HOOT: Heavy Occlusions in Object Tracking Benchmark

Gozde Sahin
University of Southern California
University Park, Los Angeles
gsahin@usc.edu

Laurent Itti
University of Southern California
University Park, Los Angeles
itti@usc.edu

Abstract

*In this paper, we present **HOOT**, the **H**eavy **O**cclusions in **O**bject **T**racking Benchmark, a new visual object tracking dataset aimed towards handling high occlusion scenarios for single-object tracking tasks. The dataset consists of 581 high-quality videos, which have 436K frames densely annotated with rotated bounding boxes for targets spanning 74 object classes. The dataset is geared for development, evaluation and analysis of visual tracking algorithms that are robust to occlusions. It is comprised of videos with high occlusion levels, where the median percentage of occluded frames per-video is 68%. It also provides critical attributes on occlusions, which include defining a taxonomy for occluders, providing occlusion masks for every bounding box, per-frame partial/full occlusion labels and more. HOOT has been compiled to encourage development of new methods targeting occlusion handling in visual tracking, by providing training and test splits with high occlusion levels. This makes HOOT the first densely-annotated, large dataset designed for single-object tracking under severe occlusion. We evaluate 15 state-of-the-art trackers on this new dataset to act as a baseline for future work focusing on occlusions.*

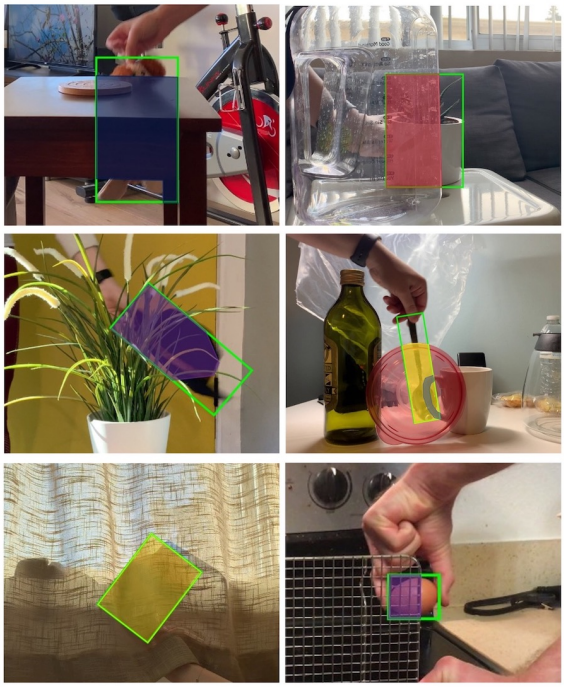


Figure 1: Sample frames from the HOOT benchmark showing different classes with a variety of occluder masks provided with the dataset, colored according to the defined occluder taxonomy (solid: dark blue, sparse: purple, semi-transparent: yellow, transparent: red). Images cropped to regions of interest to better view the target rotated bounding boxes and occluder masks.

1. Introduction

Visual object tracking is one of the most fundamental problems in computer vision and a building block to larger scale applications such as surveillance, assistive robotics, smart home devices and self-driving vehicles [1, 2, 12, 13]. These real-world applications require very robust visual tracking, especially if the algorithms are deployed for tasks such as elder care or self-driving vehicles where safety is the most important aspect. While major progress has been made in recent years with the wide adoption of deep learning in visual tracking applications [6, 20, 22, 37, 39, 41], there are still confounding factors (such as rotations, deformations, occlusions and fast motion) that are widely known to degrade tracking performance. These factors have been provided as video and sometimes per-frame attributes in pioneering single-object tracking (SOT) benchmarks like OTB [35, 36] and VOT [16, 15], as well as in more recent ones like LaSOT [9] and GOT-10k [11]. In this paper, we aim to create a new benchmark devoted to training, evaluation and analysis of visual trackers under severe occlusions.

Occlusions have always been a difficult challenge for visual trackers, as they represent a lack of visual signal coming from the target object. This makes occlusions diffi-

cult to model compared to other factors. Over the years, many algorithms have focused on occlusions in tracking applications [8, 18, 27, 30, 33]. In fact, occlusions have been a crucial part of pedestrian and vehicle tracking and datasets that are curated for these tasks [7, 34]. However, representation have lagged behind in training and evaluation benchmarks for generic single-object tracking (SOT) until recently [11, 17]. Lack of heavy occlusion scenarios in popular training and evaluation datasets has created difficulty in development of new algorithms robust to occlusion, since it is difficult to properly evaluate trackers' performance against occlusions on low occlusion benchmarks. This is starting to change with the recent HOB dataset [17], the first evaluation benchmark in SOT to focus on high occlusion scenarios. HOB consists of 20 high occlusion sequences and is annotated with a selection of per-video occlusion-related attributes. However, it is very limited in terms of dense annotations and dataset size compared to HOOT, nor does it provide a training split.

HOOT, or the **H**eavy **O**cclusions in **O**bject **T**racking Benchmark, is a new dataset for training and evaluation of SOT algorithms under heavy occlusion scenarios. It consists of 581 high-quality videos, totaling 436K frames. The contributions of HOOT to visual tracking are as follows:

- The videos are curated such that 67.7% of all frames have occlusions to varying degrees, examples of which can be seen in Fig. 1. Target classes focus on everyday items to encourage generic object tracking, since high occlusion for subjects like persons and vehicles have more specific datasets curated [7, 34].
- The videos are densely annotated with rotated bounding boxes, and have a variety of occlusion-related labels annotated *per-frame*. These include different types of partial and full occlusion labels, as well as occlusion by similar objects to the target. These attribute annotations allow for extensive analysis of tracker performance under occlusions and can be beneficial for supervision during training.
- Dense occluder masks are given for all annotated bounding boxes, as presented in Fig. 1. We define an occluder taxonomy to go with these occluder masks, to enable further analysis of tracking performance against different types of occluders (e.g., transparent vs. solid), as also identified in [17].
- The benchmark provides a resource to extensively evaluate trackers against occlusions, which is not possible with current benchmarks with limited occlusion representation and labels.

In addition to the introduction of this high occlusion dataset for visual object tracking, we also benchmark a variety of state-of-the-art trackers on HOOT. We provide analyses for different occluder types, and present a baseline for future work on occlusion invariant visual tracking algorithms. While we provide both test and training splits, we have kept training a baseline tracker on HOOT out of the scope of the paper, since effectively using HOOT occlusion labels for training would likely require a new generation of trackers that can use occlusion labels during training, which is a separate research direction.

2. Related Work

2.1. Single Object Tracking Benchmarks

This section provides a general overview of the datasets aimed for single object tracking and the information they provide about occlusions. We present an overview of widely-used and recent related works in Table 1 and further discuss how HOOT stands out amongst them below.

Starting from 2013, many datasets have been released that specifically address visual object tracking in videos. While some of them include frame-level occlusion information, others provide only video-level attributes. ALOV300++, was one of the largest datasets in the field at the time and contained a single video-level occlusion (OCC) attribute [29]. OTB-2015 [36], a pioneering evaluation dataset with 100 videos (extended from the 50 videos in OTB-2013 [35]) includes video-level occlusion (OCC) and out-of-view (OV) attributes, but do not provide per-frame information on occlusions. Other datasets like NfS [14], UAV213 [23] and TrackingNet [24] follow a similar approach to annotating occlusions and provide video-level annotations for varying attributes such as out-of-view (OV), full occlusion (FOC) and partial occlusion (POC).

In addition to the video-level occlusion attributes mentioned above, many other datasets provide per-frame annotations related to occlusions to varying extents. VOT, one of the pioneering datasets in the field, has an annual challenge [15, 16], which has consisted of 60 videos that trackers are evaluated on. While VOT provides per-frame binary occlusion tags that can indicate either partial or full occlusion, they do not provide absence labels that indicate frames where the target leaves the frame. These occlusion tags are helpful to evaluate trackers on occlusions; however, occlusion representation in VOT videos are generally low (around 10% over the years). This makes it difficult to evaluate trackers specifically addressing occlusion on the VOT challenge. Similarly to VOT, NUS-PRO [19] provides by-frame occlusion labels. However, instead of one binary occlusion label, they provide separate labels for partial and full occlusion cases.

More recently, the popular benchmark LaSOT [9] only provides absence labels per-frame for its 1.4K videos. Absence labels alone are not suitable to train or analyze trackers against per-frame partial or full occlusions, which degrade tracker performance significantly. OxUvA [31], an

General Dataset Statistics									
	OTB2015 [36]	VOT2021 [15]	UAV123 [23]	TrackingNet [24]	GOT-10k [11]	LaSOT [9]	HOB [17]	TOTB [10]	**HOOT**
Num. of Videos	100	60	123	31K	10K	1.4K	20	225	581
Num. of Frames	59K	20K	113K	14M	1.5M	3.2M	55K	87K	436K
Num. of Classes	22	30	9	21	563	70	9	15	74
Frame Rate (fps)	30	30	30	30	10	30	-	30	30-60
Avg. Duration (sec)	20	11	31	16	15	84	-	12.7	22
Train/Test Set	✗ / ✓	✗ / ✓	✗ / ✓	✓ / ✓	✓ / ✓	✓ / ✓	✗ / ✓	✗ / ✓	✓ / ✓
Occlusion Related Information									
	OTB2015 [36]	VOT2021 [15]	UAV123 [23]	TrackingNet [24]	GOT-10k [11]	LaSOT [9]	HOB [17]	TOTB [10]	**HOOT**
Video-Level Attr.	✓	✓	✓	✓	✓	✓	✓	✓	✓
Frame Absence	✗	✗	✗	✗	✓	✓	✓	✓	✓
Frame Full Occ.	✗	✓	✗	✗	✓	✗	✗	✓	✓
Frame Partial Occ.	✗	✓	✗	✗	✓	✗	✗	✗	✓
Frame Occ. Level	✗	✗	✗	✗	✓	✗	✗	✗	✓
Occluder Types	✗	✗	✗	✗	✗	✗	✓	✗	✓
Occlusion Masks	✗	✗	✗	✗	✗	✗	✗	✗	✓

Table 1: An overview of recent and widely-used visual object tracking benchmarks compared to HOOT. First part of the table focuses on general statistics, while the second part focuses on occlusion specific information provided by these benchmarks. HOOT stands out as the dataset that provides the most detailed occlusion data per-frame.

evaluation benchmark focused on long-term tracking, also provides absence labels per-frame. Another evaluation benchmark focused on transparent targets, TOTB [10], also provides per-frame absence and full-occlusion labels.

On the other hand, GOT-10k [11], a 10K video benchmark aimed towards generic, one-shot visual tracking, became one of the first visual tracking datasets that annotated occlusions in most detail. Along with absence and cut-by-frame labels per-frame, GOT-10k also provides an occlusion level for the target in each frame in the form of 9 labels ranging from fully-occluded to full-visible. While these were the most detailed occlusion labels yet, they did not provide location of the occlusion or the type of occluder. Occluder types only became of interest with the recent release of the evaluation dataset HOB [17], where videos were tagged with labels indicating the target was occluded by a similar object or a transparent object. Unfortunately, HOB does not provide per-frame partial occlusion information, nor does it provide occluder types per-frame. Moreover, it is only annotated every 15th frame, compared to the dense annotations in HOOT. With HOOT, we extend the occlusion level labels from GOT-10k by providing occluder masks for each frame, and inspired by some occluder types introduced in HOB, define a taxonomy for occluders, to help analyze tracker performance against each of them separately.

2.2. Other Benchmarks

Heavy occlusion representation has also started to gain attention in other computer vision tasks. One of most related works to HOOT in this aspect is the OVIS (or Occluded Video Instance Segmentation) Benchmark [26]. OVIS is the first benchmark in video instance segmentation to focus on heavy occlusions. They provide segmentation masks for objects of 25 different classes that consists of animals, vehicles and persons. In terms of objects classes, we believe that OVIS can be highly complementary to HOOT, which consists mainly of everyday objects. OVIS has the following occlusions distributions for the instances in the benchmark: 18.2% no occlusion, 55.5% slight occlusion, and 26.3% severe occlusion. While OVIS and HOOT are both occlusion heavy, OVIS computes occlusion levels considering intersections of instance bounding boxes, whereas HOOT annotates all types of occlusions on the target objects to represent occluded regions as accurately as possible.

Multi-object tracking is another closely related field which has been paying attention to occlusions. Unlike single-object tracking benchmarks mentioned in the previous section, the most widely-used multi-object tracking benchmark MOT, annotates all occluders and visibility ratios [7]. Another multi-object tracking benchmark, UA-DETRAC [34], also labels per-frame occlusions using bounding box intersections, and provides visibility ratios. However, both of these benchmarks focus solely on pedestrian and vehicles, whereas HOOT covers a wider number of objects from everyday life.

Lastly, addressing pedestrian detection, CityPersons [38] is one of the first datasets to focus on high occlusion representation and provides visibility regions for each bounding box. The rough occlusion masks provided by CityPersons were successfully used to predict occluded regions in works like [25]. These examples show that just like visual object tracking, many other computer vision communities are paying more and more attention to heavy occlusions.

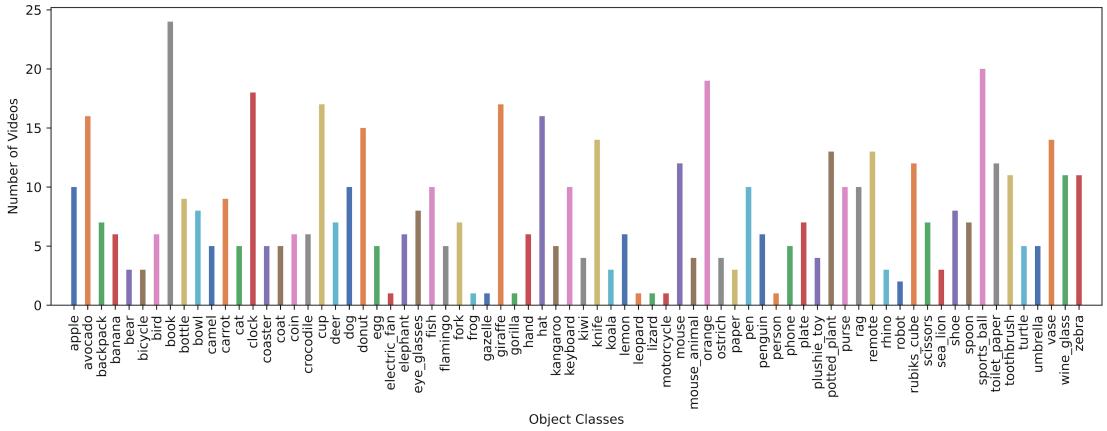


Figure 2: Target class distribution in HOOT.

3. HOOT Benchmark

In this section, we present the **H**eavy **O**cclusions in **O**bject **T**racking (or HOOT in short) Benchmark in detail. This main introduction to HOOT will include the design choices made for the benchmark, a general overview of statistics, details about data collection and annotation phases, in-depth statistics on occlusion-related attributes and evaluation protocols. The full dataset, along with evaluation results will be released at *https://www.hootbenchmark.org*.

3.1. Benchmark Design

As we discussed in Section 1, HOOT is aimed to be the first occlusion-heavy benchmark in visual object tracking that has dense annotations for occlusions and to provide a space to evaluate new algorithms against occlusions. State-of-the-art trackers still suffer huge performance drops when evaluated on high occlusion scenarios (see supp. material Section 1), and HOOT can facilitate further development of occlusion robust trackers in the field. In collection and annotation of the benchmark, we observed the following design choices:

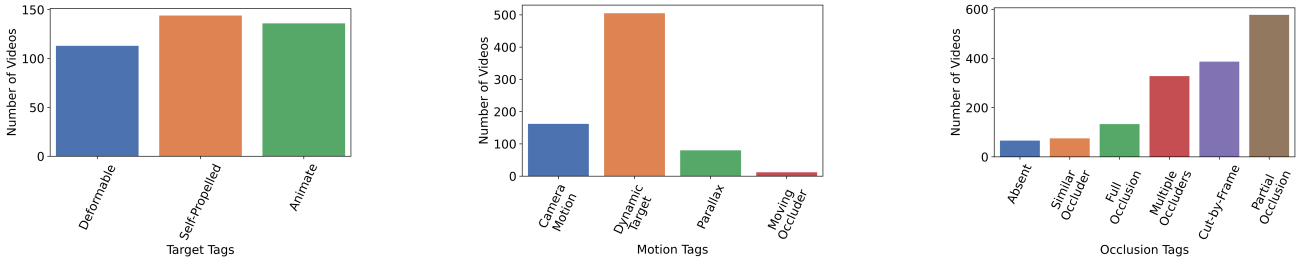
Heavily Occluded Targets: To encourage development and extensive analysis of occlusion invariant trackers, we designed the benchmark to be occlusion heavy. 67.7% of all frames in HOOT have occlusion while previous SOT benchmarks with per-frame occlusion annotations like VOT and GOT-10k have much lower occlusion representation (around 10% for VOT and 15.43% for GOT-10k [11]). The median percentage of occlusion in HOOT videos is 68%.

Dense Occlusion Attributes: Since HOOT highlights addressing occlusions in tracking, we designed the benchmark to densely annotate types of occlusions that exists in each frame. Therefore, instead of focusing on attributes like illumination variance or rotations, we curated HOOT to include 6 occlusion attributes annotated per-frame: *absent*, *full occlusion*, *cut-by-frame*, *partial occlusion*, *occluded-by-similar-object* and *occluded-by-multiple-occluder-types*. Moreover, we designed a taxonomy for occluders which are further detailed in Section 3.2.

Dense Occluder Masks: Instead of pixel-level target segmentation, HOOT provides a rotated bounding box for the target in each frame, as well as occluder masks for every bounding box. Occluder masks were created using polygons, instead of pixel-wise labeling, due to the cost of pixel-level annotations. These occluder masks (Fig. 1), coupled with the occluder taxonomy defined for the benchmark, provide valuable information on the level of visual signal coming from the target in every frame. They can also be helpful for training occlusion-aware visual trackers and performing in depth analyses for new tracking algorithms.

Class Distribution: As discussed in Section 2, outside of SOT benchmarks, much attention has been paid to occlusions for targets like persons or vehicles. Therefore, we curated HOOT such that it can be complementary to these other datasets. Thus, most of the videos in HOOT come from everyday objects that appear in common detection or tracking datasets. The variety of classes in HOOT makes it a benchmark that is more geared towards generic object tracking. The class distribution can be found in Fig. 2.

Both Training and Evaluation: The benchmark is also designed to be large enough to provide options for both training and evaluation of trackers. The videos in HOOT can be used along with other low-occlusion datasets to train more occlusion-invariant trackers. We believe using HOOT annotations effectively to train more occlusion-robust trackers can be a wider research topic and have kept it out of the scope of this paper. We expect HOOT to grow in the future and continue being an important resource to address the difficult problem of occlusions in visual object tracking.



(a) Target Attr. (b) Motion Attr. (c) Occlusion Attr.

Figure 3: Video-level distribution for target, motion and occlusion attributes in HOOT.

3.2. Benchmark Overview

This section gives further details on the benchmark, including general statistics and more importantly, extensive statistics for its occlusion-related attributes.

HOOT comprises 581 high-quality videos (1080p or higher, with frame rates of 30-60fps), with on average 22.5 seconds of duration per video. The dataset has over 3 hours of footage, and provides almost 436K frames. Further details can be found in Table 2. The distribution of the 74 object classes in the benchmark can be found in Fig. 2.

The benchmark has 13 labeled attributes which include target (3), motion (4), and occlusion (6) related attributes. Target and motion attributes are labeled only per-video, while occlusion attributes are labeled both per-frame and video.

3.2.1 Target & Motion Attributes

Target related attributes define whether the target is deformable, self-propelled (not moved by a human or device) or animate. The *deformable* attribute allows us to keep track of videos where bounding boxes may be less accurate when a deformable object changes shape while being occluded. As expected for controlled occlusion scenarios, many of the videos for everyday objects include targets moved by human subjects, which is annotated using the *self-propelled* attribute. During our evaluations, we did not observe a pattern where trackers locked on the hand moving the objects that are not self-propelled.

We also label 4 motion attributes per-video. The *camera-motion* tags videos that might exhibit varying amounts of camera motion. While most of the targets in HOOT are dynamic (represented by the attribute *dynamic-target*), there are some static target scenarios, where occlusions are either caused by parallax (camera moving to cause occlusions) or moving occluders (while camera is static). These cases are represented by the attributes *parallax* and *moving-occluder*. The distributions of target and motion-related attributes in HOOT are given in Fig. 3a and Fig. 3b.

Total # of Videos	581	**Min. # of Frames**	41	**Min. Duration**	0.98sec
Total # of Classes	74	**Max. # of Frames**	4596	**Max. Duration**	1min 38sec
Total # of Attributes	13	**Avg. # of Frames**	750	**Avg. Duration**	22.5sec
Total # of Occ. Attributes	6	**Median # of Frames**	708	**Median Duration**	21.6sec
Percentage of Occ. Frames	68%	**Total # of Frames**	435,790	**Total Duration**	3hr 38min

Table 2: General statistics for the HOOT Benchmark.

3.2.2 Occlusion Attributes

The main contribution of HOOT to the visual object tracking field is the dense occlusion-related annotations it provides. These are:

- 6 occlusion attributes, labeled by-frame,
- A taxonomy of occluder types, and
- Occlusion masks for every target bounding box, labeled by the defined taxonomy.

As mentioned in Section 3.1, the 6 occlusion attributes annotated per-frame are: *absent*, *full occlusion*, *cut-by-frame*, *partial occlusion*, *occluded-by-similar-object* and *occluded-by-multiple-occluder-types*. Video level distributions for these attributes can be seen in Fig. 3c and frame-level distributions are given in Fig. 5a, which shows target is partially occluded in 59.9% of the frames. Full occlusion and absent cases are represented, but occur considerably less often than other scenarios, since long-term tracking was not in the scope of the project. The targets are only out of the frame for 0.8% of the frames, which is 3.6K frames.

Along with heavy partial occlusions in the frame, the benchmark also has a significant representation for *cut-by-frame* (where object moves partially out of the frame). Frames labeled *multiple-occluder-types* can help analyze trackers against increasingly complex occluders. Moreover, the *occluded-by-similar-object* labels can help assess performance when the target is occluded by objects that can also be considered distractors.

In addition to these 6 occlusion attributes labeled per-frame, HOOT is also densely annotated with 4 occluder types. This occluder taxonomy has been defined as follows, with visual examples for each type given in Fig. 4:

(a) Solid occluders (for targets apple, bird, clock and remote).

(b) Sparse occluders (for targets coin, potted plant, cup and book).

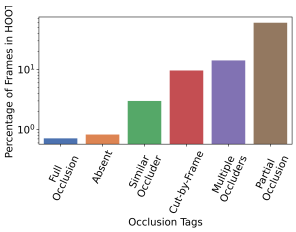
(c) Semi-transparent occluders (for targets ball, shoe, Rubik's cube and potted plant.

(d) Transparent occluders (for targets rag, orange, plate and glass.

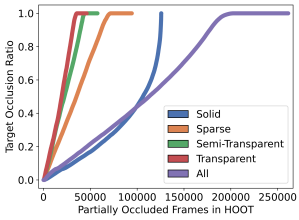
Figure 4: Sample images of different types of occluders defined in the benchmark taxonomy.

- **Solid** - where occluder completely blocks visual information from target (e.g. tree trunk, wall).
- **Sparse** - where occluder is formed of sparsely distributed solids that allow varying levels of visual information from the target (e.g. foliage, railings, blinds). This allows us to not label these complex occluders with pixel-level segmentation.
- **Semi-Transparent** - where object is covered fully by an occluder that allows some altered visual information to pass through (e.g. frosted glass, colored plastic, tight meshes).
- **Transparent** - where object is covered fully by an occluder that allows mostly unaltered visual information to pass through (e.g. glass, clear plastic).

The occluder types defined above are especially important for the dense occlusion masks provided with HOOT. As can be seen from the sample images in Fig. 1, every mask is labeled by the type of occluder corresponding to it. This ensures rough pixel-level information about the occlusion level of the target in the bounding box. For example, more visual information from the target can be obtained from areas marked by a transparent occluder, compared to a solid occluder. With these masks, we can compute a percentage level of occlusion for the target in each frame, using the intersection of occluder masks with the target bounding box. Fig. 5b shows the distribution of occlusion ratios for frames where the target is partially occluded. Overall, we find that 17.6%, 22%, 36% and 48% of the partially occluded frames in HOOT contain transparent, semi-transparent, sparse and solid occluders respectively.



(a) Occlusion attribute dist.



(b) Occlusion level dist.

Figure 5: (a) Per-frame occlusion related attributes in HOOT. (b) Target occlusion levels per occluder type across all partially occluded frames in HOOT. Full solid occlusion means target is fully occluded, which is why solid occlusion does not have partially occluded frames with occlusion ratio 1.0 while other types might.

3.3. Video Collection & Annotation

The videos in HOOT were collected by the authors and other recruited contributors (including graduate students), in a variety of environments (public and private) to increase variations in backgrounds. The recruits were given a tutorial about the general aim of the dataset and the taxonomy of occluders, as well as sample videos taken by the authors. The collected videos were cut by the authors to make sure full object visibility in the first frame and heavy occlusion.

The annotation of the collected videos were performed by a team of graduate students, using the Computer Vision Annotation Tool (CVAT) [28]. The annotation team was trained by the authors for consistent annotation and were given continuous feedback during the process. Two rounds of validation were performed before evaluation to make sure the annotations were of high standard. Further details on the procedures followed for collection and annotation can be found in the supp. material, Section 2.

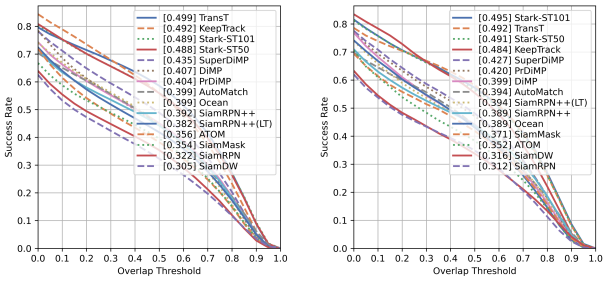
3.4. Evaluation Protocols

Inspired by LaSOT [9], we propose two protocols to evaluate trackers on the HOOT benchmark.

Protocol I This protocol includes all 581 videos in the benchmark, and aims to provide a playground to evaluate and analyze trackers against different kinds of occlusions. This protocol assumes that the evaluated trackers have not utilized any of the HOOT videos during development.

Tracker	Backbone	Venue	Protocol I (All videos)			Protocol II (Test Split)		
			Precision	Norm. Precision	Success	Precision	Norm. Precision	Success
SiamRPN [21]	AlexNet	CVPR'18	0.102	0.366	0.322	0.102	0.362	0.312
SiamMask [32]	ResNet-50	CVPR'19	0.126	0.413	0.354	0.137	0.443	0.371
ATOM [5]	ResNet-18	CVPR'19	0.121	0.415	0.356	0.121	0.420	0.352
SiamRPN++ [20]	ResNet-50	CVPR'19	0.140	0.447	0.392	0.142	0.448	0.389
SiamRPN++ (LT) [20]	ResNet-50	CVPR'19	0.135	0.417	0.382	0.148	0.440	0.394
SiamDW [40]	CIResNet-22	CVPR'19	0.092	0.348	0.305	0.106	0.361	0.316
DiMP [3]	ResNet-50	ICCV'19	0.143	0.470	0.407	0.137	0.462	0.399
PrDiMP [6]	ResNet-50	CVPR'20	0.142	0.467	0.404	0.142	0.486	0.420
Ocean [41]	ResNet-50	ECCV'20	0.142	0.475	0.399	0.134	0.467	0.389
SuperDiMP [3, 6]	ResNet-50	-	0.152	0.499	0.435	0.141	0.495	0.427
TransT [4]	ResNet-50	CVPR'21	0.230	0.597	0.499	0.235	0.589	0.492
KeepTrack [22]	ResNet-50	ICCV'21	0.177	0.578	0.492	0.169	0.570	0.484
AutoMatch [39]	ResNet-50	ICCV'21	0.158	0.480	0.399	0.160	0.478	0.394
Stark-ST50 [37]	ResNet-50	ICCV'21	0.202	0.557	0.488	0.209	0.563	0.491
Stark-ST101 [37]	ResNet-101	ICCV'21	0.212	0.564	0.489	0.216	0.571	0.495

Table 3: Overall performance results for 15 state-of-the-art trackers on HOOT protocols defined in Section 3.4. Metrics are computed as described in Section 4.1. Green, red and orange numbers represent the top 3 performers.



(a) Protocol I (b) Protocol II

Figure 6: Success curves for the state-of-the-art trackers evaluated on HOOT. The trackers are ranked according to AUC.

Protocol II For protocol II, we provide a smaller test split for evaluation of tracking algorithms on heavy occlusion scenarios. The test split contain 130 videos. Two videos were selected randomly from each object class that has at least 3 videos in the benchmark, creating a class-balanced test split. The total number of frames in the 130 test split videos is 95K. The distributions of occlusion-related attributes and occluder types can be found in the supp. material, Section 3. For this protocol, the rest of the videos in HOOT are available to use for algorithm development and training.

4. Experiments

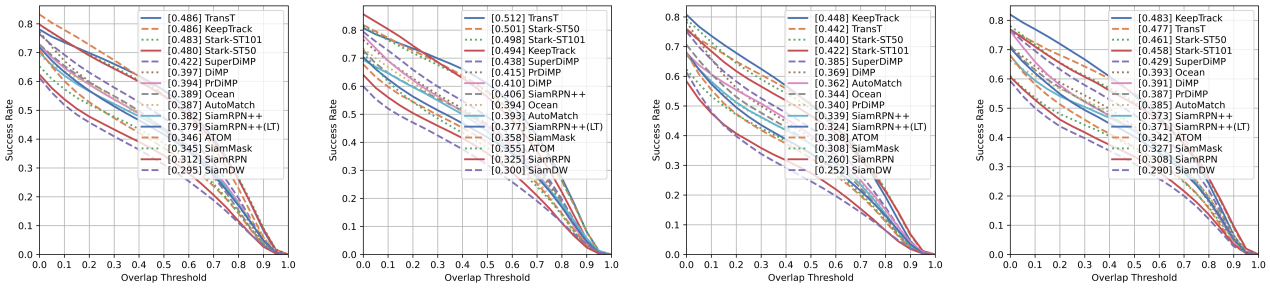
In this section, we benchmark various state-of-the-art tracking algorithms on HOOT protocols and give analyses for different occlusion attributes.

4.1. Performance Metrics

HOOT uses One-Pass Evaluation (or OPE) for evaluations, like many datasets in the field [9, 24, 36]. The metrics used to compute performance are success, precision and normalized precision. Success is computed using the Intersection over Union (IoU) between the predicted and ground truth boxes, where a success represents IoU (or overlap) being higher than some threshold. For success, tracking algorithms are ranked using Area Under the Curve (AUC) between 0 and 1. We also adopt precision and normalized precision, the latter of which was defined in [24]. Precision is calculated by looking at the percentage of frames where the distance between the predicted and ground truth boxes are under a certain threshold [35]. On the other hand, normalized precision takes into account resolution and object scale changes, by normalizing this distance with the size of the ground truth bounding box [24]. All performance results were computed by converting the rotated bounding boxes in HOOT to axis-aligned boxes, which is the output format of most trackers.

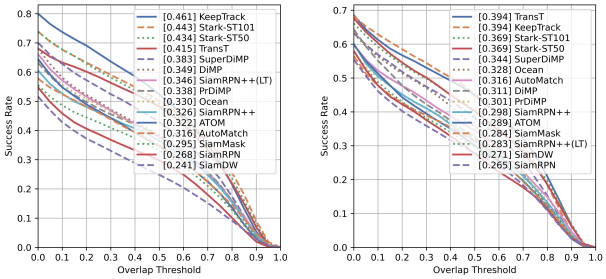
4.2. Overall Performance

In this section, we evaluate 15 recent trackers on both protocols of HOOT, and present the results on the metrics defined above in Table 3. We chose trackers that have publicly-available code and released model weights for our evaluations. The evaluated trackers represent a variety of visual tracker types. We evaluate 5 fully-convolutional Siamese trackers: SiamRPN [21], SiamMask [32], SiamRPN++ and its long term configuration SiamRPN++ (LT) [20], as well as SiamDW [40]. We also evaluate recent works that train online. These include ATOM [5], which trains an online classifier, and DiMP [3],



(a) Solid (b) Sparse (c) Semi-Transparent (d) Transparent

Figure 7: Success curves for the different occluder types annotated in HOOT, computed for Protocol I.



(a) Full Occlusion (b) Similar Object

Figure 8: Success curves for some attributes annotated in HOOT, computed for Protocol I.

which trains an online model optimizer, as well as DiMP variants PrDiMP [6] and SuperDiMP. KeepTrack [22], a recent tracker based on SuperDiMP, focuses on keeping track of distractors by utilizing a target candidate association network. Ocean [41] is an anchor-free tracker, which is also the baseline to the recent AutoMatch [39]. Finally, TransT [4] and Stark [37] use transformers for visual tracking.

Overall, trackers performed poorly on HOOT with best performers TransT, KeepTrack and Stark suffering 15-17% drops compared to the performance they achieved for LaSOT (supp. material, Section 1). Similar to [17], this shows that state-of-the-art trackers are still vulnerable to heavy occlusion scenarios. Moreover, it justifies the addition of HOOT to the field as both an extensive evaluation and a training resource (with its dense occlusion labels). Success curves for both protocols can be seen in Fig. 6, with qualitative results in the supp. material, Section 4.

4.3. Evaluation on Occlusion Attributes

We also evaluate trackers on different per-video occluder attributes, using Protocol I. Success plots for videos that contain solid, sparse, semi-transparent and transparent occluders can be viewed in Fig. 7. We notice a larger drop in performance for semi-transparent occluders, which means these will affect trackers the most in the wild even though some visual information on the object is present.

Fig. 8 shows the success plots for the HOOT videos that include cases of object being fully occluded and occluded by a similar object. We found that top transformer trackers suffered larger drops compared to KeepTrack for full occlusion, while SiamRPN++ (LT) raised in ranking since it focuses on long-term tracking. As can be seen in Fig. 8b, similar occluders affects the trackers the most. The AUC scores for all trackers suffer major drops, including KeepTrack, which focuses on handling distractors. Further results and discussions for all other attributes can be found in the supp. material Section 5 due to space limitations.

5. Conclusions

In this paper, we introduce HOOT, the Heavy Occlusions in Object Tracking Benchmark and evaluate state-of-the-art trackers on the heavy occlusion scenarios presented in the dataset. HOOT is the first dataset in single object tracking that annotates occlusions in detail for each frame. It provides occluder masks for every box in the dataset, and defines an occluder taxonomy to analyze trackers using their performance against different occluders. With two evaluation protocols, HOOT allows for training and testing of trackers against heavy occlusions, and it can facilitate the development of increasingly occlusion-robust tracking algorithms in the future.

Acknowledgements: This work was supported by C-BRIC (one of six centers in JUMP, a Semiconductor Research Corporation (SRC) program sponsored by DARPA), DARPA (HR00112190134) and the Army Research Office (W911NF2020053). The authors affirm that the views expressed herein are solely their own, and do not represent the views of the United States government or any agency thereof.

References

[1] Claudine Badue, Rânik Guidolini, Raphael Vivacqua Carneiro, Pedro Azevedo, Vinicius B Cardoso, Avelino Forechi, Luan Jesus, Rodrigo Berriel, Thiago M Paixao, Filipe Mutz, et al. Self-driving cars: A survey. *Expert Systems with Applications*, 165:113816, 2021.

[2] Frédéric Bergeron, Kevin Bouchard, Sébastien Gaboury, and Sylvain Giroux. Tracking objects within a smart home. *Expert Systems with Applications*, 113:428–442, 2018.

[3] Goutam Bhat, Martin Danelljan, Luc Van Gool, and Radu Timofte. Learning discriminative model prediction for tracking. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6182–6191, 2019.

[4] Xin Chen, Bin Yan, Jiawen Zhu, Dong Wang, Xiaoyun Yang, and Huchuan Lu. Transformer tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8126–8135, 2021.

[5] Martin Danelljan, Goutam Bhat, Fahad Shahbaz Khan, and Michael Felsberg. Atom: Accurate tracking by overlap maximization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4660–4669, 2019.

[6] Martin Danelljan, Luc Van Gool, and Radu Timofte. Probabilistic regression for visual tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7183–7192, 2020.

[7] Patrick Dendorfer, Hamid Rezatofighi, Anton Milan, Javen Shi, Daniel Cremers, Ian Reid, Stefan Roth, Konrad Schindler, and Laura Leal-Taixé. Mot20: A benchmark for multi object tracking in crowded scenes. *arXiv preprint arXiv:2003.09003*, 2020.

[8] Xingping Dong, Jianbing Shen, Dajiang Yu, Wenguan Wang, Jianhong Liu, and Hua Huang. Occlusion-aware real-time object tracking. *IEEE Transactions on Multimedia*, 19(4):763–771, 2016.

[9] Heng Fan, Liting Lin, Fan Yang, Peng Chu, Ge Deng, Sijia Yu, Hexin Bai, Yong Xu, Chunyuan Liao, and Haibin Ling. Lasot: A high-quality benchmark for large-scale single object tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5374–5383, 2019.

[10] Heng Fan, Halady Akhilesha Miththanthaya, Siranjiv Ramana Rajan, Xiaoqiong Liu, Zhilin Zou, Yuewei Lin, Haibin Ling, et al. Transparent object tracking benchmark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10734–10743, 2021.

[11] Lianghua Huang, Xin Zhao, and Kaiqi Huang. Got-10k: A large high-diversity benchmark for generic object tracking in the wild. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(5):1562–1577, 2019.

[12] Omar Javed and Mubarak Shah. Tracking and object classification for automated surveillance. In *European Conference on Computer Vision*, pages 343–357. Springer, 2002.

[13] S Hamidreza Kasaei, Miguel Oliveira, Gi Hyun Lim, Luís Seabra Lopes, and Ana Maria Tomé. Towards lifelong assistive robotics: A tight coupling between object perception and manipulation. *Neurocomputing*, 291:151–166, 2018.

[14] Hamed Kiani Galoogahi, Ashton Fagg, Chen Huang, Deva Ramanan, and Simon Lucey. Need for speed: A benchmark for higher frame rate object tracking. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1125–1134, 2017.

[15] Matej Kristan, Jiří Matas, Aleš Leonardis, Michael Felsberg, Roman Pflugfelder, Joni-Kristian Kämäräinen, Hyung Jin Chang, Martin Danelljan, Luka Cehovin, Alan Lukežič, et al. The ninth visual object tracking vot2021 challenge results. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2711–2738, 2021.

[16] Matej Kristan, Roman Pflugfelder, Ales Leonardis, Jiri Matas, Fatih Porikli, Luka Cehovin, Georg Nebehay, Gustavo Fernandez, Tomas Vojir, et al. The vot2013 challenge: overview and additional results. 2014.

[17] Thijs P Kuipers, Devanshu Arya, and Deepak K Gupta. Hard occlusions in visual object tracking. In *European Conference on Computer Vision*, pages 299–314. Springer, 2020.

[18] Beng Yong Lee, Lee Hung Liew, Wai Shiang Cheah, and Yin Chai Wang. Occlusion handling in videos object tracking: A survey. In *IOP conference series: earth and environmental science*, volume 18, page 012020. IOP Publishing, 2014.

[19] Annan Li, Min Lin, Yi Wu, Ming-Hsuan Yang, and Shuicheng Yan. Nus-pro: A new visual tracking challenge. *IEEE transactions on pattern analysis and machine intelligence*, 38(2):335–349, 2015.

[20] Bo Li, Wei Wu, Qiang Wang, Fangyi Zhang, Junliang Xing, and Junjie Yan. Siamrpn++: Evolution of siamese visual tracking with very deep networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4282–4291, 2019.

[21] Bo Li, Junjie Yan, Wei Wu, Zheng Zhu, and Xiaolin Hu. High performance visual tracking with siamese region proposal network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8971–8980, 2018.

[22] Christoph Mayer, Martin Danelljan, Danda Pani Paudel, and Luc Van Gool. Learning target candidate association to keep track of what not to track. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13444–13454, 2021.

[23] Matthias Mueller, Neil Smith, and Bernard Ghanem. A benchmark and simulator for uav tracking. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision – ECCV 2016*, pages 445–461, Cham, 2016. Springer International Publishing.

[24] Matthias Muller, Adel Bibi, Silvio Giancola, Salman Alsubaihi, and Bernard Ghanem. Trackingnet: A large-scale dataset and benchmark for object tracking in the wild. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 300–317, 2018.

[25] Yanwei Pang, Jin Xie, Muhammad Haris Khan, Rao Muhammad Anwer, Fahad Shahbaz Khan, and Ling Shao. Mask-guided attention network for occluded

pedestrian detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4967–4975, 2019.

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

[27] Gozde Sahin and Laurent Itti. Multi-task occlusion learning for real-time visual object tracking. In *2021 IEEE International Conference on Image Processing (ICIP)*, pages 524–528. IEEE, 2021.

[28] Boris Sekachev, Nikita Manovich, Maxim Zhiltsov, Andrey Zhavoronkov, Dmitry Kalinin, Ben Hoff, TOsmanov, Dmitry Kruchinin, Artyom Zankevich, DmitriySidnev, Maksim Markelov, Johannes222, Mathis Chenuet, a andre, telenachos, Aleksandr Melnikov, Jijoong Kim, Liron Ilouz, Nikita Glazov, Priya4607, Rush Tehrani, Seungwon Jeong, Vladimir Skubriev, Sebastian Yonekura, vugia truong, zliang7, lizhming, and Tritin Truong. opencv/cvat: v1.1.0, Aug. 2020.

[29] Arnold WM Smeulders, Dung M Chu, Rita Cucchiara, Simone Calderara, Afshin Dehghan, and Mubarak Shah. Visual tracking: An experimental survey. *IEEE transactions on pattern analysis and machine intelligence*, 36(7):1442–1468, 2013.

[30] Chong Sun, Dong Wang, and Huchuan Lu. Occlusion-aware fragment-based tracking with spatial-temporal consistency. *IEEE Transactions on Image Processing*, 25(8):3814–3825, 2016.

[31] Jack Valmadre, Luca Bertinetto, Joao F Henriques, Ran Tao, Andrea Vedaldi, Arnold WM Smeulders, Philip HS Torr, and Efstratios Gavves. Long-term tracking in the wild: A benchmark. In *Proceedings of the European conference on computer vision (ECCV)*, pages 670–685, 2018.

[32] Qiang Wang, Li Zhang, Luca Bertinetto, Weiming Hu, and Philip HS Torr. Fast online object tracking and segmentation: A unifying approach. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, pages 1328–1338, 2019.

[33] Xin Wang, Zhiqiang Hou, Wangsheng Yu, Lei Pu, Zefenfen Jin, and Xianxiang Qin. Robust occlusion-aware part-based visual tracking with object scale adaptation. *Pattern Recognition*, 81:456–470, 2018.

[34] 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. *Computer Vision and Image Understanding*, 193:102907, 2020.

[35] Yi Wu, Jongwoo Lim, and Ming-Hsuan Yang. Online object tracking: A benchmark. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2411–2418, 2013.

[36] Yi Wu, Jongwoo Lim, and Ming-Hsuan Yang. Object tracking benchmark. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(9):1834–1848, 2015.

[37] Bin Yan, Houwen Peng, Jianlong Fu, Dong Wang, and Huchuan Lu. Learning spatio-temporal transformer for visual tracking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10448–10457, 2021.

[38] Shanshan Zhang, Rodrigo Benenson, and Bernt Schiele. Citypersons: A diverse dataset for pedestrian detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[39] Zhipeng Zhang, Yihao Liu, Xiao Wang, Bing Li, and Weiming Hu. Learn to match: Automatic matching network design for visual tracking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13339–13348, 2021.

[40] Zhipeng Zhang and Houwen Peng. Deeper and wider siamese networks for real-time visual tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4591–4600, 2019.

[41] Zhipeng Zhang, Houwen Peng, Jianlong Fu, Bing Li, and Weiming Hu. Ocean: Object-aware anchor-free tracking. In *European Conference on Computer Vision*, pages 771–787. Springer, 2020.