Opening up Open World Tracking

Yang Liu[1,*] Idil Esen Zulfikar[2,*] Jonathon Luiten[2,3,*] Achal Dave[3,*]
Deva Ramanan[3] Bastian Leibe[2] Aljoša Ošep[1,3] Laura Leal-Taixé[1]

[1]Technical University of Munich, Germany [2]RWTH Aachen University, Germany [3]Carnegie Mellon University, USA

[1]{yang14.liu, aljosa.osep, leal.taixe}@tum.de [3]{achald, deva}@cs.cmu.edu

[2]{zulfikar, luiten, leibe}@vision.rwth-aachen.de

openworldtracking.github.io

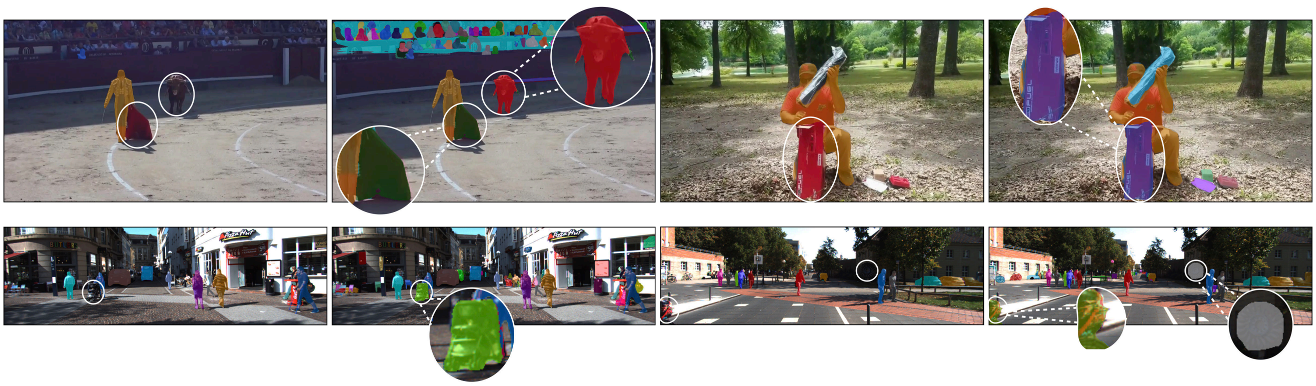

Figure 1. *Each pair left:* The standard approach to multi-object tracking is to detect, track and possibly segment objects that correspond to specific, pre-defined semantic classes, such as cars and pedestrians [78]. *Each pair right:* The output of our tracking baseline, that can track objects, such as child stroller, that was not labeled in the model training set. The significant contribution of this paper is the first benchmark, designed for studying the performance of object trackers in such open world conditions, in which trackers are only given a partial knowledge about the visual world, embracing the fact that *one could never train object detectors for every possible semantic class*.

Abstract

Tracking and detecting any object, including ones never-seen-before during model training, is a crucial but elusive capability of autonomous systems. An autonomous agent that is blind to never-seen-before objects poses a safety hazard when operating in the real world – and yet this is how almost all current systems work. One of the main obstacles towards advancing tracking any *object is that this task is notoriously difficult to evaluate. A benchmark that would allow us to perform an apples-to-apples comparison of existing efforts is a crucial first step towards advancing this important research field. This paper addresses this evaluation deficit and lays out the landscape and evaluation methodology for detecting and tracking both known and unknown objects in the open-world setting. We propose a new benchmark, TAO-OW: Tracking Any Object in an Open World, analyze existing efforts in multi-object tracking, and construct a baseline for this task while highlighting future challenges. We hope to open a new front in multi-object tracking research that will hopefully bring us a step closer to intelligent systems that can operate safely in the real world.*

* These authors contributed equally to this work.

1. Introduction

Understanding common scenarios is easy. Vision systems, trained on millions of examples of cars and pedestrians, work pretty well at detecting these objects, determining what and where they are, and tracking them through a scene. *Understanding never-seen-before scenarios is* ***extremely*** *hard.* What happens when a plane lands on the road in front an autonomous vehicle? Or a new children's toy is thrown onto the road? How will current vision systems be able to handle these previously unseen and unknown situations? Will a system designed to detect and track potentially hazardous objects pick up on these at all? Or will they be completely ignored with disastrous consequences (such as a vehicle hitting the *child stroller* in Fig. 1, *bottom-left*)?

Tracking and detection methods work reasonably well for objects that have a huge amount of data collected on them. But without building systems that can deal with never-seen-before objects, vision systems will never be safe enough to work in the real world and collecting more data can never scale up to address the infinite variety of possible unknown things that can happen. Many anecdotal examples indicate that current vision systems perform poorly in previously unseen scenarios [60], but we cannot quantitatively measure this phenomenon, or even evaluate progress, because there are no benchmarks on which to evaluate.

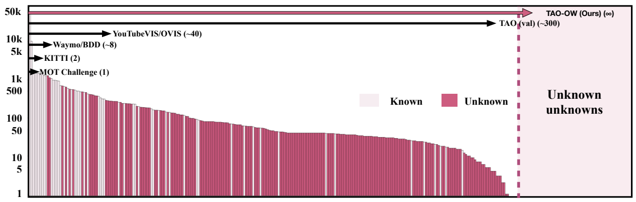

Figure 2. **TAO-OW Benchmark** class distribution in the validation set, showing *known* classes for which training data is given, and the *unknown* classes which serve as a proxy for the infinite variety (*unknown unknowns*) of objects which may appear in an open-world. Note the y-axis is **log-scaled**.

In this paper we present a new benchmark (**TAO-OW**: *Tracking Any Object in an Open World*) for measuring detection and tracking performance in an open-world setting. Closed-world multi-object tracking benchmarks [16,18,24] and methods [7,40,78] focus on tracking object classes that belong to a predefined set of frequently observed classes. In contrast, in our Open-World Tracking (OWT) task, *all* object must be tracked, and methods are specifically evaluated on how well they can track object classes that they weren't allowed to train on (*unknown* objects), as well as objects which were in the training set (*known* objects).

Open-World evaluation is inherently difficult. One has to restrict the set of objects that algorithms are allowed to train on. These *known* objects should be varied and diverse enough to represent the set of objects that could typically be expected to have data collected for, but there should be plentiful examples of further *unknown* objects, not presented as labeled samples to the models being evaluated. We base our work upon the recently introduced TAO dataset [16][1], which contains a large corpus of videos from many diverse scenarios such as driving, movies, and everyday scenes. Such a wide diversity is important in order to be able to capture a wide range of *unknown* objects. For *known* classes we use the 80 classes from COCO [42], which cover a wide range of common objects, while leaving over 700 *unknown* object categories to evaluate the performance of algorithms on objects for which they have not been trained. In Fig. 2 we show our TAO-OW benchmark, with its inherently long tailed distribution of object categories, its *known* and *unknown* split, and a comparison to previous tracking benchmarks [18,24,64,75,88,91], which are all limited to closed-world evaluation on a small number of categories.

Another inherent difficulty with open-world evaluation is dealing with the fact that it is impossible to exhaustively annotate the complete set of objects which should be detected and tracked (by definition, we do not want to penalize trackers for tracking unknown, unannotated objects). To tackle this issue, we propose a new evaluation metric called Open-World Tracking Accuracy (OWTA) which naturally decomposes detection and tracking evaluation components allowing the evaluation of tracking accuracy in the setting where extra unannotated object detections are not penalized. Such evaluation is enabled by the constraint that proposed objects must be supplied as non-overlapping segmentation masks.

Armed with our Open-World Tracking benchmark and evaluation methodology, we analyze several methods which have attempted this task but have lacked a common evaluation protocol [17, 47, 56]. A significant contribution of this paper is our thorough analysis of a wide variety of approaches. This analysis leads us to propose an open-world tracking approach which currently performs the best on our Open-World Tracking Benchmark, while also performing very competitively on previous closed-world benchmarks, even though it was not designed or tuned for these.

In summary, the **main contribution** of this work is to open up a new direction in vision-based multi-object tracking that goes beyond current closed-world benchmarks. We formalize the Open-World Tracking problem, (i) propose a benchmark with a suitable recall-based evaluation to measure progress, (ii) analyze existing design paradigms, providing a large collection of baselines based on state-of-the-art approaches from the closed-world setting, and (iii) present a strong method which works well for both open- and closed-world tracking. Our experiments show that closed-world detectors work surprisingly well for *localizing* even unknown objects. However, *tracking* unknown objects remains more challenging than known objects.

2. Related Work

Related tasks and benchmarks. Multi-object tracking (MOT) is a challenging task which involves localizing objects in both space and time, often in dense, crowded environments. Existing MOT datasets focus on closed-set tracking on video [18, 24, 83, 90] or LiDAR streams [14, 75]. Recent efforts move towards pixel-precise segmentation of tracked objects in video [36, 48, 78, 82, 90] or LiDAR sequences [3], and study performance in the long tail of object classes [16]. Closer to our work is unsupervised video object segmentation (UVOS) [13] and motion segmentation [9, 30, 71], where multiple objects that are present throughout the video and exhibit dominant motion need to be tracked and segmented. However, almost all classes in these benchmarks exist in COCO, and almost all methods [2,47] achieve excellent performance by training on COCO. Our work explicitly evaluates on classes *beyond* COCO.

Multi-object tracking. Early methods in vision-based tracking [27, 59, 85] and robotic perception [54, 76] utilized class-agnostic, bottom-up segmentation as a tracking cue, *e.g.*, based on LiDAR point cloud clustering [53, 79] or background modeling and foreground grouping [32, 73, 85]. A step forward in vision-based MOT was the *tracking-by-detection* paradigm, which relies on pre-trained object de-

[1]License available at taodataset.org.

tectors. Early effort focused on developing strong data association techniques [11,39,41,50,63,65] and hand-crafting appearance [23, 49, 52] and motion cues [15, 38]. More recent efforts are largely data-driven, learning strong appearance models [35, 37], learning to regress targets [7] and to associate detections using graph neural networks [10]. This progress in closed-set multi-object tracking is largely thanks to efforts in releasing new datasets, benchmarks, and evaluation metrics. However, MOT is currently only evaluated in well-controlled, closed-set domains, where object classes are known a priori and are present in training sets.

Beyond closed-world tracking. Tracking-by-detection approaches have been generalized to generic objects [17, 56, 58] and UVOS [17, 46, 47, 87], using object proposal methods trained in a category-agnostic manner [28, 62]. However, until recently, there was no suitable evaluation methodology for the open-world domain, making it unclear how such methods generalized to arbitrary objects.

Recent **parallel work** [80] focuses on labeling a variety of object classes in human-centric Kinetics400 dataset [34]. This work focuses on data collection and proposes to use the existing closed-world Track mAP [90] metric for evaluation. This metric has recently been heavily criticized [45] due poor interpretability, lack of sensitivity and a lack of error-type differentiability, which are especially problematic for evaluating tracking in the open world. Furthermore, by default, this metric requires exhaustively labeling all objects, which is infeasible in practice. The data is also limited to human-centric activities. In contrast, we study open-world tracking in a significantly more diverse setting including videos from multiple different domains, which is crucial for studying open-world problems, resulting in less bias and more generalization (*e.g.* avoiding that objects always appear in the center of frames). Finally, we analyze prior work on open-world tracking and identify building blocks of these methods to perform a thorough evaluation of these efforts and devise a new baseline, shown to work very well in both, open- and closed-world conditions.

Open-set recognition, detection and segmentation. Open-set recognition methods [6, 31, 69, 70] focus on minimizing the confusion between *known* object classes, presented to the model during the training, and *unknown* object classes, that may (only) appear in the open world. Object detection has recently been studied in open-set conditions [20, 51]. By contrast, open-world recognition methods, as defined by [5, 43] must explicitly recognize *unknown* object instances that were not observed during training, and update object detectors to recognize these unknown instances. Learning to detect *unknown* objects in automotive scenarios was tackled in [55], where object detectors were re-trained using clusters of *unknown* object tracks [56, 57], mined from video. Similarly, [29] learns to detect *unknown* object instances by sampling and clustering object proposals from the *void* regions from labeled images and using these clusters as pseudo-labels during model training. *Joseph* et al. [33] propose an extension to Faster R-CNN [66] for distinguishing known/unknown classes by adding a contrastive objective, that maximizes the margin between *known* and *unknown* objects in feature space. Unlike these previous works, we do not study how to minimize the confusion between *known* or *unknown* semantic classes or tackle incremental learning. We study how well we can identify and track objects from both *known* and *unknown* classes, and we do not require semantic interpretation of tracked objects. Instead, we advocate for the view that any-object tracking is a fundamental problem that should precede recognition. We see our work as a basis for applying such techniques to the video domain that intelligent agents observe.

3. Opening up Open World Tracking

Current trackers are limited to specific object classes, such as people or cars, that are labeled in training datasets (which we refer to as *known* objects). We wish to additionally evaluate trackers on *unknown* objects, which were not labeled in the training set. An open-world tracker must segment and track *all* objects (both *known* and *unknown*) in videos. Evaluating trackers in this setting is notoriously challenging. First, densely labeling every object in a video is prohibitively expensive. Virtually *no* real-world dataset labels all objects, typically limiting the labeling cost by labeling only a subset of classes (*e.g.* KITTI [24], MOTChallenge [18]) or instances (*e.g.* TAO [16]). Second, defining a generic but consistent notion of an *object* is difficult [1].

We address these two challenges simultaneously by relying on a *recall*-based evaluation, inspired by early work on object proposal evaluation [1, 22] and also adopted for zero-shot object detection [4] and open-world LiDAR segmentation [84]. Although a precise definition of an *object* is difficult to specify, people have a general notion of what an object is and can label arbitrary objects in a scene [26]. Therefore, we can obtain positive object instances as those on which multiple human annotators reach a consensus that *something is an object*. This allows us to measure how many ground truth instances a tracker can *recall*.

Defining the notion of a false positive (FP) is non-trivial as we can only expect a *subset* of objects to be labeled. If we consider unlabeled regions as non-objects (FPs), we may be penalizing the tracking system for tracking regions that could still be considered to be valid objects. See Fig. 3 for an example of objects not labeled in the TAO [16] dataset, but correctly tracked by our baseline tracker.

Open-World Tracking Accuracy (OWTA). We propose the OWTA (Open-World Tracking Accuracy) metric for this task, which is a generalization of the recently proposed

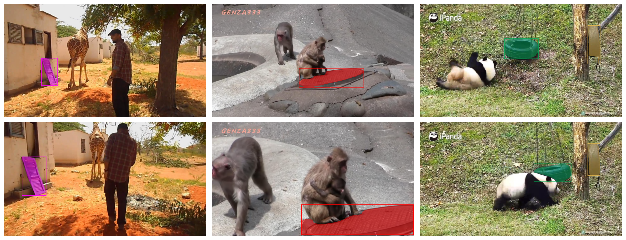

Figure 3. **Unknown unknowns.** Examples of unlabeled objects outside of the TAO [16] vocabulary which are correctly tracked by our tracker.

HOTA metric [45] for closed-world tracking. OWTA consists of two intuitive terms, the *association accuracy (AssA)* and *detection recall (DetRe)*. Both terms are evaluated with respect to localization threshold α, and the final OWTA metric is integrated over localization thresholds α:

$$\text{OWTA}_\alpha = \sqrt{\text{DetRe}_\alpha \cdot \text{AssA}_\alpha}\,, \quad \text{DetRe}_\alpha = \frac{|\text{TP}_\alpha|}{|\text{TP}_\alpha| + |\text{FN}_\alpha|}.$$

The recall term *DetRe* does not penalize false positives. This recall-based evaluation is inspired by prior work for evaluating tasks in the open-world such as zero-shot object detection or LiDAR instance segmentation [4, 84].

The association accuracy *AssA* term was recently introduced in [45]. It measures the number of frames in which the predicted track overlaps with the matched ground truth track. For each true positive detection in a predicted track p_t which is matched to a ground truth track g_t, *AssA* computes the number of TP associations (TPA, detections in p_t which overlap with g_t), FP associations (FPA, detections in p_t which do not overlap with g_t), and FN associations (FNA, ground truth annotations in g_t which do not overlap with p_t). *AssA* is evaluated as intersection-over-union over TPA, FPA and FNA sets, and averaged over TPs:

$$\text{AssA}_\alpha = \frac{1}{|\text{TP}_\alpha|} \sum_{c \in \text{TP}_\alpha} \frac{\text{TPA}_\alpha(c)}{\text{TPA}_\alpha(c) + \text{FPA}_\alpha(c) + \text{FNA}_\alpha(c)}.$$

The adoption of this *association* term is built on the insight that it is class-agnostic and does not require a densely labeled dataset. This is possible because the FPA term in *AssA* is not affected by FP tracks that are not matched to ground truth. Such a factorization is not possible with other metrics such as Track mAP [90] and IDF1 [68].

Note that at test time, we require methods to output tracks as non-overlapping masks, such that each pixel in each frame must be uniquely assigned to a track or the background. Thus, to achieve high recall a method must correctly group and track pixels over time. A trivial solution that would (theoretically) predict infinitely many tracks is not possible, as the prediction of any track implies that no other track can occupy the same pixels. This also aligns our OWT task with the current trend in tracking research to focus on tracking objects at a pixel-accurate segmentation level, and to move away from coarse bounding-box level tracking. Our task can be understood as an Open-World version of MOTS (Multi-Object Tracking and Segmentation [78]) or VIS (Video Instance Segmentation [90]).

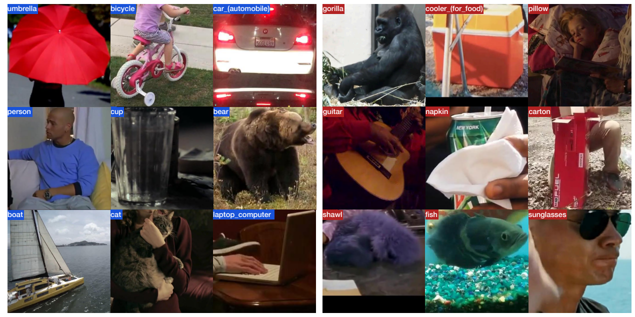

Figure 4. Examples of *known* object categories (left) and *unknown* object categories (right).

4. TAO-OW Benchmark

Defining a precise and reliable benchmark is critical for enabling progress. Therefore, we propose the Tracking Any Object in an Open World (TAO-OW) benchmark.

Dataset. Unlike most existing MOT benchmarks [64, 75, 78, 90, 91], the recently introduced TAO [16] dataset covers a wide range of classes . TAO contains almost 3,000 videos (including 593 train, 988 validation and 1,419 test), comprising 100,000 annotated frames and 800 object categories. Importantly, TAO is annotated without pre-defining object classes: annotators were asked to tag *any objects that move in the video*. This results in a long-tailed class distribution (see Fig. 2), which serves as a proxy for the wide range of objects that could appear in the real world. If we can build trackers that can track every object in this large video corpus, we can expect them to generalize to a large variety of unconstrained and *open-world* scenarios.

By default, TAO focuses on a *closed-world* setting, where all classes are defined with examples that are given during training. We re-purpose this data for the open-world setting, by holding out certain classes from training, while still evaluating on them. We also evaluate on a further 143 classes which are only present in the test set and not the validation set, which we refer to as *unknown unknowns*. This enables evaluation in open-world conditions for classes that were not used for validating model parameters.

The *known* and the *unknown*. When selecting a set of classes for *known* and *unknown* sets there are several factors to consider: (i) there should be a large enough and varied enough amount of data covering the *known* classes, such that we can train models capable of generalizing to a wider set of classes; (ii) there should be adequate number of *unknown* classes remaining to perform a thorough analysis of tracking results for these; and finally (iii), the *known* classes should contain classes commonly used in closed-

Figure 5. **TAO-OW classes.** Word cloud showing *known* (left) and *unknown* (right) classes in our TAO-OW benchmark, with word-size proportional to frequency.

world MOT, so trackers trained for the closed-world can be directly evaluated in the open-world setting. Thus, we define classes from the COCO [42] dataset as *known*, containing 80 common classes, including people, animals, vehicles, hand-held objects and appliances.

TAO validation contains 52 of COCO's 80 classes, with a total of $87,358$ distinct object tracks – we label these as the *known* set of object tracks. TAO validation contains an additional 209 classes which do not overlap with COCO, consisting of $20,522$ distinct object tracks. Of these *unknown* object classes not present in the COCO dataset, the most common are *fish*, *towel* and *pillow* with 1274, 1128 and 688 tracks, respectively. This *unknown* set includes many interesting and worthwhile-to-track classes; some of the authors' favorites include *walrus*, *ice cream*, *drum*, *frog*, *gift wrap* and *binoculars*. Fig. 4 shows visual examples of videos for both *known* and *unknown* objects. Fig. 5 presents a word cloud of all *known* and *unknown* objects in the TAO-OW validation set, where the word size is proportional to the number of annotated tracks per class. To ensure evaluation is not biased by classes similar to *known* classes, we identify 41 related classes and mark them as 'distractors', as done in closed-world tracking benchmarks [18, 24]. These are not used for evaluation. Examples include *cab* (a special case of *car*) and *water bottle* (a special case of *bottle*). We provide details in the supplementary.

Additional considerations. TAO is not densely labeled, there are many objects with no annotations. This requires special handling for *closed-set* tracking, where metrics would penalize trackers for correctly predicting unannotated objects. However, this does not affect OWT, which uses a recall-based OWTA metric (see Sec.3). Also, TAO labels objects with bounding boxes, not segmentation masks, while OWT requires methods to produce mask results. Since the ground truth boxes are non-amodal (only cover the visible part of objects), we can evaluate by converting masks to bounding boxes during evaluation.

TAO-OW dataset split. We provide a train, validation and test split for the TAO-OW dataset, which are adapted from the original TAO dataset. For training we only retain annotations from *known* classes and remove all other objects.

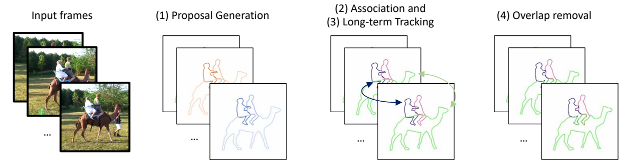

Figure 6. Open-world tracking baseline (OWTB) is inspired by tracking-by-detection pipeline: we (1) obtain object proposals, (2) compute cross-frame association scores, that are used to (3) form and manage tracks, and finally, (4) ensure that conflicts with tracks occupying same space-time volume are resolved.

The validation set contains all objects from TAO but are further labeled as either *known* or *unknown* depending on if they match a COCO class. The test set contains all objects, with classes which were present in the validation set labeled as *known* or *unknown* respectively, and remaining classes labeled as *unknown unknowns*. Since the *unknown* classes in the validation set can be used to validate design decisions, in order to test models in a truly 'hold out' scenario, we require that the test set contains classes which are not present in the validation set. These are the *unknown unknown* classes. Only by such separation, can we consider our test set as a valid proxy for the real open-world, beyond all classes in both the train and validation sets.

5. Designing Open-World Trackers

No benchmark is complete without well thought through and well-designed baselines. The most closely related methods [17, 47, 56] are not directly applicable to the TAO-OW domain: [56] require stereo video, [17] assume objects move, while [47] assumes that all objects are present in every frame.Therefore, a significant contribution of this paper is the analysis of the principles underlying these methods to distill a unified framework for open-world trackers.

To devise a strong baseline in such a challenging setting, we first study the anatomy of *tracking-by-detection* (TBD) methodology which has been the dominant MOT approach for years [18], and study how it can be adapted for the task of OWT tracking. We observe that standard TBD can be decompose into four stages (Fig. 6): (1) First, we need to obtain per-image object proposals. This is followed by (2) short-term (cross-frame) proposal similarity estimation, a direct cue for data association; (3) Based on estimated similarities we need to associate proposals and manage tracks, and finally, (4) we need to determine for each pixel a unique track-to-pixel assignment. In the following we carefully analyze each stage, using the best-performing decisions as input for later stages to reduce the exponential design space.

5.1. Proposal Generation (1)

Following *tracking-by-detection* design we first need to obtain image-level evidence the presence of potential objects. We build on intuition [17, 19, 56] that learned object proposal mechanisms, such as Region Proposal Net-

	Overall	Small size	Medium size	Large size
known/*unknown*	95.4/75.5	91.4/66.1	98.4/85.9	99.7/98.2

Table 1. **Recall/size Analysis.** Recall for varying object sizes (1k proposals/image). While models work well for known objects, and *large* unknown objects, they struggle on smaller unknown objects.

work [61, 62, 66] are explicitly trained to distinguish object-like regions from the *background* and can thus generalize beyond object classes observed in the training set, as already shown in [61, 62]. We base our analysis on the Mask R-CNN [28] and study how well it generalizes to *unknown* objects. We train our model using labels for 80 classes and first study its performance on TAO-OW's *known* and *unknown* classes separately. We evaluate this detector as a proposal generator by using a low score threshold and consider the top 1000 proposals output by the model.

Proposal recall. Tab. 1 shows the recall for both *known* and *unknown* objects of different sizes when disabling non-maximum suppression and evaluating all 1000 proposals. Object sizes are relative to the image size: Large (*ratio* $\geq$ 0.3), Medium ($0.03 \leq$ *ratio* < 0.3), Small (*ratio* < 0.03). The model performs well for large *known* and *unknown* objects, but significantly worse for small *unknown* objects. This indicates that the proposals generalize well to *unknown* objects when such objects are large and obvious but are not able to find these objects as well when they are small.

Since using all 1000 proposals as a tracking cue is not feasible, we next investigate how to distinguish *unknown* objects from the background clutter. In Fig. 7 (left) we show detection recall vs. number of object proposals for several different scoring strategies and display area under the curve. Fig. 7 show that the most confident *known* class prediction score (score, ▬) is not a very reliable ranking cue (0.89 AUC for *known* and 0.59 AUC for the *unknown*). The objectness score (obj., ▬) estimated by the RPN provides a significantly better cue (0.92 AUC for *known* and 0.67 AUC for the *unknown*). The background score (bg, ▬), estimated as score for none of the classes, *e.g.* the '$c+1$'th class for a c class detector, is reliable cue for *unknown* objects (0.67 AUC), but not for *known* objects ($0.79 AUC$). We obtain most reliable cue by combining the background and objectness scores (obj.+bg, ▬) using the arithmetic mean (0.93 AUC for *known* and 0.7 AUC for the *unknown*). We use this scoring function for the remainder of our experiments. In conclusion, 2-stage object detectors such as Mask R-CNN generalize quite well to *unknown* classes, suggesting they inherently have both an 'any object' detector built in (the RPN) and an object vs. non-object classifier.

Track recall. In addition to proposal recall, we are interested in how well *tracks* are recalled. Fig. 7 (right) shows the percentage of tracks recalled over different minimum relative track lengths. Almost every *unknown* object (97%) is recalled at least once during its track and over 80% of objects are recalled more than half the time. Around 25% of the *unknown* objects are recalled in every frame.

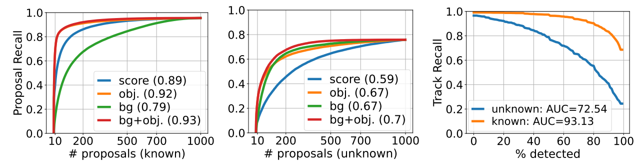

Figure 7. **Recall Analysis.** Proposal generation recall vs number of proposals for different scoring methods at IoU threshold 0.5 for *(left) known* objects and *(center) unknown* objects. *Right*: Track recall at varying % objects correctly recalled: *e.g.*, 50% detected means at least half of the track must be correctly localized.

Group	Method	Inter.	Known	Unkn.
Appearance-free	Box IoU	✓	**86.4**	70.7
		✗	73.2	72.6
	Mask IoU	✓	71.6	39.5
		✗	64.4	45.3
	GIoU	✓	**86.4**	70.5
		✗	73.0	70.9
	B. IoU + thresh	✓	81.0	**74.7**
	KF, Box IoU	✓	84.9	69.1
Regress.	Regression	✓	**88.2**	65.9
		✗	74.7	70.3
	KF, Regression	✓	**87.2**	65.5
Opt. Flow	Box IoU	✓	**87.0**	67.6
		✗	80.1	**76.8**
	Mask IoU	✓	73.3	40.8
		✗	68.4	47.9
	GIoU	✗	80.3	**75.9**
	Opt. Fl. + Regr.	✓	**88.2**	65.9
		✗	81.4	**75.3**
Re-Identification	MaskRCNN *euclidean*	✓	74.3	63.5
		✗	75.3	73.3
	MaskRCNN *cosine*	✓	73.0	64.4
		✗	75.0	74.2
	PReMVOS *euclidean* *	✗	82.3*	77.1*
	PReMVOS *cosine* *	✗	82.7*	77.5*
Mix	Flow-Box IoU + MaskRCNN *cosine*	✗	**86.3**	**81.9**

Table 2. **Association Similarity Ablation.** Top-1 accuracy on 1FPS proposal association classification for various approaches - see text. Best performing methods colored: **1st**, **2nd**, **3rd**, **4th**, **5th**. The Inter. column indicates whether 'intermediate frames' were used. *Non open-world oracle (trained on *unknown* classes)

5.2. Association Similarity (2)

Tracking requires estimating proposal *similarity* across frames to maintain object identity. Since this short-term association based on similarity is critical for accurate long-term tracking, we evaluate it in a controlled setting. We pose short-term association as a *relative classification* problem: Given a proposal corresponding to a specific query object in frame t, how well can the method identify the object among N candidate proposals in frame $t+k$? We set k to correspond to a 1 second gap, and systematically evaluate several different approaches proposed in the community [7, 8, 17, 47, 56]. We outline our analysis in Tab. 2. Note that all our methods are restricted to training on *known* classes and have not seen *unknown* classes during training.

Appearance-free. We start with simple measures that ignore image content, relying only on intersection-over-union (IoU) in the 'Appearance-free' block. This includes 'box IoU', 'mask IoU', and 'generalized IoU' (GIoU) [67]. We evaluate a strategy ('box IoU w/ assoc. thresh') of only propagating through proposals which have box IoU over a threshold (0.75) with the previous frame, skipping frames with low quality matches. We also use a Kalman Filter (KF) to forecast the box in frame $t+k$ [8, 16, 17] ('KF, Box IoU'), following parameters from [8], before computing IoU.

Regression. To incorporate appearance information in the motion estimation, we re-purpose an object detector's regressor [7] to regress the box in frame $t+k$ ('Regression').

We also consider combining this with the KF, by using the KF forecast as input to the regressor ('KF, Regression').

Flow-based. Next, we use optical flow to estimate proposal motion, following [44, 47]. We use optical flow to warp a proposal from one frame to the other and use this warped proposal with varying 'IoU' criteria. We also use this flow warp as input to the 'Regression' approach described above.

Re-Identification. We further investigate appearance-based re-identification (ReID) for similarity estimation, ensuring that the ReID is trained only on *known* classes. We repurpose the classification layer embedding (1024D) from our detector for the ReID('MaskRCNN'). We also evaluate a "non-open-world oracle" ReID which is not limited to training on *known* objects [46] ('PReMVOS').

Intermediate frames. As TAO is annotated at 1 FPS, we evaluate in two settings: direct (comparing frames 1 second apart directly) and continuous, where the similarity is propagated through intermediate frames (*i.e.*, we estimate similarity in one frame, select the most similar proposal, and repeat for all intermediate frames; denoted 'Inter. frames').

Discussion. We find that 'box IoU' performs well for both *known* (86.4) and *unknown* (70.7) objects, matching GIoU and outperforming 'mask IoU', which is sensitive to occlusions and articulated motion. Using a regressor trained on *known* objects ('Regression') improves association for *known* objects (88.2) but degrades for *unknown* objects (65.9). Using a Kalman Filter does not improve accuracy over Box IoU, and matches 'Regression' when used with the regressor. Using intermediate frames generally improves *known* accuracy, but harms *unknown* accuracy. This is because detectors have low recall for *unknown* classes, which makes dense propagation prone to drifting when propagating proposals across frames. We add 'B. IoU + thresh' entry that skips frames containing low quality matches and increases *unknown* accuracy to 74.7. Optical flow improves the results in almost all cases, and appearance-based ReID with Mask R-CNN features slightly improve results for *known* and *unknown*. The 'oracle' PReMVOS ReID [46] improves *known*, but only slightly improves *unknown* over flow methods. The most promising method uses optical flow and box IoU. We hypothesize that this method may be improved by using Mask R-CNN embeddings and evaluate a simple average of the similarity of these two approaches ('Mix'). This outperforms other approaches for *unknown* by a large margin. Not using intermediate frames works well (and is about 30$\times$ faster), therefore we ignore intermediate frames for the rest of our analysis.

5.3. Long-term Tracking (3)

After obtaining object proposals and determining a method for calculating similarity between proposals over time, we must now combine all the proposals together into long-term tracks. We compare (i) simple Hungarian matching, (ii) Hungarian matching with a keep-alive mechanism to keep tracks alive through occlusions or missing detections [17], and (iii) UnOVOST [47] that first builds tracklets using Hungarian matching, and then merges these tracklets in a second offline step. We observe that while *keep alive* strategy (39.7 OWTA for *unknown*) increases association recall over simple Hungarian matching (39.8 OWTA for *unknown*), however it does so at the loss of the association precision. *Offline tracklet merging* outperforms the two alternative strategies (40.2 OWTA for *unknown*). We provide detailed results and analysis in the supplementary.

5.4. Overlap Removal (4)

In open-world tracking scenarios we need to rely on object proposals for tracking. Thus we can hypothesize several possible explanations of the observed evidence (*i.e.*, object proposals overlap). However, OWTB tracking task requires unique assignment of pixels in the video to one of objects or background. First strategy (*'non-overlap then track'*) resolves overlaps on the proposal level, and then performs tracking. The second approach (*'track then non-overlap'*) follows [56] and performs tracking first on the set of (possibly) overlapping proposals. Each track as a whole is then scored using the mean score of each proposal in a track and track suppression is performed within the video volume. Intuitively, the second approach should perform better as it can account for temporal context, however, the association problem becomes significantly more complex. We observe that differences between the two approaches are marginal (we provide detailed results and analysis in the supplementary). The simpler *'non-overlap then track'* approach produces slightly better results. This is different to findings reported in [56] and where (i) this strategy benefits from depth information and (ii) relies on quadratic pseudo-boolean optimization [41] that is infeasible to apply at this scale.

6. Evaluation

After analyzing several design choices for open-world tracking, we settle on a tracker that uses both optical flow and re-id similarity scoring and combines these into final tracks using tracklet merging. We call our final tracker OWTB (Open-World Tracking Baseline).

Tab. 3 reports the final results of our OWTB tracker on the TAO-OW validation set. First, we compare OWTB to SORT [8] and Tracktor [7], both using same set of input proposals as OWTB (see Sec. 5.1). As can be seen, OWTB performs significantly better compared to SORT in terms of detection recall (+9.6 for *known* and +3.9 for *unknown*), association accuracy (+13.1 for *known* and +3.6 for *unknown*) and, consequentially, OWTA (+13.2 for *known* and +4.9 for *unknown*). This highlights that (i) a better tracking mechanism leads to a larger number of *unknown* objects be-

	Method	Known					Unknown					Unknown-Unknown				
		OWTA	D.Re	A.Acc	A.Re	A.Pr	OWTA	D.Re	A.Acc	A.Re	A.Pr	OWTA	D.Re	A.Acc	A.Re	A.Pr
Val	SORT [8]	46.6	67.4	33.7	39.7	56.4	33.9	43.4	30.3	34.2	57.5	–	–	–	–	–
	Tracktor [7]	57.9	80.2	42.6	43.6	94.4	22.8	54.0	10.0	10.4	96.6	–	–	–	–	–
	OWTB (Ours)	60.2	77.2	47.4	59.1	57.9	39.2	46.9	34.5	42.6	48.9	–	–	–	–	–
	OWTB (w/o N.O.) †	60.8	82.0	45.5	57.3	56.3	42.4	58.9	31.5	39.5	46.8	–	–	–	–	–
	AOA*† [21]	52.8	72.5	39.1	48.8	53.6	49.7	74.7	33.4	41.1	51.1	–	–	–	–	–
	SORT-TAO*† [16]	54.2	74.0	40.6	45.0	67.3	39.9	68.8	24.1	28.9	51.6	–	–	–	–	–
Test	SORT [8]	46.6	67.1	33.7	39.5	56.0	32.0	42.2	26.0	30.3	53.7	34.3	44.7	28.2	32.5	56.5
	Tracktor [7]	57.9	79.7	42.9	43.9	94.5	23.8	53.8	11.0	11.4	96.2	26.3	57.9	12.4	12.8	96.2
	OWTB (Ours)	60.3	76.8	47.8	59.4	58.1	38.5	45.9	33.8	42.4	49.0	41.5	48.9	36.5	45.4	52.3

Table 3. **Results of our OWTB on the TAO-OW val. and test set**. We report results in terms of our proposed OWTA metric, and additionally compare methods in terms of Detection Recall (D.Re), Association Accuracy (A.Acc), Association Recall (A.Re) and Association Precision (A.Pr). On the val set we compare our final Open-World Tracking Baseline (OWTB) to previous SOTA trackers on TAO-OW val. For the test set, *Unknown* classes are the same as those present in the val set, while *Unknown-Unknown* classes are further *unknown* classes only present in the test set. *: Non open-world (trained on *unknown* classes), †: contains overlapping results.

ing tracked, and (ii) that our method produces longer tracks in these challenging scenarios due to a higher association recall. Association precision slightly drops (for *unknown* objects) compared to SORT, which is not surprising, as we are tracking a larger number of objects. Tracktor [7] almost doesn't incorrectly merge objects at all (high A.Pr), but also doesn't merge them correctly very often either (low A.Re), resulting in an overall worse A.Acc. score than OWTB, especially for *unknown* objects. Tracktor, however, gets a boost in D.Re over OWTB because it is able to produce more proposals in each frame than the ones given to OWTB.

As an oracle, we compare to two methods (AOA [21] and SORT-TAO [16]) that are state-of-the-art on closed-world TAO. These comparisons are for reference only, as these methods are trained on *unknown* classes, and thus are not open-world trackers. They also do not produce non-overlapping results. To analyze the impact of the non-overlapping constraint, we also evaluate OWTB without forcing non-overlaps. This results in slightly better scores for both *known* and *unknown* classes, with the detection-recall improving significantly, while the association recall and precision slightly drop. OWTB performs much better than both previous SOTA closed-world trackers for *known* classes due to its strong design. However, it falls behind for *unknown* objects, since both oracle methods have been trained to both detect and track these unknown objects, resulting in higher detection recall and association accuracy. This highlights the open challenges in open-world tracking.

TAO-OW Test-set Evaluation. We evaluate and analyze OWTB on the TAO-OW validation set, where we have chosen the best design decisions via tuning. To test our tracker truly in the open-world, we further evaluate OWTB on the TAO-OW test set, which contains both *known*, *unknown* and *unknown unknown* classes (unknown classes that are not present in the validation set). Tab. 3 shows that our OWTB performs similarly between the validation and test sets for *known* and *unknown* objects. Most importantly, our method performs similarly on the set of *unknown unknown* classes (unique to the test set) as it does on classes present in validation. We hope our result will be the first of many on the TAO-OW benchmark for Open-World Tracking.

OWTB vs. previous closed-world trackers. Does a tracker designed for performing well in open-world tracking also work in the traditional closed-world scenario? To test this, we evaluate our OWTB on two previous tracking benchmarks, DAVIS unsupervised [13] and KITTI-MOTS [78]. We summarize our findings and provide detailed results discussion in the supplementary. For DAVIS we use our own proposal-generation method and rank second (65.5 $\mathcal{J}$ & $\mathcal{F}$), outperforming several recent methods [2, 25, 72, 77, 81, 92], except UnOVOST, all fine-tuned for segmenting dominantly moving regions (67.9 $\mathcal{J}$ & $\mathcal{F}$). On KITTI-MOTS we use public detections supplied by the benchmark and compare to TrackR-CNN [78] (56.5/41.9 HOTA) and recent PointTrack [89] (61.9/54.4 HOTA), that use the same detection set. Our OWTB outperforms both methods for the *car* class (64.0 HOTA) and ranks second for the *pedestrian* class (52.7 HOTA) All other methods are specifically tuned for the benchmarks and the classes, reinforcing the strong generalization capability of our method.

7. Conclusion

With this paper, we hope to light a spark in the heart of the tracking community, by opening up the potential of Open-World Tracking. We have defined Open-World Tracking as a task, motivated its importance, and presented a benchmark and precise evaluation procedure. We propose a general paradigm for tackling open-world tracking and analyze a wide range of design decisions within this paradigm. Our analysis results in a tracker that works well for *both* open-world and closed-world tracking. We are excited to announce that Open-World Tracking is open for business.

Acknowledgments. The authors would like to thank Alexander Hermans and Patrick Dendorfer for their helpful comments on our manuscript. For partial funding of this project IEZ, JL and BL would like to acknowledge the ERC Consolidator Grant DeeViSe (ERC-2017-COG-773161). YL, AO and LLT would like to acknowledge the Humboldt Foundation through the Sofja Kovalevskaja Award. AD, DR and LLT would like to acknowledge the CMU Argo AI Center for Autonomous Vehicle Research. We thank Patrick Dendorfer and Mark Weber for help with setting up the benchmark.

References

[1] Bogdan Alexe, Thomas Deselaers, and Vittorio Ferrari. Measuring the objectness of image windows. *PAMI*, 34(11):2189–2202, 2012. 3

[2] Ali Athar, Sabarinath Mahadevan, Aljosa Osep, Laura Leal-Taixé, and Bastian Leibe. Stem-seg: Spatio-temporal embeddings for instance segmentation in videos. In *ECCV*, 2020. 2, 8

[3] Mehmet Aygün, Aljoša Ošep, Mark Weber, Maxim Maximov, Cyrill Stachniss, Jens Behley, and Laura Leal-Taixé. 4d panoptic lidar segmentation. In *CVPR*, 2021. 2

[4] Ankan Bansal, Karan Sikka, Gaurav Sharma, Rama Chellappa, and Ajay Divakaran. Zero-shot object detection. In *ECCV*, 2018. 3, 4

[5] Abhijit Bendale and Terrance Boult. Towards open world recognition. In *CVPR*, 2015. 3

[6] Abhijit Bendale and Terrance E Boult. Towards open set deep networks. In *CVPR*, 2016. 3

[7] Philipp Bergmann, Tim Meinhardt, and Laura Leal-Taixé. Tracking without bells and whistles. In *ICCV*, 2019. 2, 3, 6, 7, 8

[8] Alex Bewley, Zongyuan Ge, Lionel Ott, Fabio Ramos, and Ben Upcroft. Simple online and realtime tracking. In *ICIP*, 2016. 6, 7, 8

[9] Pia Bideau and Erik Learned-Miller. A detailed rubric for motion segmentation. *arXiv preprint arXiv:1610.10033*, 2016. 2

[10] Guillem Braso and Laura Leal-Taixe. Learning a neural solver for multiple object tracking. In *CVPR*, 2020. 3

[11] William Brendel, Mohamed R. Amer, and Sinisa Todorovic. Multi object tracking as maximum weight independent set. *CVPR*, 2011. 3

[12] D. J. Butler, J. Wulff, G. B. Stanley, and M. J. Black. A naturalistic open source movie for optical flow evaluation. In *ECCV*, 2012. 1

[13] Sergi Caelles, Jordi Pont-Tuset, Federico Perazzi, Alberto Montes, Kevis-Kokitsi Maninis, and Luc Van Gool. The 2019 DAVIS challenge on VOS: unsupervised multi-object segmentation. *arXiv arXiv:1905.00737*, 2019. 2, 8

[14] Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuScenes: A multimodal dataset for autonomous driving. In *CVPR*, 2020. 2

[15] Wongun Choi. Near-online multi-target tracking with aggregated local flow descriptor. In *ICCV*, 2015. 3

[16] Achal Dave, Tarasha Khurana, Pavel Tokmakov, Cordelia Schmid, and Deva Ramanan. TAO: A large-scale benchmark for tracking any object. In *ECCV*, 2020. 2, 3, 4, 6, 8, 7

[17] Achal Dave, Pavel Tokmakov, and Deva Ramanan. Towards segmenting anything that moves. In *ICCV Workshops*, 2019. 2, 3, 5, 6, 7

[18] Patrick Dendorfer, Aljoša Ošep, Anton Milan, Konrad Schindler, Daniel Cremers, Ian Reid, and Stefan Roth Laura Leal-Taixé. Motchallenge: A benchmark for single-camera multiple target tracking. *IJCV*, 2020. 2, 3, 5, 1

[19] Akshay Dhamija, Manuel Gunther, Jonathan Ventura, and Terrance Boult. The overlooked elephant of object detection: Open set. In *WACV*, 2020. 5

[20] Akshay Raj Dhamija, Manuel Günther, and Terrance E Boult. Reducing network agnostophobia. In *NeurIPS*, 2018. 3

[21] 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. 8

[22] Ian Endres and Derek Hoiem. Category independent object proposals. In *ECCV*, 2010. 3

[23] Andreas Geiger, Martin Lauer, Christian Wojek, Christoph Stiller, and Raquel Urtasun. 3d traffic scene understanding from movable platforms. *PAMI*, 1012–1025(36):5, 2014. 3

[24] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the KITTI vision benchmark suite. In *CVPR*, 2012. 2, 3, 5, 1

[25] Shreyank N Gowda, Panagiotis Eustratiadis, Timothy Hospedales, and Laura Sevilla-Lara. Alba: Reinforcement learning for video object segmentation. *arXiv preprint arXiv:2005.13039*, 2020. 8, 2

[26] Agrim Gupta, Piotr Dollar, and Ross Girshick. LVIS: A dataset for large vocabulary instance segmentation. In *CVPR*, 2019. 3

[27] Ismail Haritaoglu, David Harwood, and Larry S. Davis. W4: Real-time surveillance of people and their activities. *PAMI*, 22:809–830, 2000. 2

[28] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask R-CNN. In *ICCV*, 2017. 3, 6

[29] Jaedong Hwang, Seoung Wug Oh, Joon-Young Lee, and Bohyung Han. Exemplar-based open-set panoptic segmentation network. In *CVPR*, 2021. 3

[30] Michal Irani and P Anandan. A unified approach to moving object detection in 2d and 3d scenes. *TPAMI*, 20(6):577–589, 1998. 2

[31] Lalit P Jain, Walter J Scheirer, and Terrance E Boult. Multi-class open set recognition using probability of inclusion. In *ECCV*, 2014. 3

[32] Ramesh Jain and H.-H Nagel. On the analysis of accumulative difference pictures from image sequences of real world scenes. *PAMI*, 1:206 – 214, 1979. 2

[33] K J Joseph, Salman Khan, Fahad Shahbaz Khan, and Vineeth N Balasubramanian. Towards open world object detection. In *CVPR*, 2021. 3

[34] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 3

[35] Chanho Kim, Fuxin Li, Arridhana Ciptadi, and James M. Rehg. Multiple hypothesis tracking revisited. In *ICCV*, 2015. 3

[36] Dahun Kim, Sanghyun Woo, Joon-Young Lee, and In So Kweon. Video panoptic segmentation. In *CVPR*, 2020. 2

[37] Laura Leal-Taixé, Cristian Canton-Ferrer, and Konrad Schindler. Learning by tracking: Siamese cnn for robust target association. *CVPR Workshops*, 2016. 3

[38] Laura Leal-Taixé, Michele Fenzi, Alina Kuznetsova, Bodo Rosenhahn, and Silvio Savarese. Learning an image-based motion context for multiple people tracking. In *CVPR*, 2014. 3

[39] Laura Leal-Taixé, Gerard Pons-Moll, and Bodo Rosenhahn. Everybody needs somebody: Modeling social and grouping behavior on a linear programming multiple people tracker. In *ICCV Workshops*, 2011. 3

[40] Bastian Leibe, Aleš Leonardis, and Bernt Schiele. Robust object detection with interleaved categorization and segmentation. *IJCV*, 77(1-3):259–289, 2008. 2

[41] Bastian Leibe, Konrad Schindler, Nico Cornelis, and Luc Van Gool. Coupled object detection and tracking from static cameras and moving vehicles. *PAMI*, 30(10):1683–1698, 2008. 3, 7

[42] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *ECCV*, 2014. 2, 5

[43] Ziwei Liu, Zhongqi Miao, Xiaohang Zhan, Jiayun Wang, Boqing Gong, and Stella X Yu. Large-scale long-tailed recognition in an open world. In *CVPR*, 2019. 3

[44] Jonathon Luiten, Tobias Fischer, and Bastian Leibe. Track to reconstruct and reconstruct to track. *RAL*, 5(2):1803–1810, 2020. 7

[45] Jonathon Luiten, Aljoša Ošep, Patrick Dendorfer, Philip Torr, Andreas Geiger, Laura Leal-Taixé, and Bastian Leibe. Hota: A higher order metric for evaluating multi-object tracking. *IJCV*, 2020. 3, 4

[46] Jonathon Luiten, Paul Voigtlaender, and Bastian Leibe. Premvos: Proposal-generation, refinement and merging for video object segmentation. In *ACCV*, 2018. 3, 7

[47] Jonathon Luiten, Idil Esen Zulfikar, and Bastian Leibe. UnOVOST: Unsupervised offline video object segmentation and tracking. In *WACV*, 2020. 2, 3, 5, 6, 7

[48] Anton Milan, Laura Leal-Taixé, Konrad Schindler, and Ian Reid. Joint tracking and segmentation of multiple targets. In *CVPR*, 2015. 2

[49] Anton Milan, Stefan Roth, and Konrad Schindler. Continuous energy minimization for multitarget tracking. *PAMI*, 36(1):58–72, 2014. 3

[50] Anton Milan, Konrad Schindler, and Stefan Roth. Multi-target tracking by discrete-continuous energy minimization. *PAMI*, 38(10):2054–2068, 2016. 3

[51] Dimity Miller, Lachlan Nicholson, Feras Dayoub, and Niko Sünderhauf. Dropout sampling for robust object detection in open-set conditions. In *ICRA*, 2018. 3

[52] Dennis Mitzel, Esther Horbert, Andreas Ess, and Bastian Leibe. Multi-person tracking with sparse detection and continuous segmentation. In *ECCV*, 2010. 3

[53] Frank Moosmann, Oliver Pink, and Christoph Stiller. Segmentation of 3d lidar data in non-flat urban environments using a local convexity criterion. In *Intel. Vehicles Symp.*, 2009. 2

[54] Frank Moosmann and Christoph Stiller. Joint self-localization and tracking of generic objects in 3d range data. In *ICRA*, 2013. 2

[55] Aljosa Osep, Paul Voigtlaender, Jonathon Luiten, Stefan Breuers, and Bastian Leibe. Towards large-scale video video object mining. In *ECCV Workshop on Interactive and Adaptive Learning in an Open World*, 2018. 3

[56] Aljoša Ošep, Wolfgang Mehner, Paul Voigtlaender, and Bastian Leibe. Track, then decide: Category-agnostic vision-based multi-object tracking. In *ICRA*, 2018. 2, 3, 5, 6, 7

[57] Aljoša Ošep, Paul Voigtlaender, Jonathon Luiten, Stefan Breuers, and Bastian Leibe. Large-scale object mining for object discovery from unlabeled video. In *ICRA*, 2019. 3

[58] Aljoša Ošep, Paul Voigtlaender, Mark Weber, Jonathon Luiten, and Bastian Leibe. 4d generic video object proposals. In *ICRA*, 2020. 3

[59] Nikos Paragios and Rachid Deriche. Geodesic active contours and level sets for the detection and tracking of moving objects. *PAMI*, 22:266–280, 2000. 2

[60] Đorđe Petrović, Radomir Mijailović, and Dalibor Pešić. Traffic accidents with autonomous vehicles: type of collisions, manoeuvres and errors of conventional vehicles' drivers. *Transportation research procedia*, 45:161–168, 2020. 1

[61] P.O. Pinheiro, R. Collobert, and P. Dollár. Learning to segment object candidates. In *NeurIPS*, 2015. 6

[62] P.H.O. Pinheiro, T.-Y. Lin, R. Collobert, and P. Dollár. Learning to refine object segments. In *ECCV*, 2016. 3, 6

[63] Hamed Pirsiavash, Deva Ramanan, and Charles C.Fowlkes. Globally-optimal greedy algorithms for tracking a variable number of objects. In *CVPR*, 2011. 3

[64] Jiyang Qi, Yan Gao, Yao Hu, Xinggang Wang, Xiaoyu Liu, Xiang Bai, Serge Belongie, Alan Yuille, Philip Torr, and Song Bai. Occluded video instance segmentation. *arXiv preprint arXiv:2102.01558*, 2021. 2, 4

[65] Donald B Reid. An algorithm for tracking multiple targets. *IEEE Trans. Automatic Control*, 24(6):843–854, 1979. 3

[66] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *NeurIPS*, 2015. 3, 6

[67] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *CVPR*, 2019. 6

[68] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *ECCV*, 2016. 4

[69] Walter J Scheirer, Anderson de Rezende Rocha, Archana Sapkota, and Terrance E Boult. Toward open set recognition. *PAMI*, 35(7):1757–1772, 2012. 3

[70] Walter J Scheirer, Lalit P Jain, and Terrance E Boult. Probability models for open set recognition. *PAMI*, 36(11):2317–2324, 2014. 3

[71] Jianbo Shi and Jitendra Malik. Motion segmentation and tracking using normalized cuts. In *ICCV*, 1998. 2

[72] Hongmei Song, Wenguan Wang, Sanyuan Zhao, Jianbing Shen, and Kin-Man Lam. Pyramid dilated deeper convlstm for video salient object detection. In *ECCV*, 2018. 8, 2

[73] Chris Stauffer and W. Eric L. Grimson. Learning patterns of activity using real-time tracking. *PAMI*, 22:747–757, 2000. 2

[74] Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. PWC-Net: CNNs for optical flow using pyramid, warping, and cost volume. In *CVPR*, 2018. 1

[75] 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. 2, 4

[76] Alex Teichman, Jesse Levinson, and Sebastian Thrun. Towards 3D object recognition via classification of arbitrary object tracks. In *ICRA*, 2011. 2

[77] Carles Ventura, Miriam Bellver, Andreu Girbau, Amaia Salvador, Ferran Marqués, and Xavier Gir'o i Nieto. RVOS: end-to-end recurrent network for video object segmentation. In *CVPR*, 2019. 8, 2

[78] Paul Voigtlaender, Michael Krause, Aljosa Osep, Jonathon Luiten, B.B.G Sekar, Andreas Geiger, and Bastian Leibe. MOTS: Multi-object tracking and segmentation. In *CVPR*, 2019. 1, 2, 4, 8

[79] Dominic Zeng Wang, Ingmar Posner, and Paul Newman. What could move? Finding cars, pedestrians and bicyclists in 3D laser data. In *ICRA*, 2012. 2

[80] Weiyao Wang, Matt Feiszli, Heng Wang, and Du Tran. Unidentified video objects: A benchmark for dense, open-world segmentation. *arXiv preprint arXiv:2104.04691*, 2021. 3

[81] Wenguan Wang, Hongmei Song, Shuyang Zhao, Jianbing Shen, Sanyuan Zhao, Steven CH Hoi, and Haibin Ling. Learning unsupervised video object segmentation through visual attention. In *CVPR*, 2019. 8, 2

[82] Mark Weber, Jun Xie, Maxwell Collins, Yukun Zhu, Paul Voigtlaender, Hartwig Adam, Bradley Green, Andreas Geiger, Bastian Leibe, Daniel Cremers, Aljoša Ošep, Laura Leal-Taixé, and Chen Liang-Chieh. STEP: segmenting and tracking every pixel. In *NeurIPS Track on Datasets and Benchmarks*, 2021. 2

[83] Longyin Wen, Dawei Du, Zhaowei Cai, Zhen Lei, Ming-Ching Chang, Honggang Qi, Jongwoo Lim, Ming-Hsuan Yang, and Siwei Lyu. Ua-detrac: A new benchmark and protocol for multi-object detection and tracking. *arXiv preprint arXiv:1511.04136*, 2015. 2

[84] Kelvin Wong, Shenlong Wang, Mengye Ren, Ming Liang, and Raquel Urtasun. Identifying unknown instances for autonomous driving. In *CoRL*, 2020. 3, 4

[85] Christopher Richard Wren, Ali Azarbayejani, Trevor Darrell, and Alex Pentland. Pfinder: Real-time tracking of the human body. *PAMI*, 19:780–785, 1997. 2

[86] Yuxin Wu, Alexander Kirillov, Francisco Massa, Wan-Yen Lo, and Ross Girshick. Detectron2. `https://github.com/facebookresearch/detectron2`, 2019. 1

[87] Christopher Xie, Yu Xiang, Zaid Harchaoui, and Dieter Fox. Object discovery in videos as foreground motion clustering. In *CVPR*, 2019. 3

[88] Ning Xu, Linjie Yang, Yuchen Fan, Jianchao Yang, Dingcheng Yue, Yuchen Liang, Brian Price, Scott Cohen, and Thomas Huang. YouTube-VOS: Sequence-to-sequence video object segmentation. In *ECCV*, 2018. 2

[89] Zhenbo Xu, Wei Zhang, Xiao Tan, Wei Yang, Huan Huang, Shilei Wen, Errui Ding, and Liusheng Huang. Segment as points for efficient online multi-object tracking and segmentation. In *ECCV*, 2020. 8, 2

[90] Linjie Yang, Yuchen Fan, and Ning Xu. Video instance segmentation. In *ICCV*, 2019. 2, 3, 4

[91] 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. 2, 4

[92] Tianfei Zhou, Jianwu Li, Shunzhou Wang, Ran Tao, and Jianbing Shen. Matnet: Motion-attentive transition network for zero-shot video object segmentation. *IEEE Transactions on Image Processing*, 29:8326–8338, 2020. 8, 2