HOOT: Heavy Occlusions in Object Tracking Benchmark - Supplementary Material

Tracker	Norm. Precision			Success		
	LaSOT	HOOT	Δ	LaSOT	HOOT	Δ
ATOM	0.576	0.420	-0.156	0.515	0.352	-0.163
SiamRPN++	0.569	0.448	**-0.121**	0.496	0.389	**-0.107**
DiMP	0.650	0.462	-0.188	0.569	0.399	-0.170
PrDiMP	0.688	0.486	**-0.202**	0.598	0.420	-0.178
Ocean	0.651	0.467	-0.184	0.560	0.389	-0.171
TransT	0.738	0.589	-0.149	0.649	0.492	-0.157
KeepTrack	0.772	0.570	**-0.202**	0.671	0.484	-0.187
AutoMatch	-	0.478	-	0.583	0.394	**-0.189**
Stark-ST101	0.770	0.571	-0.199	0.671	0.495	-0.176

Table 1: Comparison of the overall performance results between HOOT and LaSOT test sets. The table below only includes trackers that have also evaluated on LaSOT, and presents normalized precision and success numbers for both datasets. It shows a steep decline in performance for HOOT, which has much higher occlusion representation. Green numbers mark the trackers that suffered the least drops between LaSOT and HOOT, while red numbers mark the trackers that suffered the most.

1. Performance Comparisons with LaSOT

In this section, we compare the results of the state-of-the-art algorithms we evaluated on HOOT to their performance on the LaSOT test set. While not directly comparable, these results demonstrate how challenging HOOT can be, compared to the current popular benchmarks that do not have heavy occlusion distributions.

Table 1 shows the overall performance results for success and normalized precision metrics in LaSOT and HOOT test sets. We only present trackers that have previously evaluated on LaSOT, and use numbers given in their papers to compile the LaSOT results. Comparing the difference in overall performance, we observe that even strong state-of-the-art trackers can suffer drops when evaluated on HOOT. For normalized precision, this drop was found to be between 12-20%, while for success, it ranged from 10% to 19%. The tracker that suffered the least was SiamRPN++, which might be due to its fully offline training framework. On the other hand, trackers that perform online model updates (like KeepTrack and DiMP variants) suffered higher drops in performance compared to their performance in LaSOT. This shows how difficult HOOT is compared to the current datasets in the field of visual object tracking.

2. Data Collection & Annotation

In this section, we present further details on the data collection and annotation process, which include the specific instructions we asked collectors and annotators to follow during the process. We hope that this provides more transparency on how HOOT was created.

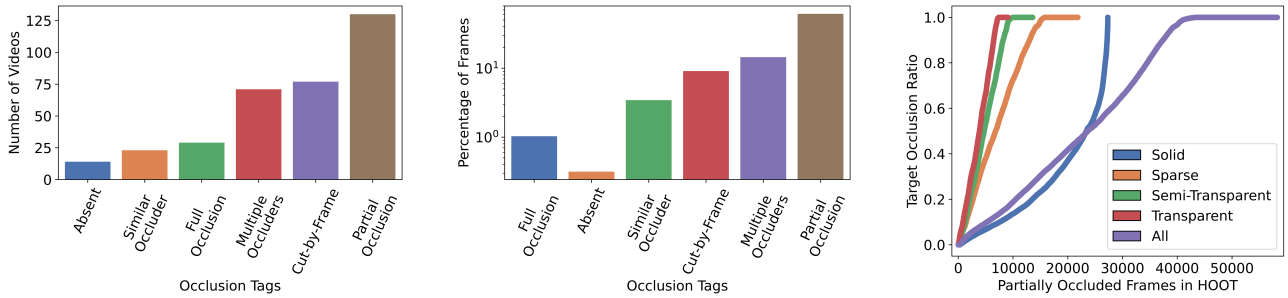
Those who were recruited for **video collection** received a tutorial explaining the aim of the project, basic definitions and sample videos taken by the authors. They were instructed to follow the instructions below:

- Set video quality to 1080p or higher (with 4k preferred) and shoot in landscape mode.
- Set frame rate to at least 30fps.
- Eliminate illumination variance by shooting in daylight or sufficiently lit indoor environments.
- Set the maximum distance of the object from the camera to a distance where the object can be clearly identified as the correct class.
- Ensure the object is fully visible in the first frame (exceptions were shooting a target through glass or water).
- Use different types of occluders and create heavy occlusion scenarios.
- Do not include any identifying information (like faces) without consent of the subjects appearing in the video.

The recruits also had access to the object class list and dataset statistics (updated frequently), so they could tailor their videos to include more of the occluders or object classes that were represented less.

During the **annotation** phase, the annotators were given extensive written instructions and an in-person tutorial for using CVAT for annotating the videos. The exact steps they followed to have a video ready for validation are as follows:

- Watch video to label per-video target and motion tags.
- Annotate the target object given in the first frame by fitting an appropriate rotated bounding box to it throughout the rest of the video. (The annotators



(a) (b) (c)

Figure 1: Distributions of occlusion related attributes in the HOOT test set. (a) Video-level attribute distributions. (b) Frame-level attribute distributions. (c) Distributions of target occlusion level per occluder type across partially occluded frames.

were allowed to utilize the interpolation tool of CVAT, but were instructed to carefully check the interpolated frames.)
- Label occlusion tags presented in the main text for every frame.
- Create occluder polygons for objects that occlude the target. (Mostly using multi-vertex polygons, depending on the complexity of the occluder, with static occluders being easier to label in more detail). Occluder masks were created out of these polygons.
- Do not create solid occluders for the limited occlusions caused by a hand moving the target (i.e. fingers blocking parts of the target).
- Estimate the full rotated bounding box position for frames where the target is partially or fully occluded by an occluder.

Measures taken to ensure annotation quality and consistency were as follows:

- Annotation tasks were distributed to annotators by object class, in order to have annotation consistency for each object type.
- Annotators were given online feedback as they annotate, until they achieved the desired annotation standards set by the authors.
- Two rounds of validation were performed by the authors after annotation to carefully check each frame for correct occlusion-related attributes and high-quality polygons and fix any errors.

3. Attribute Distributions for Protocol II

HOOT is divided into two protocols. While Protocol I is evaluation using all 581 videos in the dataset, Protocol II defines a 130-video test split and performs evaluations on those videos. This opens the remaining videos in HOOT for training and development of algorithms.

In addition to the occlusion attribute statistics for the entire dataset (given in the main text), we also present the test split occlusion attribute statistics here. Fig. 1 below shows the video and frame level attribute distributions for the annotated occlusion data in the test set, as well as the percentage of target occlusion in the partially occluded frames for the defined occluder types. We find that the distributions look very similar to the overall dataset, except for the frames when the target is absent from the video occurring in fewer percentages in the test set.

4. Qualitative Results

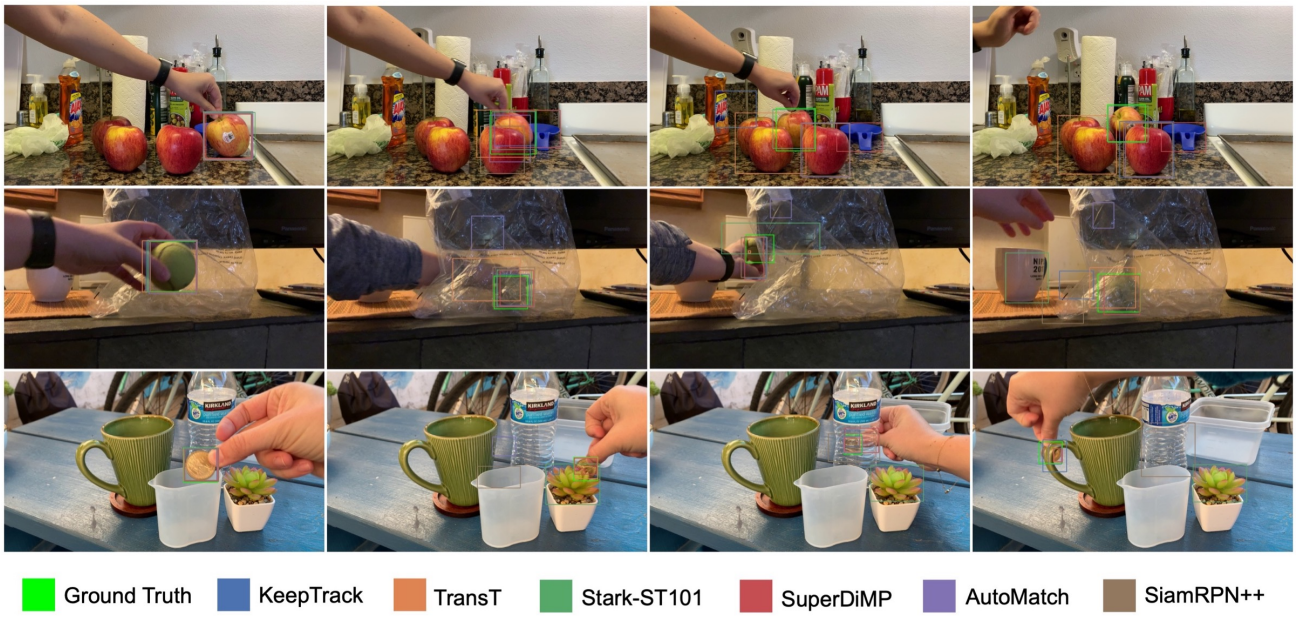
In this section, we present qualitative results for a variety of trackers we evaluated on HOOT. The trackers that we visualized in Fig. 2 cover fully-convolutional Siamese trackers like SiamRPN++, to trackers that make online model updates (such as DiMP variants), to transformer-based trackers (TransT), as well as a spatio-temporal trackers such as Stark-ST101.

We present these qualitative results in two parts. Fig. 2a contains 4 frames from 3 randomly selected videos with an average success rate of 0.418. These are videos that trackers found to be of medium difficulty. On the other hand, Fig. 2b shows results of 3 randomly selected videos that on average scored 0.128 for success. Therefore, these videos and their selected frames show examples of trackers performing significantly low on HOOT videos.

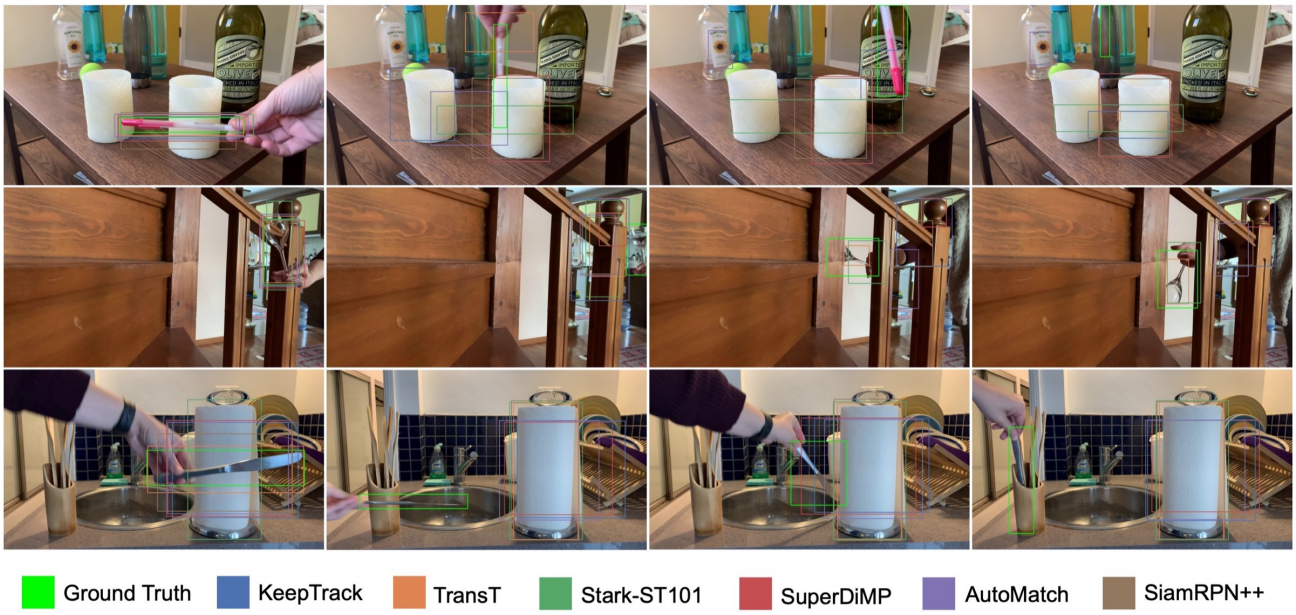
5. Additional Performance Plots

Finally, in this section, we present several performance plots for different occlusion types that we were not able to include in the main text.

Looking at the success curves for the occlusion attributes cut-by-frame, absent and multiple occluders (Fig 3), we find that tracker rankings remain mostly similar to the overall success curves. We notice that trackers suffer larger



(a) Sample frames from videos that on average scored 0.418 on the success metric.

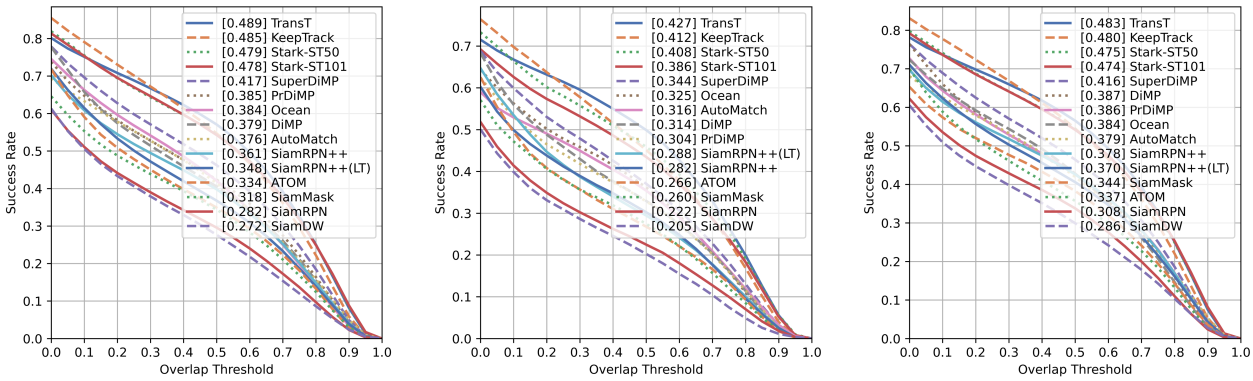


(b) Sample frames from videos that on average scored 0.128 on the success metric.

Figure 2: Qualitative tracking results of select high-performing trackers on HOOT.

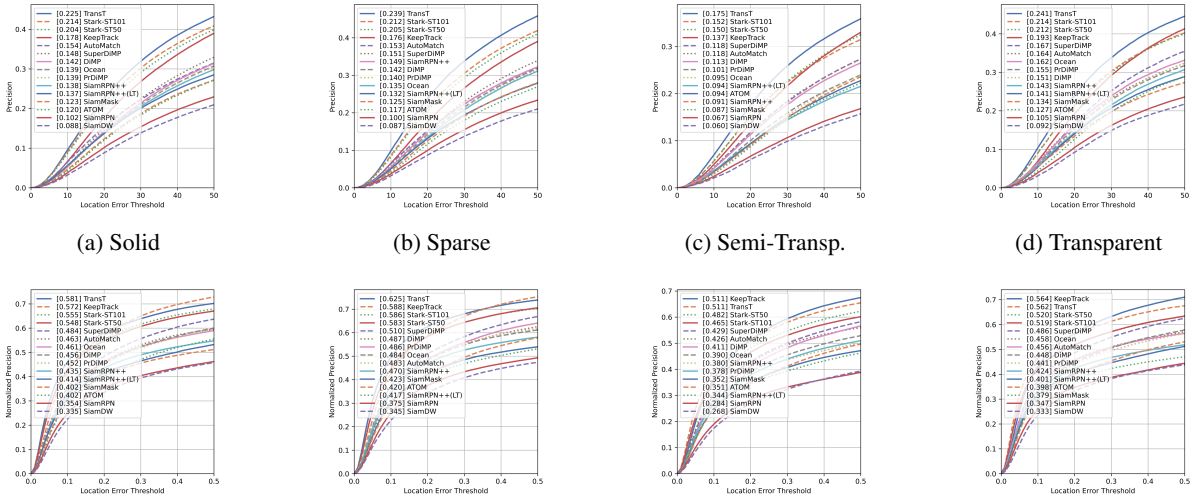
performance drops for videos that include absence, however, SiamRPN++ (LT) does perform better than the original SiamRPN++ for this attribute as expected. When computing performance for absence, absent frames were not considered since metrics cannot be computed for those. Full occlusion cases were included, since the HOOT annotations include boxes estimated by annotators even though the object is fully occluded. In addition, we find that cut-by-frame and multiple occluder scenarios were not as difficult for trackers compared to similar occluders, for which a much bigger drop in performance was observed.

We observed similar trends to success for both precision and normalized precision across all evaluations. However, minor changes in the rankings did occur. We illustrate this with the precision curves for varying location error thresholds presented for different occluder types defined in the



(a) Cut-By-Frame (b) Absent (c) Multiple Occluders

Figure 3: Success curves for the remaining occlusion attributes annotated in HOOT, computed for Protocol I.



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

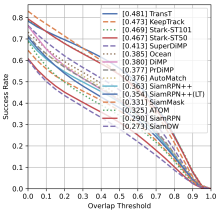
(e) Solid (f) Sparse (g) Semi-Transp. (h) Transparent

Figure 4: Precision curves for the different occluder types annotated in HOOT, computed for Protocol I. (a)-(d) Precision curves. (e)-(h) Normalized Precision curves

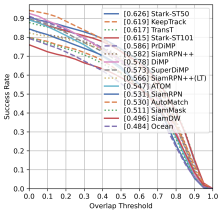
paper, Fig. 4.

In addition to the above analyses, we also computed success curves for the rest of attributes annotated in HOOT, which are the motion and target attributes. Fig. 5 shows different success curves for motion attributes, which shows that even with heavy occlusion, trackers perform better in videos that contain static objects, and worse for dynamic targets (which would be expected). Surprisingly, a major drop in performance was observed for videos with no camera motion (the highest scoring tracker's AUC for success dropped from 0.574 to 0.471). While trackers should intuitively perform better without any camera motion, this drop might imply that creating heavy occlusion conditions were probably easier for static camera scenarios, making videos with no camera motion more difficult.

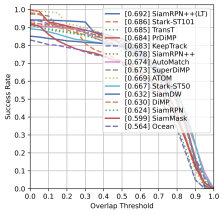
Lastly, we present analyses for target attributes annotated for HOOT in Fig. 6. We find that performance results for videos that either contain or do not contain these attributes were pretty similar, which shows the effect of occlusions do not change too much with these attributes. Notable changes were trackers performing worse for non-animate target videos. This is likely because non-animate objects can easily be put into heavier occlusion scenarios, which also explains the larger drop for videos that contain non self-propelled objects. These show that by controlling for occlusions in this manner, we were indeed able to create



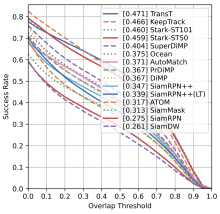
(a) Dynamic Target



(b) Static Target - Parallax

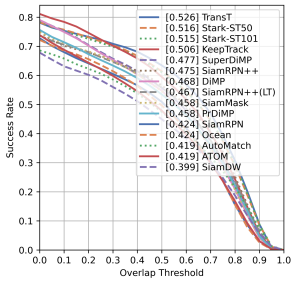


(c) Static Target - Mov. Occ.

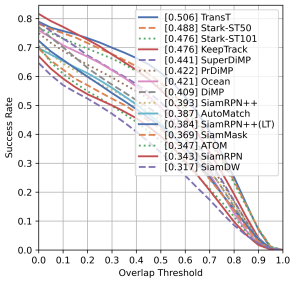


(d) No Cam. Motion

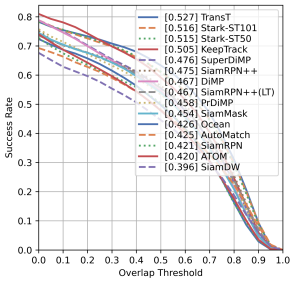
Figure 5: Success curves for videos that contain different motion tags for HOOT Protocol I. We annotate videos that contain dynamic targets and camera motion. Moreover, for static targets, we annotate tags that signify occlusion due to parallax and occlusion due to moving occluder.



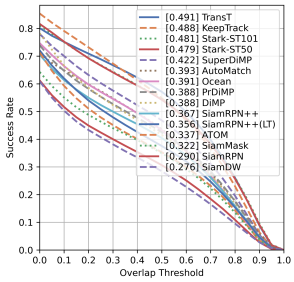
(a) Animate



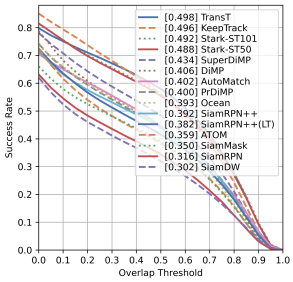
(b) Deformable



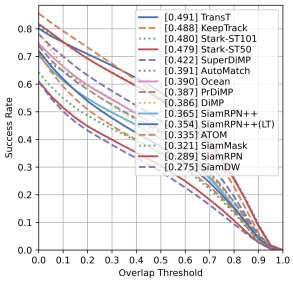
(c) Self-Propelled



(d) Non-Animate



(e) Non-Deformable



(f) Non Self-Propelled

Figure 6: Success curves for videos in HOOT annotated by different target attributes, computed for Protocol I. (a)-(c) Videos each target attribute is set to true. (d)-(f) Videos each target attribute is set to false.

difficult scenarios that affect tracker performance, and that HOOT can indeed be used to evaluate trackers on heavy occlusions.