale video object segmentation benchmark. arXiv preprint arXiv:1809.03327 (2018) 5

65. Yang, L., Fan, Y., Xu, N.: Video instance segmentation. In: ICCV (2019) 1, 3, 5, 7
66. 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: CVPR (June 2020) 8
67. Zhang, L., Li, Y., Nevatia, R.: Global data association for multi-object tracking using network flows. In: CVPR (2008) 6
68. Zhao, H., Torralba, A., Torresani, L., Yan, Z.: HACS: Human action clips and segments dataset for recognition and temporal localization. In: ICCV (2019) 2, 8
69. Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., Torralba, A.: Scene parsing through ADE20k dataset. In: CVPR (2017) 3, 7
70. Zhu, X., Wang, Y., Dai, J., Yuan, L., Wei, Y.: Flow-guided feature aggregation for video object detection. 2017 IEEE International Conference on Computer Vision (ICCV) pp. 408–417 (2017) 10
71. Zhu, Z., Wang, Q., Li, B., Wu, W., Yan, J., Hu, W.: Distractor-aware siamese networks for visual object tracking. In: ECCV (2018) 6