

LaSOT: A High-quality Benchmark for Large-scale Single Object Tracking

Heng Fan^{1*} Liting Lin^{2*} Fan Yang^{1*} Peng Chu^{1*}
 Ge Deng¹ Sijia Yu¹ Hexin Bai¹ Yong Xu² Chunyuan Liao³ Haibin Ling^{1†}
¹Temple University, Philadelphia, PA USA
²South China University of Technology, Guangzhou, China
³HiScene Information Technologies, Shanghai, China

Abstract

In this paper, we present **LaSOT**, a high-quality benchmark for Large-scale Single Object Tracking. LaSOT consists of 1,400 sequences with more than 3.5M frames in total. Each frame in these sequences is carefully and manually annotated with a bounding box, making LaSOT the largest, to the best of our knowledge, densely annotated tracking benchmark. The average sequence length of LaSOT is more than 2,500 frames, and each sequence comprises various challenges deriving from the wild where target objects may disappear and re-appear again in the view. By releasing LaSOT, we expect to provide the community a large-scale dedicated benchmark with high-quality for both the training of deep trackers and the veritable evaluation of tracking algorithms. Moreover, considering the close connections of visual appearance and natural language, we enrich LaSOT by providing additional language specification, aiming at encouraging the exploration of natural linguistic feature for tracking. A thorough experimental evaluation of 35 tracking algorithms on LaSOT is presented with detailed analysis, and the results demonstrate that there is still a big room to improvements. The benchmark and evaluation results are made publicly available at <https://cis.temple.edu/laсот/>.

1. Introduction

Visual tracking, aiming to locate an arbitrary target in a video with an initial bounding box in the first frame, has been one of the most important problems in computer vision with many applications such as video surveillance, robotics, human-computer interaction and so forth [31, 46, 53]. With considerable progresses in the tracking community, numerous algorithms have been proposed. In this process, tracking benchmarks have played a vital role in objectively eval-

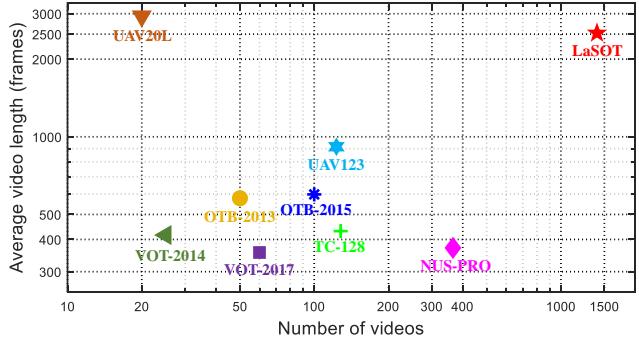


Figure 1. Summary of existing tracking benchmarks with high-quality dense (per frame) annotations, including TC-128 [34], OTB-2013 [51], OTB-2015 [52], NUS-PRO [27], UAV123 [38], UAV20L [38], VOT-2014 [25], VOT-2017 [26] and LaSOT. The proposed LaSOT is larger than all other benchmarks, and focused on long-term tracking. Best viewed in color.

uating and comparing different trackers. Nevertheless, further development and assessment of tracking algorithms are restricted by existing benchmarks with several issues:

- (1) **Small-scale.** In the era of deep learning, more and more researchers have used deep features for object tracking, and demonstrated state-of-the-art performances. However, an issue with current tracking community is that it is difficult to train a deep tracker using tracking-specific videos due to the scarcity of a *large-scale* benchmark. As shown in Fig. 1, existing datasets seldom have more than 400 sequences. As a result, researchers are restricted to leverage either the pre-trained models (*e.g.*, [45] and [18]) from image classification for deep feature extraction or the sequences from video object detection (*e.g.*, [44] and [42]) for deep feature learning, which may result in suboptimal tracking performance because of the intrinsic differences among different tasks [54].
- (2) **Lack of high-quality dense annotations.** For tracking, dense (*i.e.*, per frame) annotations with high precision are of importance for several reasons. (i) They

*Authors make equal contributions to this work.

†Corresponding author.

ensure more accurate and reliable evaluations of different trackers; (ii) they offer desired training samples for the training of tracking algorithms; and (iii) they provide rich temporal contexts among consecutive frames which are of importance for tracking task. It is worth noting that there are recently proposed benchmarks toward large-scale and long-term tracking, such as (*e.g.*, [40] and [50]), their annotations are however either semi-automatic (*e.g.*, generated by a tracking algorithm) or sparse (*e.g.*, labeled every 30 frames), limiting their usabilities.

- (3) **Short-term tracking.** A desired tracker is expected to be capable of locating the target in a relative long period, in which the target may disappear and re-enter the view. However, most existing benchmarks have been focused on *short-term* tracking where the average sequence length is less than 600 frames (*i.e.*, 20 seconds in the framerate of 30 fps, see again Fig. 1) and the target almost always appears in the video frame. The evaluations on such *short-term* benchmarks may not reflect the real performance of a tracker in real-world applications, and thus restrain the deployment in the wild.
- (4) **Category bias.** A robust tracking system should exhibit stable performance insensitive to the category the target belongs to, which signifies that the *category bias* (or *class imbalance*) should be inhibited in both training and evaluating tracking algorithms. However, existing benchmarks usually comprise a few categories (see Tab. 1) with each consisting of different number of videos, hampering assessing and developing trackers.

In the literatures, many datasets have been proposed to deal with the issues above (*e.g.*, [38] and [50] for long-term tracking, [40] for large-scale dataset, [51, 34, 24] for precise dense annotations). Nevertheless, none of them addresses all the problems, which motivates the proposal of LaSOT.

1.1. Contribution

With the goal of further advancing visual object tracking, we provide the community a novel benchmark for **Large-scale Single Object Tracking** (LaSOT) with multi-fold contributions:

- Our dataset contains 1,400 videos with an average sequence length of 2,512 frames. Each frame in each video is carefully inspected and manually labeled, and the result visually double-checked and corrected when needed. This way, we generate around 3.52 million high-quality bounding box annotations. Moreover, LaSOT contains 70 categories with each consisting of twenty sequences. To the best of our knowledge, LaSOT is the largest benchmark with high-quality manual dense annotations for object tracking to date. By releasing LaSOT, we expect it to offer a dedicated plat-

form for the development and assessment of tracking algorithms.

- Different from existing datasets, LaSOT provides both visual bounding box annotations as well as rich natural language specification, which has recently been proven to be beneficial for various computer vision tasks (*e.g.*, [21] and [30]) including visual tracking [33]. By doing so, we aim to encourage and facilitate explorations of integrating visual and lingual features for robust tracking performance.
- To assess existing trackers and provide extensive baselines for future comparisons on LaSOT, we evaluate 35 representative tracking algorithms under different protocols, and analyze their performances in details using different metrics.

2. Related Work

With considerable progresses in the tracking community, many tracking algorithms and benchmarks have been proposed in recent decades. In this section, we mainly focus on the tracking benchmarks that are relevant to our work, and refer the readers to surveys [31, 46, 53, 29] for tracking algorithms.

For a systematic review, we intentionally classify existing tracking datasets into two types: one with dense manual annotations (referred as *dense benchmark* for short) and the other one with sparse and/or (semi-)automatic annotations. In the following, we review each of these two categories.

2.1. Dense Benchmarks

Dense tracking benchmark provides dense bounding box annotations for each video sequence. To ensure high quality, the bounding boxes are usually manually labeled with careful inspection. For the visual tracking task, these highly precise annotations are desired for both training and assessing trackers. Currently, the popular dense benchmarks contain OTB [51, 52], TC-128 [34], VOT [24], NUS-PRO [27], UAV [38] and NfS [14].

OTB. OTB-2013 [51] firstly contributes a testing dataset by collecting 51 videos with manual annotated bounding box in each frame. The sequences are labeled with 11 attributes for further analysis of tracking performance. Later, OTB-2013 is extended to the larger OTB-2015 [52] by introducing extra 50 sequences.

TC-128. TC-128 [34] comprises 128 videos that are specifically designated to evaluate color-enhanced trackers. The videos in TC-128 are labeled with 11 similar attributes as in OTB [51].

VOT. VOT [24] introduces a series of tracking competitions with up to 60 sequences in each of them, aiming to evaluate the performance of a tracker in a relative short duration.

Table 1. Comparison of LaSOT with the most popular dense benchmarks in the literatures.

Benchmark	Videos	Min frames	Mean frames	Median frames	Max frames	Total frames	Total duration	frame rate	Absent labels	Object classes	Class balance	Num. of attributes	Lingual feature
OTB-2013 [51]	51	71	578	392	3,872	29K	16.4 min	30 fps	✗	10	✗	11	✗
OTB-2015 [52]	100	71	590	393	3,872	59K	32.8 min	30 fps	✗	16	✗	11	✗
TC-128 [34]	128	71	429	365	3,872	55K	30.7 min	30 fps	✗	27	✗	11	✗
VOT-2014 [25]	25	164	409	307	1,210	10K	5.7 min	30 fps	✗	11	✗	n/a	✗
VOT-2017 [26]	60	41	356	293	1,500	21K	11.9 min	30 fps	✗	24	✗	n/a	✗
NUS-PRO [27]	365	146	371	300	5,040	135K	75.2 min	30 fps	✗	8	✗	n/a	✗
UAV123 [38]	123	109	915	882	3,085	113K	62.5 min	30 fps	✗	9	✗	12	✗
UAV20L [38]	20	1,717	2,934	2,626	5,527	59K	32.6 min	30 fps	✗	5	✗	12	✗
NfS [14]	100	169	3,830	2,448	20,665	383K	26.6 min	240 fps	✗	17	✗	9	✗
LaSOT	1,400	1,000	2,506	2,053	11,397	3.52M	32.5 hours	30 fps	✓	70	✓	14	✓

Each frame in the VOT datasets is annotated with a rotated bounding box with several attributes.

NUS-PRO. NUS-PRO [27] contains 365 sequences with a focus on human and rigid object tracking. Each sequence in NUS-PRO is annotated with both target location and occlusion level for evaluation.

UAV. UAV123 and UAV20L [38] are utilized for unmanned aerial vehicle (UAV) tracking, comprising 123 short and 20 long sequences, respectively. Both UAV123 and UAV20L are labeled with 12 attributes.

NfS. NfS [14] provides 100 sequences with a high framerate of 240 fps, aiming to analyze the effects of appearance variations on tracking performance.

LaSOT belongs to the category of dense tracking dataset. Compared to others, LaSOT is the *largest* with 3.52 million frames and an average sequence length of 2,512 frames. In addition, LaSOT provides extra lingual description for each video while others do not. Tab. 1 provides a detailed comparison of LaSOT with existing dense benchmarks.

2.2. Other Benchmarks

In addition to the dense tracking benchmarks, there exist other benchmarks which may not provide high-quality annotations for each frame. Instead, these benchmarks are either annotated sparsely (*e.g.*, every 30 frames) or labeled (semi-)automatically by tracking algorithms. Despite reduction of annotation cost, the evaluations on these benchmarks may not faithfully reflect the true performances of tracking algorithms. Moreover, it may cause problems for some trackers that need to learn temporal models from annotations, since the temporal context in these benchmarks may be either *lost* because of sparse annotations or *inaccurate* due to potential unreliable annotation (by tracking) results. Representatives of this type of benchmarks include ALOV [46], TrackingNet [40] and OxUvA [50].

ALOV [46] consists of 314 sequences labeled in 14 attributes. Instead of densely annotating each frame, ALOV provides annotations every 5 frames. **TrackingNet** [40] is a subset of the video object detection benchmark YT-BB [42] by selecting 30K videos, each of which is anno-

tated by a tracker. Though the tracker used for annotation is proven to be reliable in a short period (*i.e.*, 1 second) on OTB 2015 [52], it is difficult to guarantee the same performance on a harder benchmark. Besides, the average sequence length of TrackingNet does not exceed 500 frames, which may not demonstrate the performance of a tracker in long-term scenarios. **OxUvA** [50] also comes from YT-BB [42]. Different from TrackingNet, OxUvA is focused on the long-term tracking. It consists of 366 video sequences with an average length of around 4,200 frames. However, a problem with OxUvA is that it does not provide dense annotations in consecutive frames. Each video in OxUvA is annotated every 30 frames, ignoring rich temporal context between consecutive frames when developing a tracking algorithm.

Different from the aforementioned tracking benchmarks, LaSOT provides a large set of sequences with high-quality dense bounding box annotations, which makes it more suitable for developing deep trackers as well as evaluating long-term tracking in practical application.

3. The Proposed LaSOT Benchmark

3.1. Design Principle

LaSOT aims to offer the community a dedicated dataset for training and assessing trackers. To such purpose, we follow five principles in constructing LaSOT, including *large-scale*, *high-quality dense annotations*, *long-term tracking*, *category balance* and *comprehensive labeling*.

- **Large-scale.** One of the key motivations of LaSOT is to provide a dataset for training data-hungry deep trackers, which requires a large set of annotated sequences. Accordingly, we expect such a dataset to contain at least a thousand videos with at least a million frames.
- **High-quality dense annotations.** As mentioned before, a tracking dataset is desired to have high-quality dense bounding box annotations, which are crucial for training robust trackers as well as for more faithful tracking evaluation. Following this principle, each se-

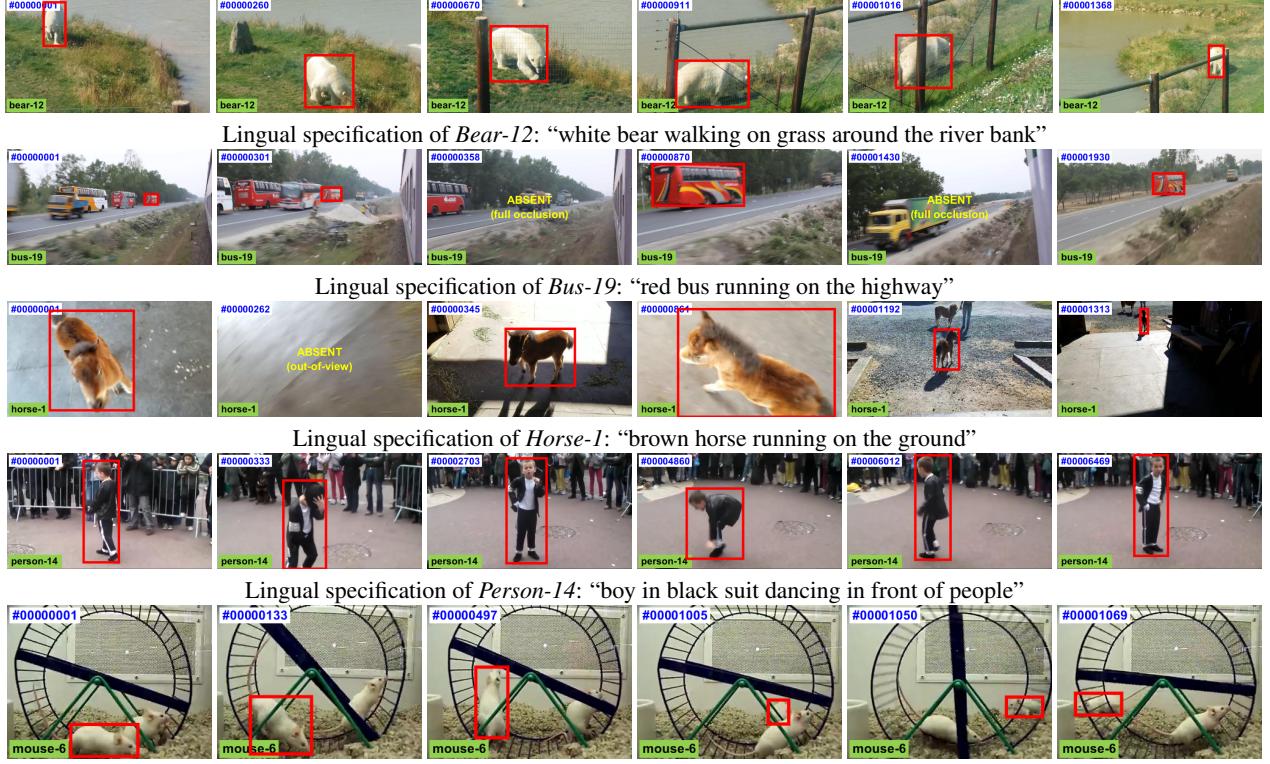


Figure 2. Example sequences and annotations of our LaSOT. We focus on long-term sequences in which the target objects may disappear, and then re-enter the view again. In addition, we provide natural language specification for each sequence. Best viewed in color.

quence in LaSOT is manually annotated with additional careful inspection and fine-tuning.

- **Long-term tracking.** In comparison with short-term tracking, long-term tracking can reflect more practical performance of a tracker in the wild. For this reason, we ensure that the shortest sequence comprises *at least* 1,000 frames, and the average sequence length of LaSOT is around 2,500 frames.
- **Category balance.** A robust tracker is expected to perform consistently regardless of the category the target object belongs to. For this purpose, in LaSOT we include a diverse set of objects from 70 classes and each class contains equal number of videos.
- **Comprehensive labeling.** As a complex task, tracking has recently seen improvements from natural language specification. To stimulate more explorations, a principle of LaSOT is to provide comprehensive labeling for videos, including both visual and lingual annotations.

3.2. Data Collection

Our benchmark covers a wide range of object categories in diverse contexts. Specifically, LaSOT consists of 70 object categories. Most of the categories are selected from the 1,000 classes from ImageNet [12], with a few exceptions,

e.g., *drone*, which are carefully chosen for popular tracking tasks. Different from existing dense benchmarks that have less than 30 categories and typically are unevenly distributed, LaSOT provides the same number of sequences for each category to alleviate potential category bias. Details of the dataset can be seen in the **supplementary material**.

After determining the 70 object categories in LaSOT, we have searched for the videos of each class from YouTube. Initially, we collect over 5,000 videos. With a joint consideration of the quality of videos for tracking and the design principles of LaSOT, we pick out 1,400 videos. However, these 1,400 sequences are not immediately available for the tracking task because of a large amount of irrelevant contents. For example, for the video of *person* category (e.g., a sporter), it often contains some introduction content of each sporter in the beginning, which is undesirable for tracking. Therefore, we carefully filter out these unrelated contents in each video and retain an usable clip for tracking. In addition, each category in LaSOT consists of 20 targets, reflecting the category balance and varieties of natural scenes.

Eventually, we have compiled a large-scale benchmark, LaSOT, for tracking by gathering 1,400 sequences with 3.52 million frames from YouTube with Creative Commons licence. The average video length of LaSOT is 2,512 frames (*i.e.*, 84 seconds based on a framerate of 30 fps). The short-

est video contains 1,000 frames (*i.e.*, 33 seconds), while the longest one consists of 11,397 frames (*i.e.*, 378 seconds).

3.3. Annotation

In order to provide consistent bounding box annotation, we define a deterministic annotation strategy. Given a video with a specific tracking target, for each frame, if the target object appears in the frame, a labeler manually draw/edit its bounding box as the tightest up-right one to fit any visible part of the target; otherwise, the labeler gives an absent label, either *out-of-view* or *full occlusion*, to the frame. Note that, such strategy can not guarantee to minimize the background area in the box, as observed in any other benchmarks. However, the strategy does provide a consistent annotation that is relatively stable for learning the dynamics.

While the above strategy works great most of the time, exceptions exist. Some objects, *e.g.* a mouse, may have long and thin and highly deformable part, *e.g.* a tail, which not only causes serious noise in object appearance and shape, but also provides little information for localizing of the target object. We carefully identify such objects and associated videos in LaSOT, and design specific rules for their annotation (*e.g.*, exclude the tails of mice when drawing their bounding boxes). An example of such cases is shown in the last row of Fig. 2.

The natural language specification of a sequence is represented by a sentence that describes the color, behavior and surroundings of the target in the whole video. For LaSOT, we provide 1,400 sentences for all the videos.

The greatest effort for constructing a high-quality dense tracking dataset is, apparently, the manual labeling, double-checking, and error correcting. For this task, we have assembled an annotation team containing several Ph.D. students working on related areas and about 10 volunteers. To guarantee high-quality annotation, each video is processed by teams: a labeling team and a validation team. A labeling team is composed of a volunteer and an expert (Ph.D. student). The volunteer manually draws/edits the target bounding box in each frame, and the expert inspects the results and adjusting them if necessary. Then, the annotation results are reviewed by the validation team, which are composed of several (typically three) experts. If an annotation result is not unanimously agreed by the members of validation team, it will be send back to the labeling team to revise.

To improve the annotation quality as much as possible, our team checks the annotation results very carefully and revises them frequently. Around 40% of the initial annotations fail in the first round of validation. Any many frames are revised more than three times. Some challenging examples of frames that are initially labeled incorrectly or inaccurately are given in Fig. 3. With all these efforts, we finally reach a benchmark with high-quality dense annotation, with some examples shown in Fig. 2.



Figure 3. Examples of fine-tuning initial annotations. Best viewed in color.

3.4. Attributes

To enable further performance analysis of trackers, we label each sequence with 14 attributes, including illumination variation (IV), full occlusion (FOC), partial occlusion (POC), deformation (DEF), motion blur (MB), fast motion (FM), scale variation (SV), camera motion (CM), rotation (ROT), background clutter (BC), low resolution (LR), viewpoint change (VC), out-of-view (OV) and aspect ration change (ARC). The definition of each attribute is shown in Tab. 2, and Fig. 4 (a) demonstrates the distribution of videos in each attribute.

From Fig. 4 (a), we observe that the most common challenge factors in LaSOT are scale changes (SV and ARC), occlusion (POC and FOC), deformation (DEF) and rotation (ROT), which are well-known challenges for tracking in real-world applications. Besides, Fig. 4 (b) demonstrates the distribution of attributes of LaSOT compared to OTB-2015 [52] and TC-128 [34] on overlapping attributes. From the figure we observe that more than 1,300 videos in LaSOT are involved with scale variations. Compared with OTB-2015 and TC-128 with less than 70 videos with scale changes, LaSOT is more challenging for scale changes. In addition, on the out-of-view attribute, LaSOT comprises 477 sequences, much larger than existing benchmarks.

3.5. Evaluation Protocols

Though there is no restriction to use LaSOT, we suggest two evaluation protocols for evaluating tracking algorithms, and conduct evaluations accordingly.

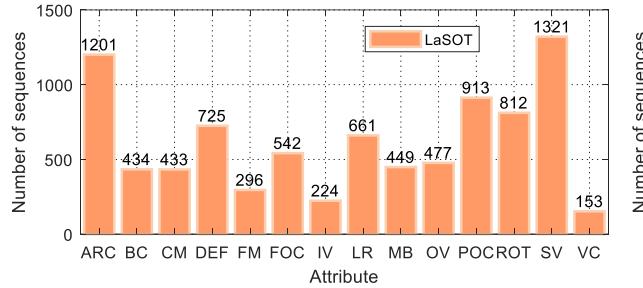
Protocol I. In protocol I, we use all 1,400 sequences to evaluate tracking performance. Researchers are allowed to employ any sequences except for those in LaSOT to develop tracking algorithms. Protocol I aims to provide large-scale evaluation of trackers.

Protocol II. In protocol II, we split LaSOT into *training* and *testing* subsets. According to the 80/20 principle (*i.e.*, the *Pareto* principle), we select 16 out of 20 videos in each category for training, and the rest is for testing¹. In specific, the *training* subset contains 1,120 videos with 2.83M frames, and the *testing* subset consists of 280 sequences with 69K

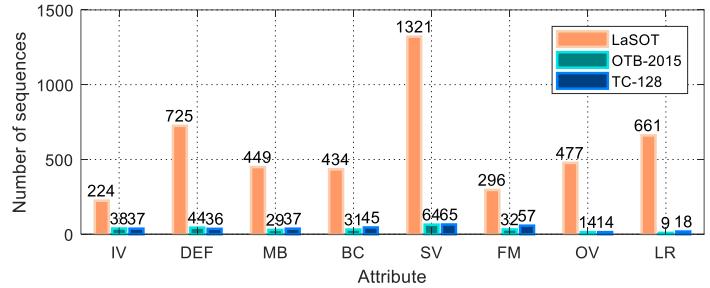
¹Due to limited space, the details of training/testing split are illustrated in the **supplementary material**.

Table 2. Descriptions of 14 different attributes in LaSOT.

Attribute	Definition	Attribute	Definition
CM	Abrupt motion of the camera	VC	Viewpoint affects target appearance significantly
ROT	The target rotates in the image	SV	The ratio of bounding box is outside the range [0.5, 2]
DEF	The target is deformable during tracking	BC	The background has the similar appearance as the target
FOC	The target is fully occluded in the sequence	MB	The target region is blurred due to target or camera motion
IV	The illumination in the target region changes	ARC	The ration of bounding box aspect ratio is outside the range [0.5, 2]
OV	The target completely leaves the video frame	LR	The target box is smaller than 1000 pixels in at least one frame
POC	The target is partially occluded in the sequence	FM	The motion of the target is larger than the size of its bounding box



(a) Distribution of sequences in each attribute on LaSOT



(b) Distribution comparison in common attributes on different benchmarks

Figure 4. Distribution of sequences in each attribute on LaSOT and comparison with other dense benchmarks. Best viewed in color.

frames. The evaluation of trackers is performed on the *testing* subset. Protocol II aims to provide a large set of videos for training and assessing trackers in the mean time.

4. Evaluation

4.1. Evaluation Metric

Following popular protocols (*e.g.* OTB-2015 [52]), we perform an One-Pass Evaluation (OPE) and measure the **precision**, **normalized precision** and **success** of different tracking algorithms under two protocols.

The precision is computed by comparing the distance between the tracking result and the groundtruth bounding box in pixels. Different tracking algorithms are ranked with this metric on a threshold (*e.g.*, 20 pixels). Since the precision metric is sensitive to target size and image resolution, we adopt the strategy as in [40] to normalize the precision. With the normalized precision metric, we rank tracking algorithms using the Area Under the Curve (AUC) between 0 to 0.5. Please refer to [40] for more about the normalized precision metric. The success is calculated as the Intersection over Union (IoU) between the tracking result and the groundtruth bounding box. The tracking algorithms are ranked using the AUC between 0 to 1.

4.2. Evaluated Trackers

We evaluate 35 algorithms on LaSOT to provide extensive baselines, comprising deep trackers (*e.g.*, MDNet [41], TRACA [5], CFNet [49], SiamFC [4], StructSiam [58], DSiam [16], SINT [48] and VITAL [47]), correlation filter trackers with hand-crafted features (*e.g.*, ECO_HC [7],

Table 3. Summary of evaluated trackers. Representation: Sparse - Sparse Representation, Color - Color Names or Histograms, Pixel - Pixel Intensity, HOG - Histogram of Oriented Gradients, H or B - Haar or Binary, Deep - Deep Feature. Search: PF - Particle Filter, RS - Random Sampling, DS - Dense Sampling.

	IVT [43]	JICV08	Representation						Search		
			PCA	Sparse	Color	Pixel	HoG	H or B	Deep	PF	RS
Struck [17]	MIL [1]	CVPR09	✓							H	
L1APG [2]	IVT [43]	ICCV11								H	
ASLA [22]	Struck [17]	CVPR12		✓						✓	
CSK [19]	L1APG [2]	CVPR12		✓						✓	
CT [57]	ASLA [22]	ECCV12									
TLD [23]	CSK [19]	ECCV12									
CN [11]	TLD [23]	PAMI12									
DSST [8]	CN [11]	CVPR14									
MEEM [55]	DSST [8]	BMVC14									
STC [56]	MEEM [55]	ECCV14									
SAMF [32]	STC [56]	ECCV14									
LCT [37]	SAMF [32]	ECCVW14									
SRDCF [10]	LCT [37]	CVPR15									
HCFT [36]	SRDCF [10]	ICCV15									
KCF [20]	HCFT [36]	ICCV15									
Staple [3]	KCF [20]	PAMI15									
SINT [48]	Staple [3]	CVPR16									
SCF4 [6]	SINT [48]	CVPR16									
MDNet [41]	SCF4 [6]	CVPR16									
SiamFC [4]	MDNet [41]	ECCV16									
Staple.CA [39]	SiamFC [4]	ECCVW16									
ECO_HC [7]	Staple.CA [39]	CVPR17									
ECO [7]	ECO_HC [7]	CVPR17									
CFNet [49]	ECO [7]	CVPR17									
CSRDCF [35]	CFNet [49]	CVPR17									
PTAV [13]	CSRDCF [35]	CVPR17									
DSiam [16]	PTAV [13]	ICCV17									
BACF [15]	DSiam [16]	ICCV17									
fDSST [9]	BACF [15]	ICCV17									
VITAL [47]	fDSST [9]	PAMI17									
TRACA [5]	VITAL [47]	CVPR18									
STRCF [28]	TRACA [5]	CVPR18									
StructSiam [58]	STRCF [28]	CVPR18									
	StructSiam [58]	ECCV18									

DSST [8], CN [11], CSK [19], KCF [20], fDSST [9],

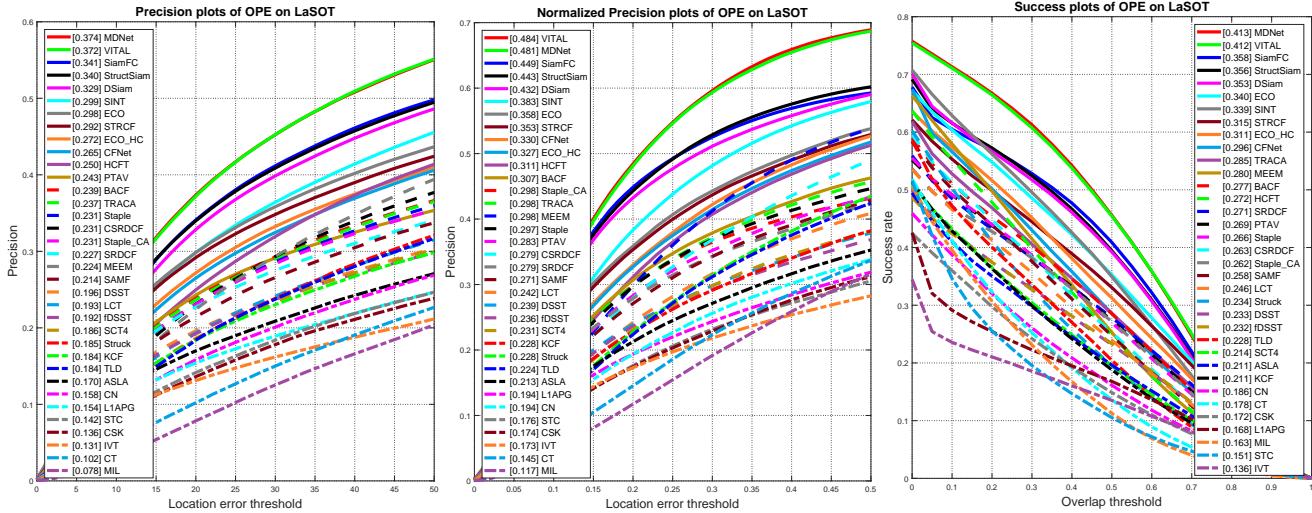


Figure 5. Evaluation results of trackers on LaSOT under protocol I using precision, normalized precision and success. Best viewed in color.

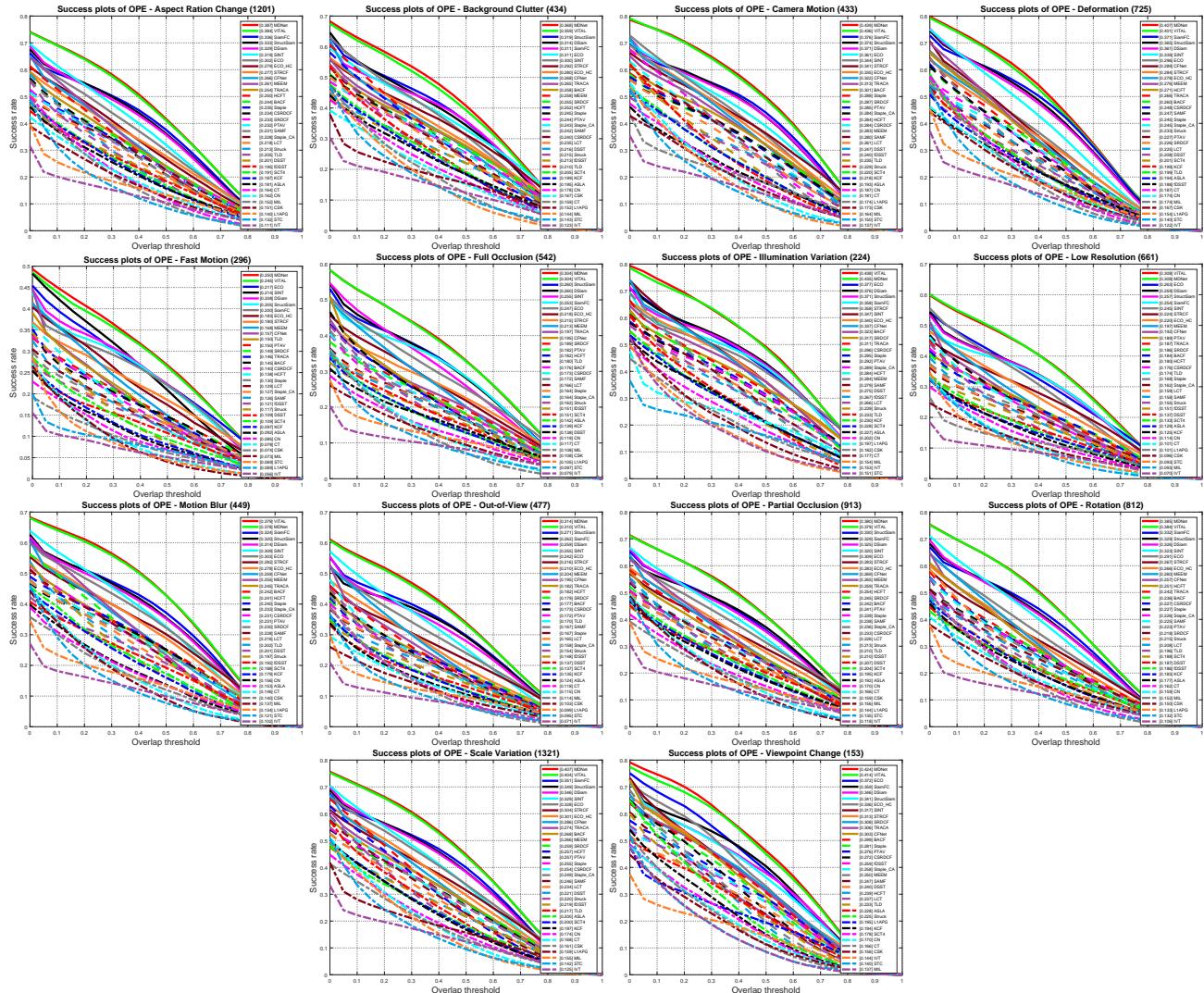


Figure 6. Performances of trackers over 14 challenges on LaSOT under protocol I using success. Best viewed in color and zoomed-in.

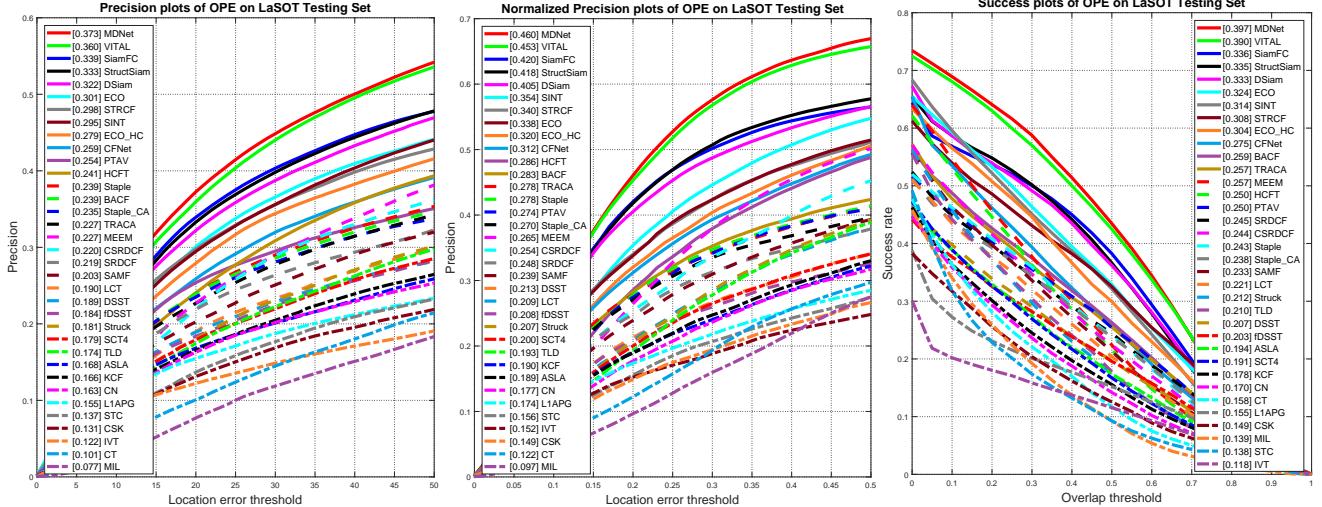


Figure 7. Evaluation on LaSOT testing set under protocol II using precision, normalized precision and success. Best viewed in color.

SAMF [32], SCT4 [6], STC [56] and Staple [3]) or deep features (*e.g.*, HCFT [36] and ECO [7]) and regularization techniques (*e.g.*, BACF [15], SRDCF [10], CSRDCF [35], Staple_CA [39] and STRCF [28]), ensemble trackers (*e.g.*, PTAV [13], LCT [37], MEEM [55] and TLD [23]), sparse trackers (*e.g.*, L1APG [2] and ASLA [22]), other representatives (*e.g.*, CT [57], IVT [43], MIL [1] and Struck [17]). Tab. 3 summarizes these trackers with their representation schemes and search strategies in a chronological order.

4.3. Evaluation Results with Protocol I

Overall performance. Protocol I aims at providing large-scale evaluations of 35 trackers on all 1,400 videos in LaSOT. Each tracker is used as is for evaluation, without any modification. We report the evaluation results in OPE using precision, normalized precision and success, as shown in Fig. 5. MDNet achieves the best precision score of 0.374 and success score of 0.413, and VITAL obtains the best normalized precision score of 0.484. Both MDNet and VITAL are trained in an online fashion, resulting in expensive computation and slow running speeds. SiamFC tracker, which learns off-line a matching function from a large set of videos using deep network, achieves competitive results with 0.341 precision score, 0.449 normalized precision score and 0.358 success score, respectively. Without time-consuming online model adaption, SiamFC runs efficiently in real-time. Motivated by SiamFC, other Siamese trackers including Struct-Siam, DSiam show good performance as well. The best correlation filter tracker is ECO with 0.298 precision score, 0.358 normalized precision score and 0.34 success score.

Compared to the typical tracking performances on existing dense benchmarks (*e.g.*, OTB-2015 [52]), the performances on LaSOT are severely degraded because of a large amount of non-rigid target objects and challenging factors in-

volved in LaSOT. An interesting observation from Fig. 5 is that all the top seven trackers leverage deep feature, demonstrating its advantages in handling appearance changes.

Attribute-based performance. Fig. 6 shows the performances of trackers over 14 challenging attributes². Overall, the performance on attributes BC, POC, LR, ARC, FOC, FM, OV and MB are poorer than that on other six attributes. The sequences with *fast motion* and *out-of-view* are difficult since existing trackers usually perform localization from a small local region. The challenges *full or partial occlusion*, *motion blur*, *aspect ration change* and *low resolution* heavily change target appearance, leading to less effective representation. The videos with *background clutter* are prone to cause drift.

4.4. Evaluation Results with Protocol II

Under protocol II, we split LaSOT into *training* and *testing* sets. Researchers are allowed to leverage the sequences in *training* set to develop their trackers and assess their performances on *testing* set. In order to provide baselines and comparisons on the *testing* set, we evaluate the 35 tracking algorithms. Each tracker is used as is for evaluation without any modification or re-training. The evaluation results are shown in Fig. 7 using precision, normalized precision and success. We observe consistent results as in protocol I. MDNet and VITAL show top performances with precision scores of 0.373 and 0.36, normalized precision scores of 0.46 and 0.453 and success scores of 0.397 and 0.39. Next, SiamFC achieves the third-ranked performance with a 0.339 precision score, a 0.42 normalized precision score and a 0.336 success score, respectively. Despite a slightly slower performance than MDNet and VITAL, SiamFC runs

²Due to limited space, please refer to **supplementary material** for detailed attribute-based evaluations of precision and normalized precision.

much faster and achieves real-time running speed, showing good balance between accuracy and efficiency. Likewise other Siamese trackers such as StructSiam and DSiam show competitive performances. For attribute-based evaluation of trackers on LaSOT *testing* set, we refer the readers to **supplementary material** because of limited space.

In addition to evaluating each tracking algorithm as it is, we conduct experiments by re-training two representative deep trackers, MDNet [41] and SiamFC [4], on the *training* set of LaSOT and assessing them. The evaluation results show similar performances for these trackers as without re-training. A potential reason is that our re-training may not follow the same configurations used by the original authors. Besides, since LaSOT are in general more challenging than previous datasets (*e.g.*, all sequences are *long-term*), dedicated configuration may be needed for training these trackers. We leave this part as a future work since it is beyond the scope of this benchmark.

5. Conclusion

In this paper we present LaSOT with high-quality dense bounding box annotations for visual object tracking. To the best of our knowledge, LaSOT is the *largest* tracking benchmark with high quality annotations to date. By releasing LaSOT, we expect to provide the tracking community a dedicated platform for training deep trackers and assessing long-term tracking performance. Besides, LaSOT provides lingual annotations for each sequence, aiming to encourage the exploration on integrating visual and lingual features for robust tracking. By releasing LaSOT, we hope to narrow the gap between the increasing number of deep trackers and the lack of large dedicated datasets for training, and meanwhile provide more veritable evaluations for different trackers in the wild. Extensive evaluations on LaSOT under two protocols imply a large room to improvement for visual tracking.

Acknowledgement. We sincerely thank Bingyao Huang, Xinyi Li, Qin Zhou, Lin Chen, Jinxiu Liang, Jingwen Wang, and anonymous volunteers for their help in constructing the benchmark.

References

- [1] B. Babenko, M.-H. Yang, and S. Belongie. Visual tracking with online multiple instance learning. In *CVPR*, 2009. **6, 8**
- [2] C. Bao, Y. Wu, H. Ling, and H. Ji. Real time robust l1 tracker using accelerated proximal gradient approach. In *CVPR*, 2012. **6, 8**
- [3] L. Bertinetto, J. Valmadre, S. Golodetz, O. Miksik, and P. H. Torr. Staple: Complementary learners for real-time tracking. In *CVPR*, 2016. **6, 8**
- [4] L. Bertinetto, J. Valmadre, J. F. Henriques, A. Vedaldi, and P. H. Torr. Fully-convolutional siamese networks for object tracking. In *ECCV*, 2016. **6, 9**
- [5] J. Choi, H. J. Chang, T. Fischer, S. Yun, K. Lee, J. Jeong, Y. Demiris, and J. Y. Choi. Context-aware deep feature compression for high-speed visual tracking. In *CVPR*, 2018. **6**
- [6] J. Choi, H. Jin Chang, J. Jeong, Y. Demiris, and J. Young Choi. Visual tracking using attention-modulated disintegration and integration. In *CVPR*, 2016. **6, 8**
- [7] M. Danelljan, G. Bhat, F. S. Khan, and M. Felsberg. Eco: Efficient convolution operators for tracking. In *CVPR*, 2017. **6, 8**
- [8] M. Danelljan, G. Häger, F. Khan, and M. Felsberg. Accurate scale estimation for robust visual tracking. In *BMVC*, 2014. **6**
- [9] M. Danelljan, G. Häger, F. S. Khan, and M. Felsberg. Discriminative scale space tracking. *TPAMI*, 39(8):1561–1575, 2017. **6, 8**
- [10] M. Danelljan, G. Hager, F. Shahbaz Khan, and M. Felsberg. Learning spatially regularized correlation filters for visual tracking. In *ICCV*, 2015. **6, 8**
- [11] M. Danelljan, F. Shahbaz Khan, M. Felsberg, and J. Van de Weijer. Adaptive color attributes for real-time visual tracking. In *CVPR*, 2014. **6**
- [12] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. **4, 12**
- [13] H. Fan and H. Ling. Parallel tracking and verifying: A framework for real-time and high accuracy visual tracking. In *ICCV*, 2017. **6, 8**
- [14] H. K. Galoogahi, A. Fagg, C. Huang, D. Ramanan, and S. Lucey. Need for speed: A benchmark for higher frame rate object tracking. In *ICCV*, 2017. **2, 3, 11**
- [15] H. K. Galoogahi, A. Fagg, and S. Lucey. Learning background-aware correlation filters for visual tracking. In *ICCV*, 2017. **6, 8**
- [16] Q. Guo, W. Feng, C. Zhou, R. Huang, L. Wan, and S. Wang. Learning dynamic siamese network for visual object tracking. In *ICCV*, 2017. **6**
- [17] S. Hare, A. Saffari, and P. H. S. Torr. Struck: Structured output tracking with kernels. In *ICCV*, 2011. **6, 8**
- [18] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016. **1**
- [19] J. F. Henriques, R. Caseiro, P. Martins, and J. Batista. Exploiting the circulant structure of tracking-by-detection with kernels. In *ECCV*, 2012. **6**
- [20] J. F. Henriques, R. Caseiro, P. Martins, and J. Batista. High-speed tracking with kernelized correlation filters. *TPAMI*, 37(3):583–596, 2015. **6**
- [21] R. Hu, H. Xu, M. Rohrbach, J. Feng, K. Saenko, and T. Darrell. Natural language object retrieval. In *CVPR*, 2016. **2**
- [22] X. Jia, H. Lu, and M.-H. Yang. Visual tracking via adaptive structural local sparse appearance model. In *CVPR*, 2012. **6, 8**
- [23] Z. Kalal, K. Mikolajczyk, J. Matas, et al. Tracking-learning-detection. *TPAMI*, 34(7):1409, 2012. **6, 8**
- [24] M. Kristan, J. Matas, A. Leonardis, T. Vojíř, R. Pflugfelder, G. Fernandez, G. Nebehay, F. Porikli, and L. Čehovin. A novel performance evaluation methodology for single-target trackers. *TPAMI*, 38(11):2137–2155, 2016. **2**

- [25] M. Kristan et al. The visual object tracking vot2014 challenge results. In *ECCVW*, 2014. 1, 3
- [26] M. Kristan et al. The visual object tracking vot2017 challenge results. In *ICCVW*, 2017. 1, 3, 11
- [27] A. Li, M. Lin, Y. Wu, M.-H. Yang, and S. Yan. Nus-pro: A new visual tracking challenge. *TPAMI*, 38(2):335–349, 2016. 1, 2, 3, 11
- [28] F. Li, C. Tian, W. Zuo, L. Zhang, and M.-H. Yang. Learning spatial-temporal regularized correlation filters for visual tracking. In *CVPR*, 2018. 6, 8
- [29] P. Li, D. Wang, L. Wang, and H. Lu. Deep visual tracking: Review and experimental comparison. *PR*, 76:323–338, 2018. 2
- [30] S. Li, T. Xiao, H. Li, B. Zhou, D. Yue, and X. Wang. Person search with natural language description. In *CVPR*, 2017. 2
- [31] X. Li, W. Hu, C. Shen, Z. Zhang, A. Dick, and A. V. D. Hengel. A survey of appearance models in visual object tracking. *ACM TIST*, 4(4):58, 2013. 1, 2
- [32] Y. Li and J. Zhu. A scale adaptive kernel correlation filter tracker with feature integration. In *ECCVW*, 2014. 6, 8
- [33] Z. Li, R. Tao, E. Gavves, C. G. Snoek, A. W. Smeulders, et al. Tracking by natural language specification. In *CVPR*, 2017. 2
- [34] P. Liang, E. Blasch, and H. Ling. Encoding color information for visual tracking: Algorithms and benchmark. *TIP*, 24(12):5630–5644, 2015. 1, 2, 3, 5, 11, 12
- [35] A. Lukezic, T. Vojir, L. C. Zajc, J. Matas, and M. Kristan. Discriminative correlation filter with channel and spatial reliability. In *CVPR*, 2017. 6, 8
- [36] C. Ma, J.-B. Huang, X. Yang, and M.-H. Yang. Hierarchical convolutional features for visual tracking. In *ICCV*, 2015. 6, 8
- [37] C. Ma, X. Yang, C. Zhang, and M.-H. Yang. Long-term correlation tracking. In *CVPR*, 2015. 6, 8
- [38] M. Mueller, N. Smith, and B. Ghanem. A benchmark and simulator for uav tracking. In *ECCV*, 2016. 1, 2, 3, 11
- [39] M. Mueller, N. Smith, and B. Ghanem. Context-aware correlation filter tracking. In *CVPR*, 2017. 6, 8
- [40] M. Müller, A. Bibi, S. Giancola, S. Al-Subaihi, and B. Ghanem. Trackingnet: A large-scale dataset and benchmark for object tracking in the wild. In *ECCV*, 2018. 2, 3, 6
- [41] H. Nam and B. Han. Learning multi-domain convolutional neural networks for visual tracking. In *CVPR*, 2016. 6, 9
- [42] E. Real, J. Shlens, S. Mazzocchi, X. Pan, and V. Vanhoucke. Youtube-boundingboxes: A large high-precision human-annotated data set for object detection in video. In *CVPR*, 2017. 1, 3
- [43] D. A. Ross, J. Lim, R.-S. Lin, and M.-H. Yang. Incremental learning for robust visual tracking. *IJCV*, 77(1-3):125–141, 2008. 6, 8
- [44] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, et al. Imagenet large scale visual recognition challenge. *IJCV*, 115(3):211–252, 2015. 1
- [45] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015. 1
- [46] A. W. Smeulders, D. M. Chu, R. Cucchiara, S. Calderara, A. Dehghan, and M. Shah. Visual tracking: An experimental survey. *TPAMI*, 36(7):1442–1468, 2014. 1, 2, 3
- [47] Y. Song, C. Ma, X. Wu, L. Gong, L. Bao, W. Zuo, C. Shen, R. Lau, and M.-H. Yang. Vital: Visual tracking via adversarial learning. In *CVPR*, 2018. 6
- [48] R. Tao, E. Gavves, and A. W. Smeulders. Siamese instance search for tracking. In *CVPR*, 2016. 6
- [49] J. Valmadre, L. Bertinetto, J. Henriques, A. Vedaldi, and P. H. Torr. End-to-end representation learning for correlation filter based tracking. In *CVPR*, 2017. 6
- [50] J. Valmadre, L. Bertinetto, J. F. Henriques, R. Tao, A. Vedaldi, A. Smeulders, P. Torr, and E. Gavves. Long-term tracking in the wild: A benchmark. In *ECCV*, 2018. 2, 3
- [51] Y. Wu, J. Lim, and M.-H. Yang. Online object tracking: A benchmark. In *CVPR*, 2013. 1, 2, 3
- [52] Y. Wu, J. Lim, and M.-H. Yang. Object tracking benchmark. *TPAMI*, 37(9):1834–1848, 2015. 1, 2, 3, 5, 6, 8, 11
- [53] A. Yilmaz, O. Javed, and M. Shah. Object tracking: A survey. *ACM CSUR*, 38(4):13, 2006. 1, 2
- [54] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson. How transferable are features in deep neural networks? In *NIPS*, 2014. 1
- [55] J. Zhang, S. Ma, and S. Sclaroff. Meem: robust tracking via multiple experts using entropy minimization. In *ECCV*, 2014. 6, 8
- [56] K. Zhang, L. Zhang, Q. Liu, D. Zhang, and M.-H. Yang. Fast visual tracking via dense spatio-temporal context learning. In *ECCV*, 2014. 6, 8
- [57] K. Zhang, L. Zhang, and M.-H. Yang. Real-time compressive tracking. In *ECCV*, 2012. 6, 8
- [58] Y. Zhang, L. Wang, J. Qi, D. Wang, M. Feng, and H. Lu. Structured siamese network for real-time visual tracking. In *ECCV*, 2018. 6

LaSOT: A High-quality Benchmark for Large-scale Single Object Tracking

— Supplementary Material —

Table 4. Details of 70 object categories in LaSOT and comparison with existing dense benchmark

NUS-PRO [27]		OTB-2015 [52]		TC-128 [34]		UAV123 [38]		VOT-17 [26]		NFS [14]		LaSOT	
class	# entries	class	# entries	class	# entries	class	# entries	class	# entries	class	# entries	class	# entries
person	193	person	36	person	45	person	48	person	19	ball	21	airplane	20
head	60	head	26	head	16	car	30	head	5	person	20	basketball	20
car	31	car	12	sphere	8	drone	10	fish	4	animal	10	bear	20
airplane	20	toy	8	2D print	5	wakeboard	10	motorcycle	4	vehicle	9	bicycle	20
boat	20	2D print	4	bicycle	5	boat	9	car	3	shuffleboard	8	bird	20
helicopter	20	cuboid	3	car	5	building	5	drone	3	face	6	boat	20
motorcycle	20	bird	2	ball	4	truck	5	ant	2	cup	4	book	20
drone	1	motorcycle	1	toy	4	bicycle	3	ball	2	dollar	4	bottle	20
-	-	deer	1	hand	3	bird	3	bird	2	aircraft	4	bus	20
-	-	bottle	1	kite	3	-	-	toy	2	airboard	2	car	20
-	-	panda	1	logo	3	-	-	bag	1	fish	2	cat	20
-	-	board	1	cuboid	3	-	-	book	1	motorcycle	2	cattle	20
-	-	can	1	boat	2	-	-	butterfly	1	drone	2	chameleon	20
-	-	dog	1	cup	2	-	-	cable	1	bicycle	2	coin	20
-	-	transformer	1	fish	2	-	-	crab	1	bird	2	crab	20
-	-	bicycle	1	guitar	2	-	-	cat	1	bag	1	crocodile	20
-	-	-	-	bird	2	-	-	flamingo	1	yoyo	1	cup	20
-	-	-	-	microphone	2	-	-	frisbee	1	-	-	deer	20
-	-	-	-	torso	2	-	-	glove	1	-	-	dog	20
-	-	-	-	motorcycle	2	-	-	hand	1	-	-	drone	20
-	-	-	-	airplane	2	-	-	heliocopter	1	-	-	electricFan	20
-	-	-	-	board	1	-	-	leaf	1	-	-	elephant	20
-	-	-	-	bottle	1	-	-	rabbit	1	-	-	flag	20
-	-	-	-	can	1	-	-	sheep	1	-	-	fox	20
-	-	-	-	deer	1	-	-	-	-	-	-	frog	20
-	-	-	-	ring	1	-	-	-	-	-	-	gameTarget	20
-	-	-	-	torus	1	-	-	-	-	-	-	gecko	20
-	-	-	-	-	-	-	-	-	-	-	-	giraffe	20
-	-	-	-	-	-	-	-	-	-	-	-	goldfish	20
-	-	-	-	-	-	-	-	-	-	-	-	gorilla	20
-	-	-	-	-	-	-	-	-	-	-	-	guitar	20
-	-	-	-	-	-	-	-	-	-	-	-	hand	20
-	-	-	-	-	-	-	-	-	-	-	-	hat	20
-	-	-	-	-	-	-	-	-	-	-	-	helmet	20
-	-	-	-	-	-	-	-	-	-	-	-	hippo	20
-	-	-	-	-	-	-	-	-	-	-	-	horse	20
-	-	-	-	-	-	-	-	-	-	-	-	kangaroo	20
-	-	-	-	-	-	-	-	-	-	-	-	kite	20
-	-	-	-	-	-	-	-	-	-	-	-	leopard	20
-	-	-	-	-	-	-	-	-	-	-	-	licensePlate	20
-	-	-	-	-	-	-	-	-	-	-	-	lion	20
-	-	-	-	-	-	-	-	-	-	-	-	lizard	20
-	-	-	-	-	-	-	-	-	-	-	-	microphone	20
-	-	-	-	-	-	-	-	-	-	-	-	monkey	20
-	-	-	-	-	-	-	-	-	-	-	-	motorcycle	20
-	-	-	-	-	-	-	-	-	-	-	-	mouse	20
-	-	-	-	-	-	-	-	-	-	-	-	person	20
-	-	-	-	-	-	-	-	-	-	-	-	pig	20
-	-	-	-	-	-	-	-	-	-	-	-	pool	20
-	-	-	-	-	-	-	-	-	-	-	-	rabbit	20
-	-	-	-	-	-	-	-	-	-	-	-	racing	20
-	-	-	-	-	-	-	-	-	-	-	-	robot	20
-	-	-	-	-	-	-	-	-	-	-	-	rubicCube	20
-	-	-	-	-	-	-	-	-	-	-	-	sepia	20
-	-	-	-	-	-	-	-	-	-	-	-	shark	20
-	-	-	-	-	-	-	-	-	-	-	-	shark	20
-	-	-	-	-	-	-	-	-	-	-	-	sheep	20
-	-	-	-	-	-	-	-	-	-	-	-	skateboard	20
-	-	-	-	-	-	-	-	-	-	-	-	spider	20
-	-	-	-	-	-	-	-	-	-	-	-	squirrel	20
-	-	-	-	-	-	-	-	-	-	-	-	surfboard	20
-	-	-	-	-	-	-	-	-	-	-	-	swing	20
-	-	-	-	-	-	-	-	-	-	-	-	tank	20
-	-	-	-	-	-	-	-	-	-	-	-	tiger	20
-	-	-	-	-	-	-	-	-	-	-	-	train	20
-	-	-	-	-	-	-	-	-	-	-	-	truck	20
-	-	-	-	-	-	-	-	-	-	-	-	turtle	20
-	-	-	-	-	-	-	-	-	-	-	-	umbrella	20
-	-	-	-	-	-	-	-	-	-	-	-	volleyball	20
-	-	-	-	-	-	-	-	-	-	-	-	yoyo	20

1. Details of 70 Object Categories in LaSOT and Comparison with Existing Dense Benchmark

LaSOT consists of 70 object categories with each containing 20 videos, as shown in Tab. 4. Most of 70 classes are chosen from the 1,000 classes in ImageNet [12], with a few exceptions such as *drone* and *gametarget*, which are carefully selected by the experts for tracking. The selection of each category must be agreed upon by all the experts to ensure its usability for visual tracking. In addition, we also compare the object categories of different dense benchmarks. As shown in Tab. 4, the number of object categories in LaSOT is two times more than that of existing benchmarks (*e.g.*, TC-128 [34] with 27 classes). Moreover, LaSOT eliminates the category bias of dataset for tracking while others do not.

2. Training/Testing Split in Protocol II

In protocol II, we split LaSOT into *training* and *testing* sets. The *training* set contains of 1,120 videos (*i.e.*, 16 sequences for each category) with 2.83M frames in total. The rest 280 videos (*i.e.*, 4 sequences for each category) with 69K frames are used for testing.

Table 5. Comparison between *training* and *testing* sets of LaSOT.

	Video	Min frames	Mean frames	Median frames	Max frames	Total frames	Total duration
LaSOT _{training}	1,120	1,000	2,529	2,043	11,397	283M	26.2 hours
LaSOT _{testing}	280	1,000	2,448	2,102	9,999	69K	6.3 hours
LaSOT	1,400	1,000	2,506	2,053	11,397	3.52M	32.5 hours

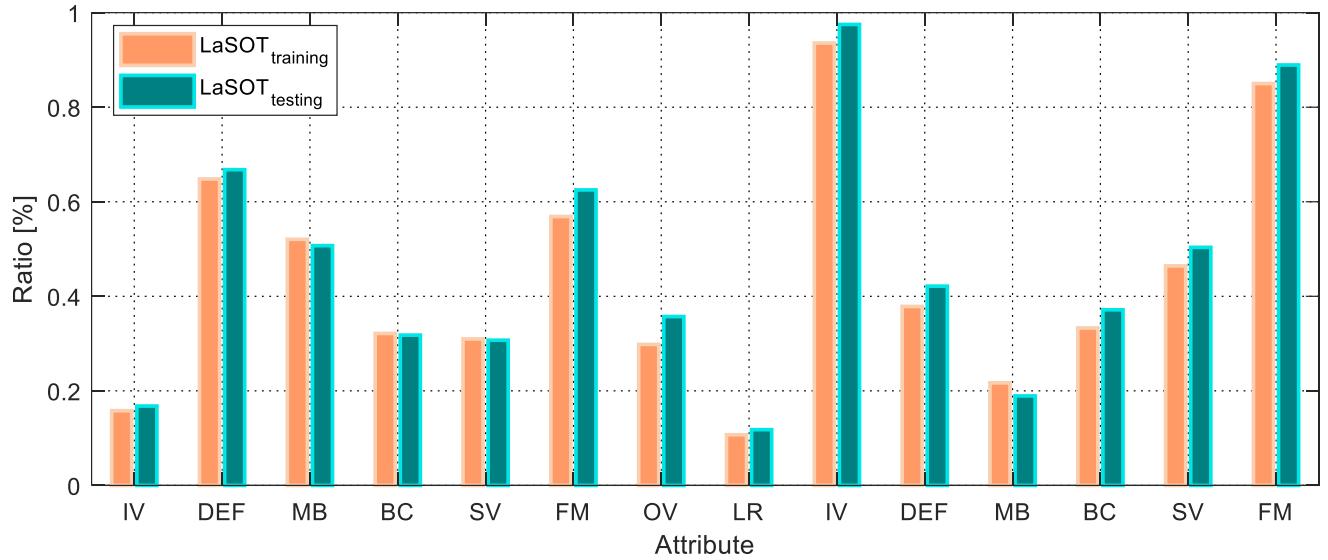


Figure 8. Comparison of sequence distribution in each attribute between *training* and *testing* sets. Best viewed in color.

Tab. 5 reports the detailed comparison between the *training* and the *testing* sets of LaSOT. We observe that the *min frames*, *mean frames*, *median frames* and *max frames* are similar between these two subsets. In addition, as shown in Fig. 8, we can see that the ratios of sequences in all 14 attributes are similar. Both Tab. 5 and Fig. 8 evidence the consistency of our training/testing split.

3. Detailed Attribute-based Performance under Protocol I

Fig. 9 shows the performance of trackers on each attribute using precision under protocol I.

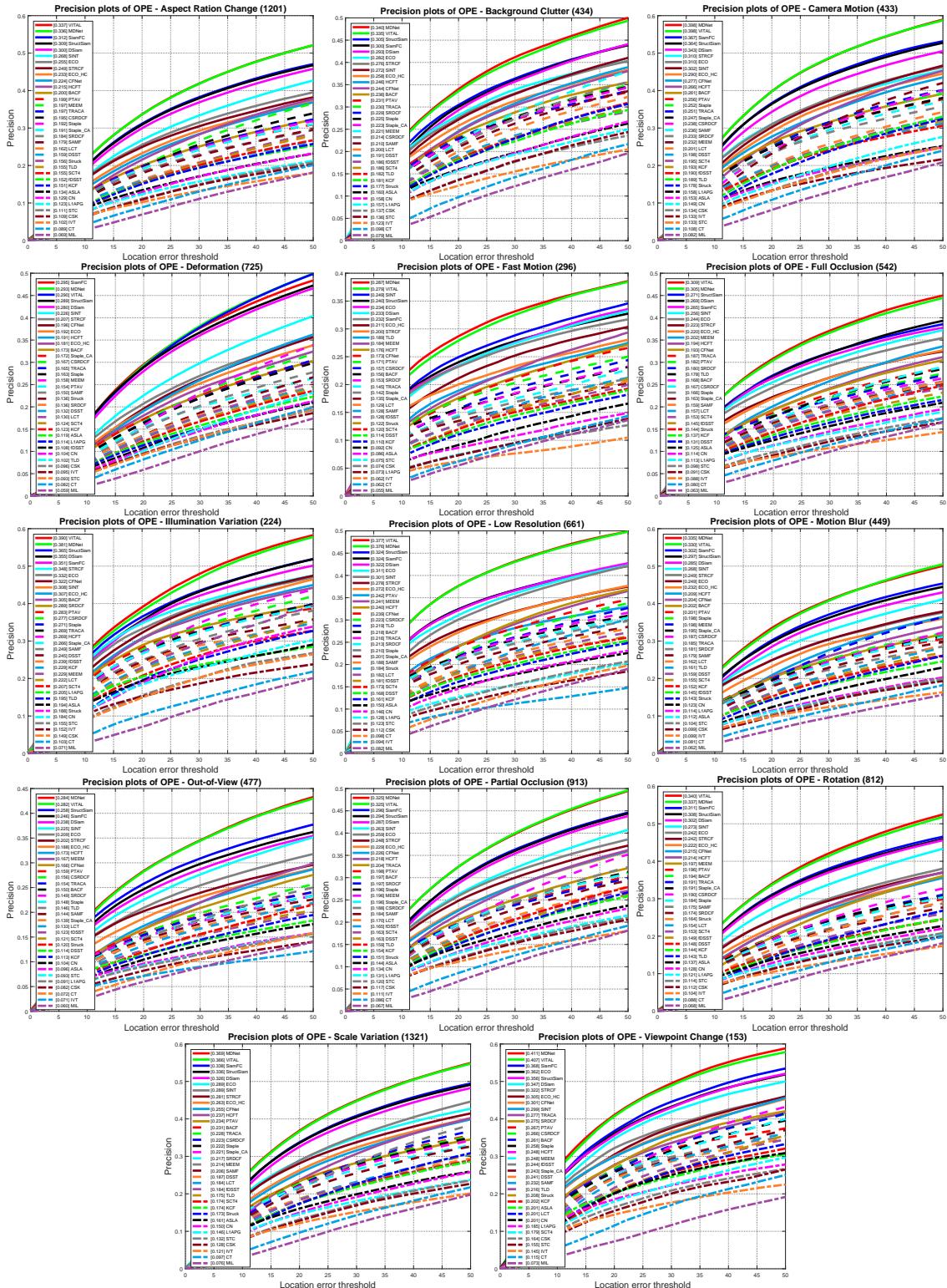


Figure 9. Performance of trackers on each attribute using precision under protocol I. Best viewed in color.

Fig. 10 shows the performance of trackers on each attribute using normalized precision under protocol I.

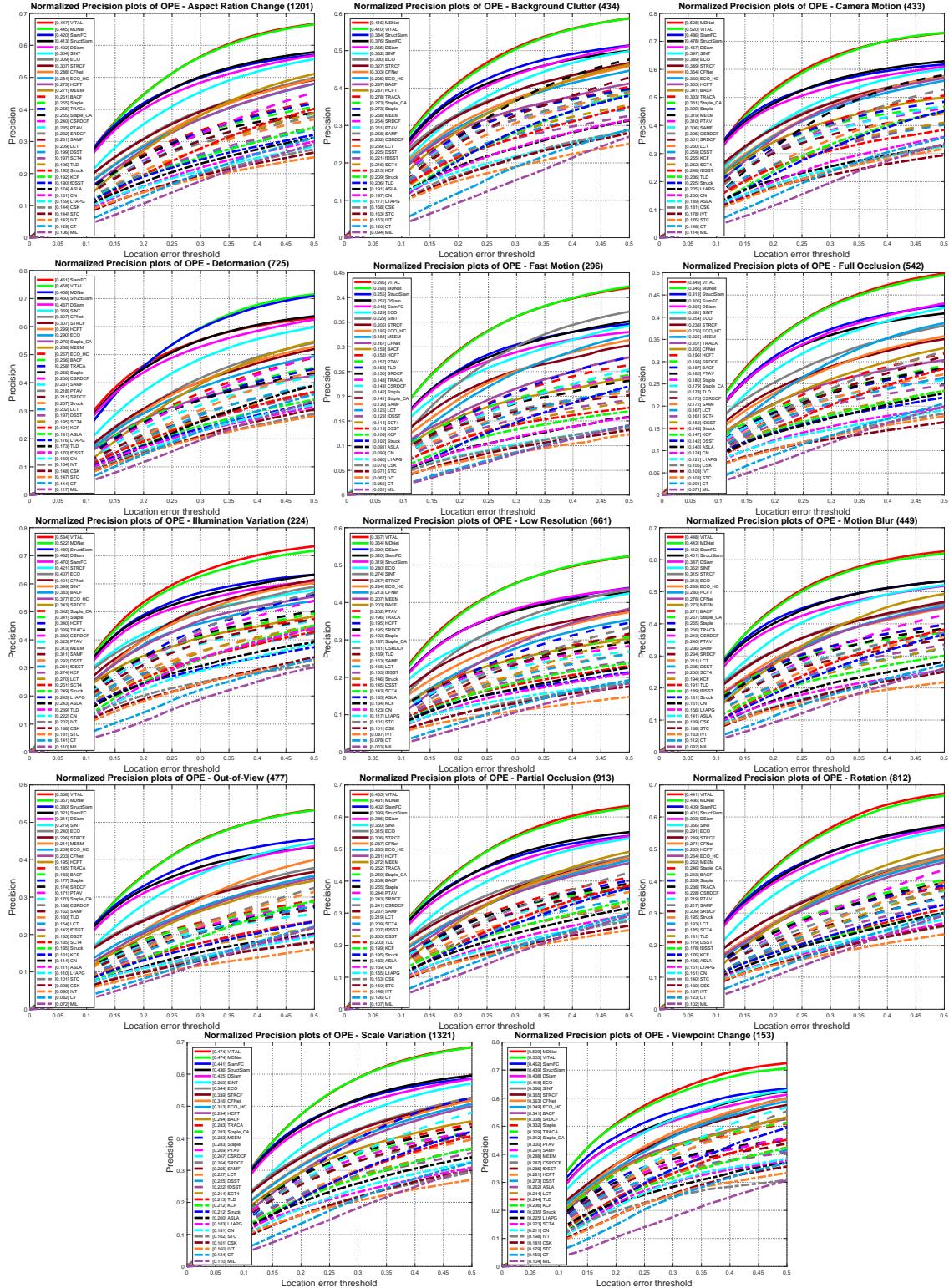


Figure 10. Performance of trackers on each attribute using precision under protocol I. Best viewed in color.

4. Detailed Attribute-based Performance under Protocol II

Fig. 11 shows the performance of trackers on each attribute using precision under protocol II.

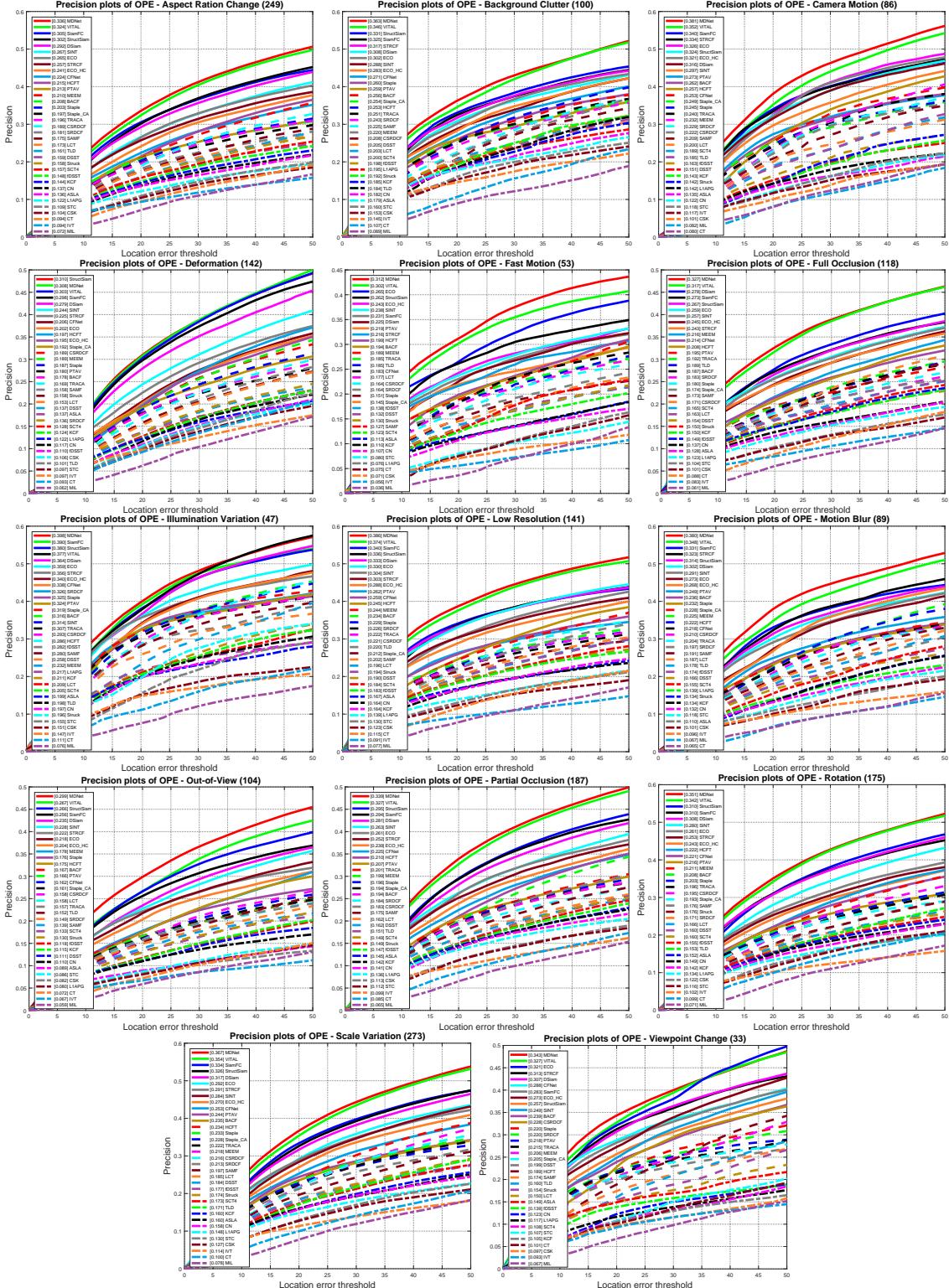


Figure 11. Performance of trackers on each attribute using precision under protocol II. Best viewed in color.

Fig. 12 shows the performance of trackers on each attribute using normalized precision under protocol II.

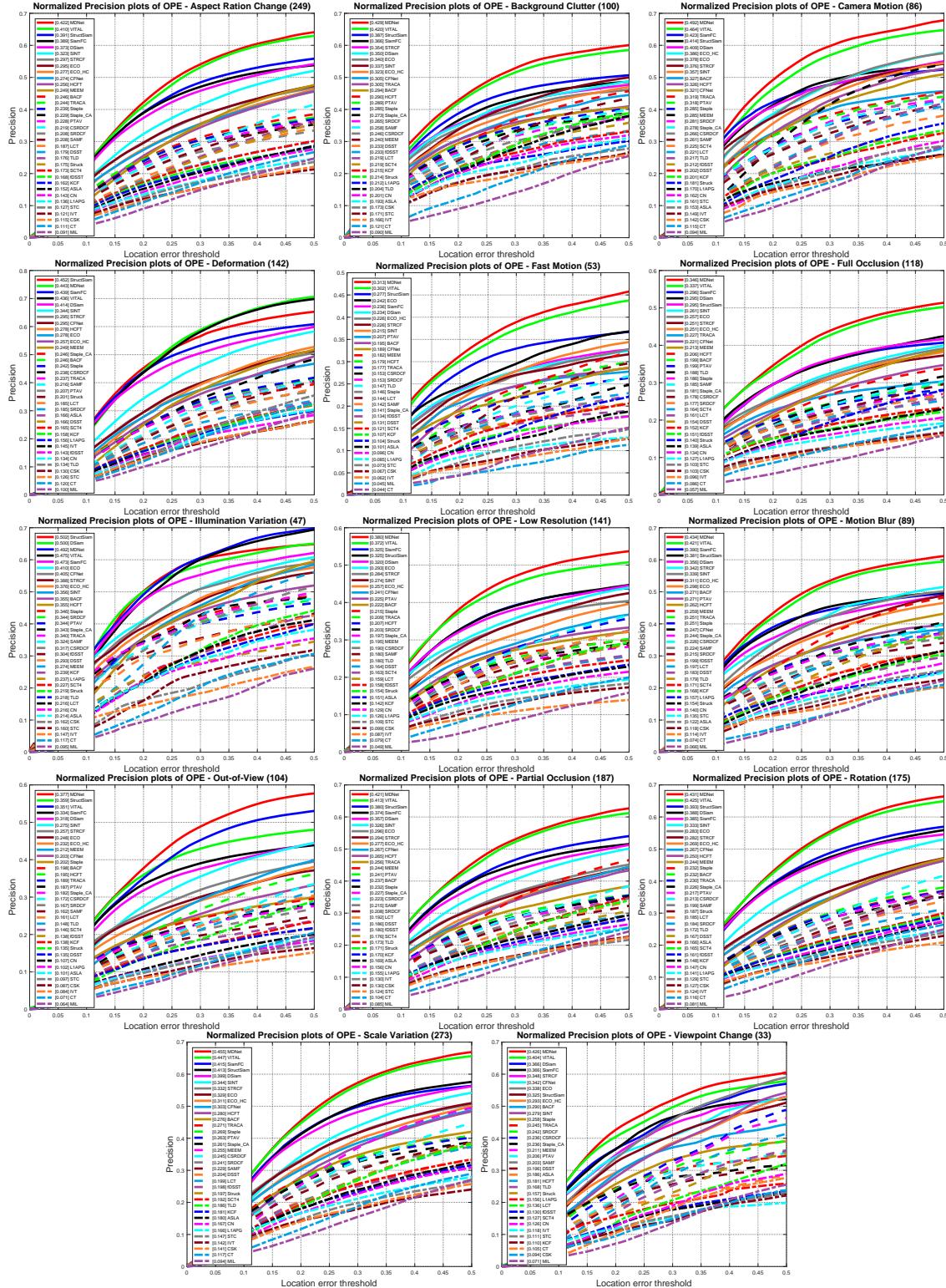


Figure 12. Performance of trackers on each attribute using precision under protocol II. Best viewed in color.

Fig. 13 shows the performance of trackers on each attribute using success under protocol II.

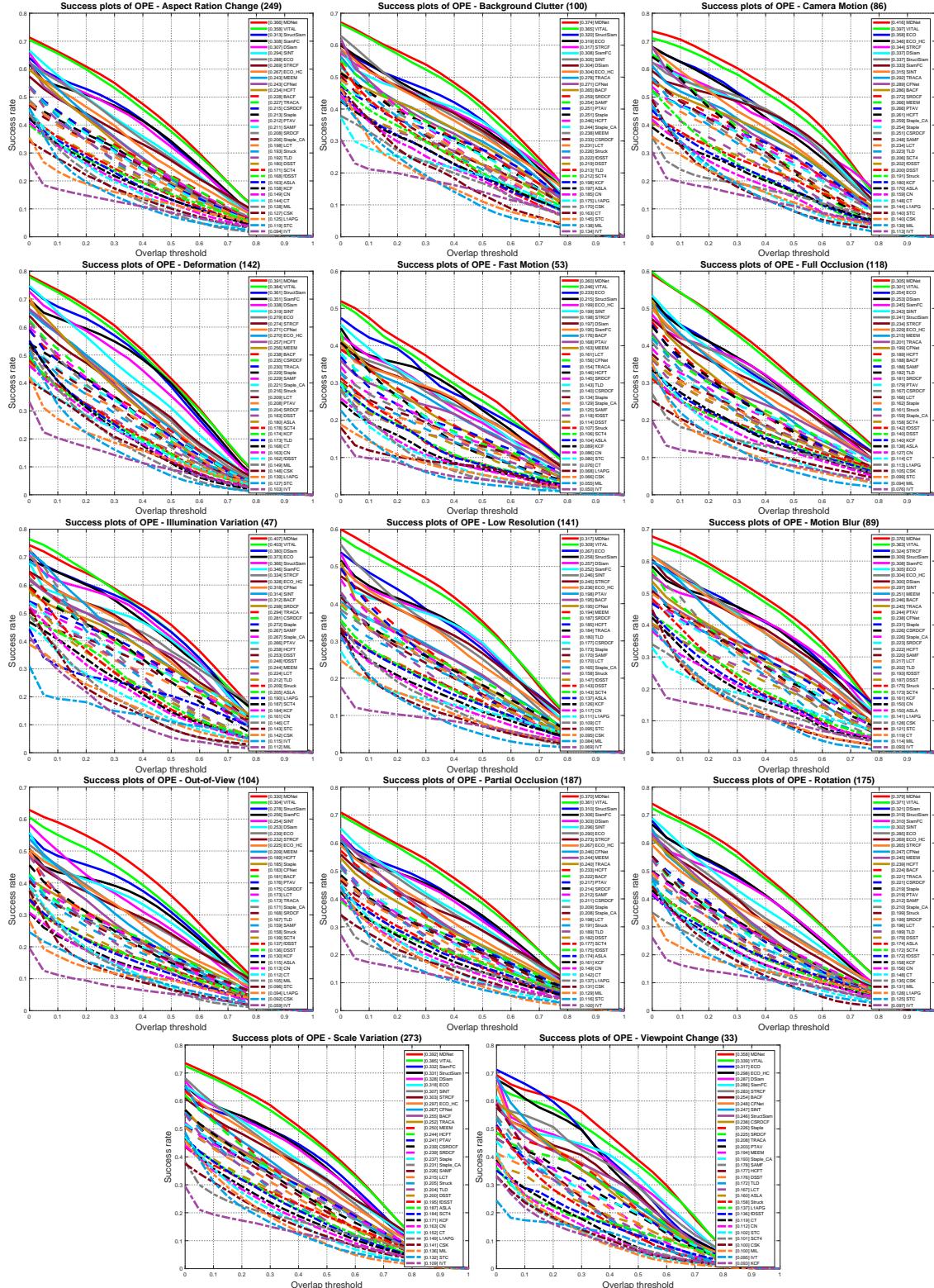


Figure 13. Performance of trackers on each attribute using success under protocol II. Best viewed in color.