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

Heng Fan[1*] Liting Lin[2*] Fan Yang[1*] Peng Chu[1*] Ge Deng[1] Sijia Yu[1] Hexin Bai[1]
Yong Xu[2] Chunyuan Liao[3] Haibin Ling[1†]

[1]Department of Computer and Information Sciences, Temple University, Philadelphia, USA
[2]School of Computer Science & Engineering, South China Univ. of Tech., Guangzhou,
Peng Cheng Laboratory, Shenzhen, China
[3]Meitu HiScene Lab, HiScene Information Technologies, Shanghai, China

https://cis.temple.edu/lasot/

## Abstract

*In this paper, we present* **LaSOT**, *a high-quality benchmark for* **La***rge-scale* **S***ingle* **O***bject* **T***racking. 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 video 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 with 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 for improvements.*

## 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 [32, 47, 54]. With considerable progresses in the tracking community, numerous algorithms have been proposed. In this process, tracking benchmarks have played a vital role in objectively evaluating and comparing different trackers. Nevertheless, further development and assessment of tracking algorithms are restricted by existing benchmarks with several issues:

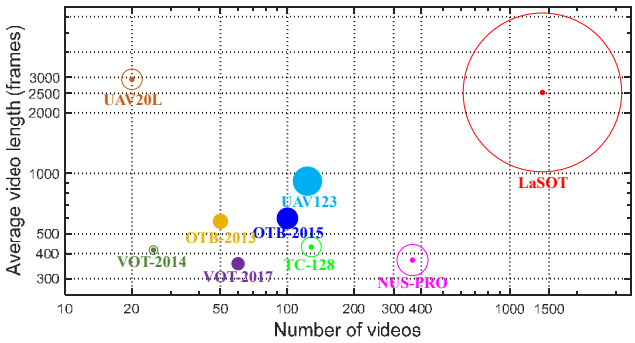



Figure 1. Summaries of existing tracking benchmarks with high-quality dense (per frame) annotations, including OTB-2013 [52], OTB-2015 [53], TC-128 [35], NUS-PRO [28], UAV123 [39], UAV20L [39], VOT-2014 [26], VOT-2017 [27] and LaSOT. The circle diameter is in proportion to the number of frames of a benchmark. The proposed LaSOT is *larger* than all other benchmarks, and focused on *long-term* tracking. Best viewed in color.

**Small-scale.** Deep representations have been popularly applied to modern object tracking algorithms, and demonstrated state-of-the-art performances. However, it is difficult to train a deep tracker using *tracking-specific* videos due to the scarcity of *large-scale* tracking datasets. 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.*, [46] and [18]) from image classification for deep feature extraction or the sequences from video object detection (*e.g.*, [45] and [43]) for deep feature learning, which may result in suboptimal tracking performance because of the intrinsic differences among different tasks [55]. Moreover, large scale benchmarks are desired for more reliable evaluation results.

**Lack of high-quality dense annotations.** For tracking,

*Authors make equal contributions to this work.
†Corresponding author.

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** [52] | 51 | 71 | 578 | 392 | 3,872 | 29**K** | 16.4 **min** | 30 fps | ✗ | 10 | ✗ | 11 | ✗ |
| **OTB-2015** [53] | 100 | 71 | 590 | 393 | 3,872 | 59**K** | 32.8 **min** | 30 fps | ✗ | 16 | ✗ | 11 | ✗ |
| **TC-128** [35] | 128 | 71 | 429 | 365 | 3,872 | 55**K** | 30.7 **min** | 30 fps | ✗ | 27 | ✗ | 11 | ✗ |
| **VOT-2014** [26] | 25 | 164 | 409 | 307 | 1,210 | 10**K** | 5.7 **min** | 30 fps | ✗ | 11 | ✗ | n/a | ✗ |
| **VOT-2017** [27] | 60 | 41 | 356 | 293 | 1,500 | 21**K** | 11.9 **min** | 30 fps | ✗ | 24 | ✗ | n/a | ✗ |
| **NUS-PRO** [28] | 365 | 146 | 371 | 300 | 5,040 | 135**K** | 75.2 **min** | 30 fps | ✗ | 8 | ✗ | n/a | ✗ |
| **UAV123** [39] | 123 | 109 | 915 | 882 | 3,085 | 113**K** | 62.5 **min** | 30 fps | ✗ | 9 | ✗ | 12 | ✗ |
| **UAV20L** [39] | 20 | 1,717 | 2,934 | 2,626 | 5,527 | 59**K** | 32.6 **min** | 30 fps | ✗ | 5 | ✗ | 12 | ✗ |
| **NfS** [14] | 100 | 169 | 3,830 | 2,448 | 20,665 | 383**K** | 26.6 **min** | 240 fps | ✗ | 17 | ✗ | 9 | ✗ |
| **GOT-10k** [22] | 10,000 | - | - | - | - | 1.5**M** | - | 10 fps | ✓ | 563 | ✗ | 6 | ✗ |
| **LaSOT** | 1,400 | 1,000 | 2,506 | 2,053 | 11,397 | 3.52**M** | 32.5 **hours** | 30 fps | ✓ | 70 | ✓ | 14 | ✓ |

dense (*i.e.*, per frame) annotations with high precision are of importance for several reasons. (i) They ensure more accurate and reliable evaluations; (ii) they offer desired training samples for the training of tracking algorithms; and (iii) they provide rich temporal contexts among consecutive frames that are important for tracking tasks. It is worth noting that there are recently proposed benchmarks toward large-scale and long-term tracking, such as [41] and [51], 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.

**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 for 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 practice.

**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 only a few categories (see Tab. 1) with unbalanced numbers of videos.

In the literature, many datasets have been proposed to deal with the issues above: *e.g.*, [39, 51] for long-term tracking, [41] for large-scale, [52, 35, 25] for precise dense annotations. Nevertheless, none of them addresses all the issues, which motivates the proposal of LaSOT.

### 1.1. Contribution

With the above motivations, we provide the community a novel benchmark for **La**rge-scale **S**ingle **O**bject **T**racking (LaSOT) with multi-fold contributions:

1) LaSOT consists of 1,400 videos with average 2,512 frames per sequence. Each frame 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 our knowledge, LaSOT is the largest benchmark with high-quality manual dense annotations for object tracking to date. By releasing LaSOT, we aim to offer a dedicated platform for the development and assessment of tracking algorithms.
2) Different from existing datasets, LaSOT provides both visual bounding box annotations and rich natural language specification, which has recently been proven to be beneficial for various vision tasks (*e.g.*, [21, 31]) including visual tracking [34]. By doing so, we aim to encourage and facilitate explorations of integrating visual and lingual features for robust tracking performance.
3) To assess existing trackers and provide extensive baselines for future comparisons on LaSOT, we evaluate 35 representative trackers under different protocols, and analyze their performances using different metrics.

## 2. Related Work

With considerable progresses in the tracking community, many trackers 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 [32, 47, 54, 30] for tracking algorithms.

For a systematic review, we intentionally classify tracking benchmarks into two types: one with dense manual annotations (referred to 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 [52, 53], TC-128 [35], VOT [25], NUS-PRO [28], UAV [39], NfS [14] and GOT-10k [22].

**OTB.** OTB-2013 [52] firstly contributes a testing dataset

by collecting 51 videos with manually 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 [53] by introducing extra 50 sequences.

**TC-128.** TC-128 [35] comprises 128 videos that are specifically designated to evaluate color-enhanced trackers. The videos are labeled with 11 similar attributes as in OTB [52].

**VOT.** VOT [25] 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. Each frame in the VOT datasets is annotated with a rotated bounding box with several attributes.

**NUS-PRO.** NUS-PRO [28] 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 [39] 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.

**GOT-10k.** GOT-10k [22] consists of 10,000 videos, aiming to provide rich motion trajectories for developing and evaluating trackers.

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. Representatives of this type of benchmarks include ALOV [47], TrackingNet [41] and OxUvA [51]. **ALOV** [47] consists of 314 sequences labeled in 14 attributes. Instead of densely annotating each frame, ALOV provides annotations every 5 frames. **TrackingNet** [41] is a subset of the video object detection benchmark YT-BB [43] by selecting 30K videos, each of which is annotated 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 [53], 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** [51] also comes from YT-BB [43]. Unlike TrackingNet, OxUvA is focused on long-term tracking. It contains 366 videos 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.

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* due to sparse annotation or *inaccurate* due to potentially unreliable annotation. By contrast, 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*.

1) **Large-scale.** One of the key motivations of LaSOT is to provide a dataset for training data-hungry deep trackers, which require 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.
2) **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 faithful evaluation. For this purpose, each sequence in LaSOT is manually annotated with additional careful inspection and fine-tuning.
3) **Long-term tracking.** In comparison with short-term tracking, long-term tracking can reflect more practical performance of a tracker in the wild. We ensure that each sequence comprises *at least* 1,000 frames, and the average sequence length in LaSOT is around 2,500 frames.
4) **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.
5) **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*) that are carefully chosen for popular tracking applications. 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 found 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 dataset by gathering 1,400 sequences with 3.52 million frames from YouTube under Creative Commons licence. The average video length of LaSOT is 2,512 frames (*i.e.*, 84 seconds for 30 fps). The shortest 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 draws/edits 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.

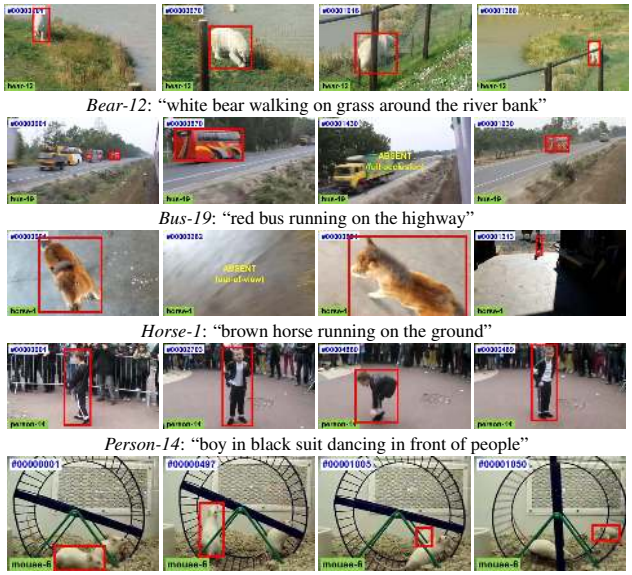



Figure 2. Example sequences and annotations of our LaSOT. We focus on long-term videos in which target objects may disappear, and then re-enter the view again. In addition, we provide natural language specification for each sequence. Best viewed in color.

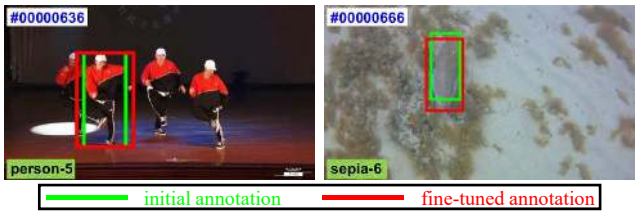


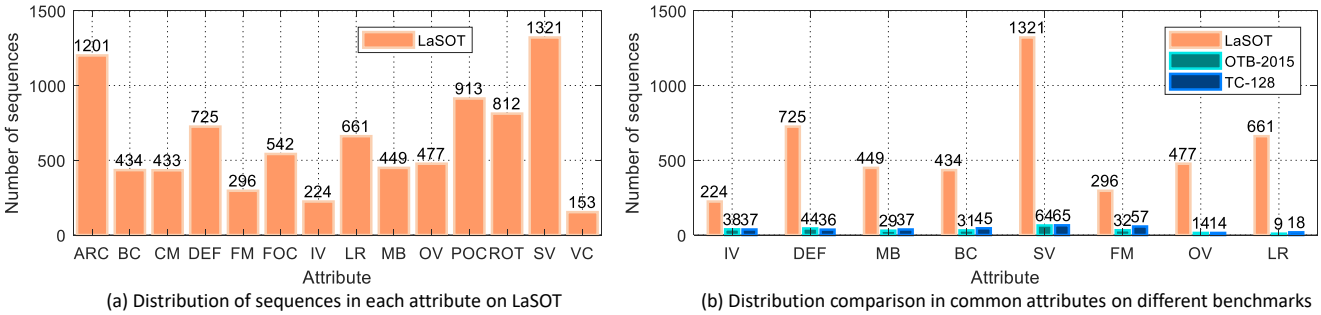

Figure 3. Examples of fine-tuning initial annotations.

The natural language specification of a sequence is represented by a sentence that describes the color, behavior and surroundings of the target. For LaSOT, we provide 1,400 sentences for all videos. Note that the lingual description aims to provide auxiliary help for tracking. For instance, if a tracker generates proposals for further processing, the lingual specification can assist in reducing the ambiguity among them by serving as a global semantic guidance.

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 adjusts them if necessary. Then, the annotation results are reviewed by the validation team containing several (typ-

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 rage [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 ratio of bounding box aspect ratio is outside the rage [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 benchmarks. Best viewed in color.

ically three) experts. If an annotation result is not unanimously agreed by the members of validation team, it will be sent back to the original 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. And 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.

### 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 ratio change (ARC). The attributions are defined in Tab. 2, and Fig. 4 (a) shows 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 [53] and TC-128 [35] 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 testing[1]. In specific, the *training* subset contains 1,120 videos with 2.83M frames, and the *testing* subset consists of 280 sequences with 690K 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 [53]), we perform an One-Pass Evaluation (OPE) and measure the **precision**, **normalized precision** and **success** of different tracking algorithms under two protocols.

[1]The training/testing split is shown in the **supplementary material**.

The precision is computed by comparing the distance between tracking result and groundtruth bounding box in pixels. Different trackers 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 normalize the precision as in [41]. With the normalized precision metric, we rank tracking algorithms using the Area Under the Curve (AUC) between 0 to 0.5. Please refer to [41] about the normalized precision metric. The success is computed as the Intersection over Union (IoU) between tracking result and 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 [42], TRACA [5], CFNet [50], SiamFC [4], StructSiam [59], DSiam [16], SINT [49] and VITAL [48]), correlation filter trackers with hand-crafted features (*e.g.*, ECO_HC [7], DSST [8], CN [11], CSK [19], KCF [20], fDSST [9], SAMF [33], SCT4 [6], STC [57] and Staple [3]) or deep features (*e.g.*, HCFT [37] and ECO [7]) and regularization techniques (*e.g.*, BACF [15], SRDCF [10], CSRDCF [36], Staple_CA [40] and STRCF [29]), ensemble trackers (*e.g.*, PTAV [13], LCT [38], MEEM [56] and TLD [24]), sparse trackers (*e.g.*, L1APG [2] and ASLA [23]), other representatives (*e.g.*, CT [58], IVT [44], MIL [1] and Struck [17]). Tab. 3 summarizes these trackers with their representation schemes and search strategies in a chronological order.

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.

| | | Representation | | | | | | | Search | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | PCA | Sparse | Color | Pixel | HoG | H or B | Deep | PF | RS | DS |
| IVT [44] | IJCV08 | ✓ | | | | | | | ✓ | | |
| MIL [1] | CVPR09 | | | | | | H | | | | ✓ |
| Struck [17] | ICCV11 | | | | | | H | | | | ✓ |
| L1APG [2] | CVPR12 | | ✓ | | | | | | ✓ | | |
| ASLA [23] | CVPR12 | | ✓ | | | | | | ✓ | | |
| CSK [19] | ECCV12 | | | | ✓ | | | | | | ✓ |
| CT [58] | ECCV12 | | | | | | H | | | | ✓ |
| TLD [24] | PAMI12 | | | | | | B | | | | ✓ |
| CN [11] | CVPR14 | | | ✓ | ✓ | | | | | | ✓ |
| DSST [8] | BMVC14 | | | | ✓ | ✓ | | | | | ✓ |
| MEEM [56] | ECCV14 | | | | ✓ | | | | | ✓ | |
| STC [57] | ECCV14 | | | | ✓ | | | | | | ✓ |
| SAMF [33] | ECCVW14 | | | ✓ | ✓ | ✓ | | | | | ✓ |
| LCT [38] | CVPR15 | | | | ✓ | ✓ | | | | | ✓ |
| SRDCF [10] | ICCV15 | | | | | ✓ | | | | | ✓ |
| HCFT [37] | ICCV15 | | | | | | | ✓ | | | ✓ |
| KCF [20] | PAMI15 | | | | | ✓ | | | | | ✓ |
| Staple [3] | CVPR16 | | | ✓ | | ✓ | | | | | ✓ |
| SINT [49] | CVPR16 | | | | | | | ✓ | | ✓ | |
| SCT4 [6] | CVPR16 | | | | | ✓ | | | | | ✓ |
| MDNet [42] | CVPR16 | | | | | | | ✓ | | ✓ | |
| SiamFC [4] | ECCVW16 | | | | | | | ✓ | | | ✓ |
| Staple_CA[40] | CVPR17 | | | ✓ | | ✓ | | | | | ✓ |
| ECO_HC [7] | CVPR17 | | | | | ✓ | | | | | ✓ |
| ECO [7] | CVPR17 | | | | | | | ✓ | | | ✓ |
| CFNet [50] | CVPR17 | | | | | | | ✓ | | | ✓ |
| CSRDCF [36] | CVPR17 | | | ✓ | ✓ | ✓ | | | | | ✓ |
| PTAV [13] | ICCV17 | | | | ✓ | ✓ | | ✓ | | | ✓ |
| DSiam [16] | ICCV17 | | | | | | | ✓ | | | ✓ |
| BACF [15] | ICCV17 | | | | | ✓ | | | | | ✓ |
| fDSST [9] | PAMI17 | | | | ✓ | ✓ | | | | | ✓ |
| VITAL [48] | CVPR18 | | | | | | | ✓ | | ✓ | |
| TRACA [5] | CVPR18 | | | | | | | ✓ | | | ✓ |
| STRCF [29] | CVPR18 | | | | | ✓ | | | | | ✓ |
| StructSiam [59] | ECCV18 | | | | | | | ✓ | | | ✓ |

### 4.3. Evaluation Results with Protocol I

**Overall performance.** Protocol I aims at providing large-scale evaluations on all 1,400 videos in LaSOT. Each tracker is used as it 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. SimaFC 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. 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 [53]), the performances on LaSOT are severely degraded because of a large mount of non-rigid target objects and challenging factors involved 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.** To analyze different challenges faced by existing trackers, we evaluate all tracking algorithms on 14 attributes. We show the results on three most challenging attributes, *i.e.*, *fast motion*, *out-of-view* and *full occlusion*, in Fig. 6 and refer the readers to **supplementary material** for detailed attribute-based evaluation.

**Qualitative evaluation.** To qualitatively analyze different trackers and provide guidance for future research, we show the qualitative evaluation results of six representative trackers, including MDNet, SiamFC, ECO, PTAV, Staple and MEEM, in six typical hard challenges containing *fast motion*, *full occlusion*, *low resolution*, *out-of-view*, *aspect ratio change* and *background clutter* in Fig. 7. From Fig. 7, we observe that, for videos with *fast motion*, *full occlusion* and *out-of-view* (*e.g.*, *Yoyo-3*, *Goldfish-4* and *Basketball-15*), the trackers are prone to lose the target because existing trackers usually perform localization from a small local region. To handle these challenges, a potential solution is to leverage an instance-specific detector to locate the target for subsequent tracking. Trackers easily drift in video with low

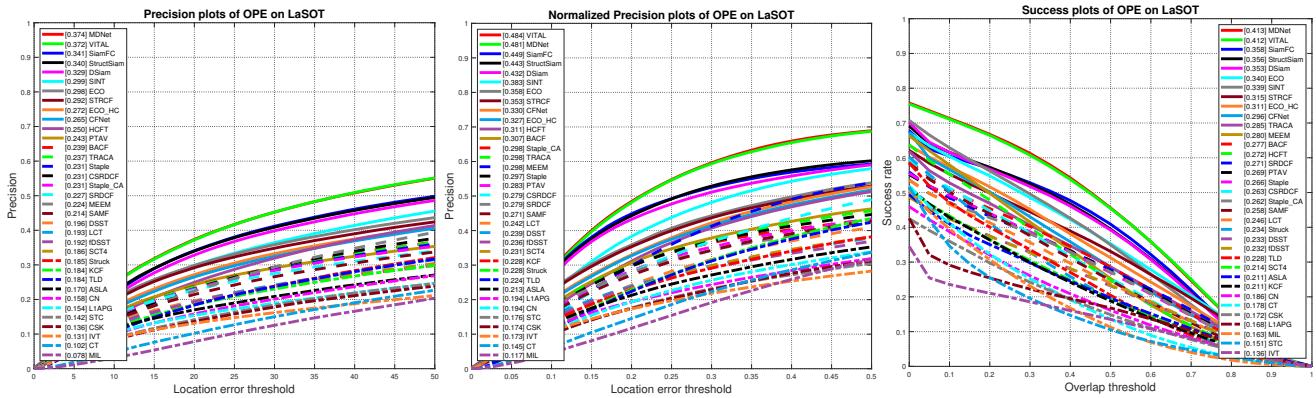



Figure 5. Evaluation results on LaSOT under protocol I using precision, normalized precision and success. Best viewed in color.

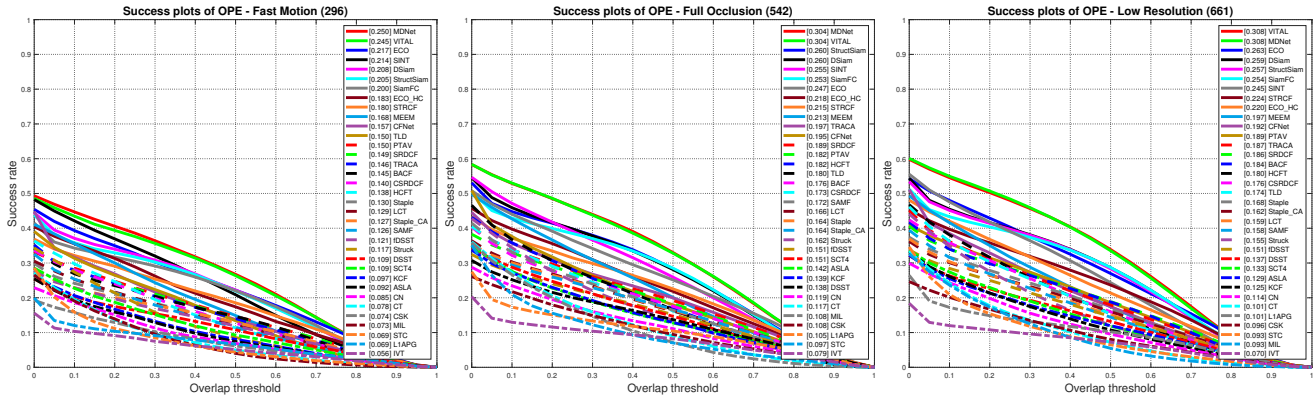



Figure 6. Performances of trackers on three most challenging attributes under protocol I using success. Best viewed in color.


Figure 7. Qualitative evaluation in six typical hard challenges: *Yoyo-3* with fast motion, *Goldfish-4* with full occlusion, *Pool-4* with low-resolution, *Basketball-15* with out-of-view, *Train-1* with aspect ration change and *Person-2* with background clutter. Best viewed in color.

resolution (*e.g.*, *Pool-4*) due to the ineffective representation for small target. A solution for deep feature based trackers is to combine features from multiple scales to incorporate details into representation. Video with *aspect ratio change* is difficult as most existing trackers either ignore this issue or adopt a simple method (*e.g.*, random search or pyramid strategy) to deal with it. Inspired from the success of deep learning based object detection, a generic regressor can be leveraged to reduce the effect of *aspect ratio change* (and *scale change*) on tracking. For sequence with *background clutter*, trackers drift due to less discriminative representation for target and background. A possible solution to alleviate this problem is to utilize the contextual information to enhance the discriminability.

### 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 the *training* set to develop their trackers and assess their performances on the *test* set. In order to provide baselines

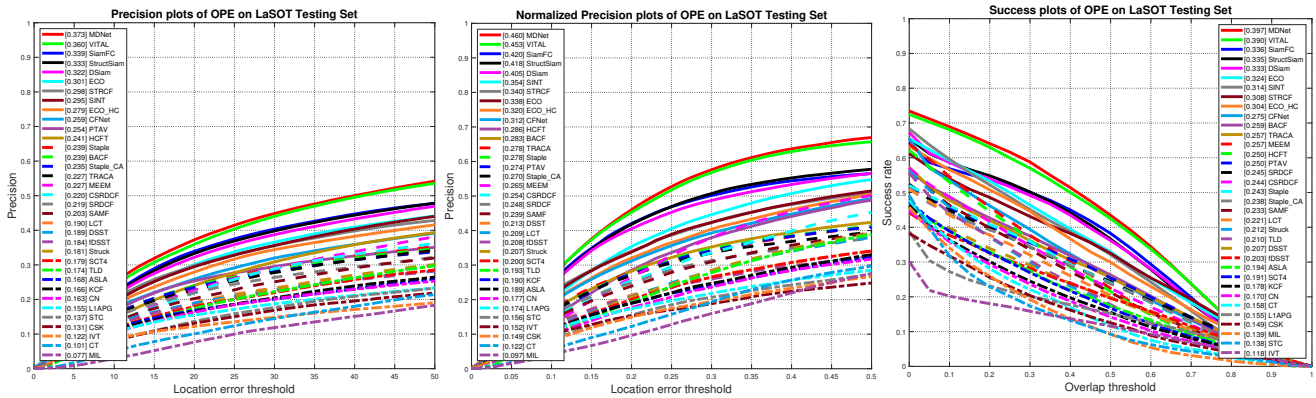



Figure 8. Evaluation results on LaSOT under protocol II using precision, normalized precision and success. Best viewed in color.

and comparisons on the *testing* set, we evaluate the 35 tracking algorithms. Each tracker is used as it is for evaluation without any modification or re-training. The evaluation results are shown in Fig. 8 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 slightly lower scores in accuracy than MDNet and VITAL, SiamFC runs much faster and achieves real-time running speed, showing good balance between accuracy and efficiency. 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 [42] 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.

### 4.5. Retraining Experiment on LaSOT

We conduct the experiment by retraining SiamFC [4] on the training set of LaSOT to demonstrate how deep learning based tracker is improved using more data. Tab. 4 reports the results on OTB-2013 [52] and OTB-2015 [53] and comparisons with the performance of original SiamFC trained on ImageNet Video [45]. Note that, we utilize color images for training, and apply a pyramid with 3 scales for tracking, *i.e.*, SiamFC-3s (color). All parameters for training and tracking are kept the same in these two experiments. From Tab. 4, we observe consistent performance gains on the two benchmarks, showing the importance of specific large-scale training set for deep trackers.

Table 4. Retraining of SiamFC [4] on LaSOT.

| Training data | | SiamFC-3s (color) | |
|---|---|---|---|
| | | ImageNet Video [45] | LaSOT training set |
| OTB-2013 [52] | Precision | 0.803 | 0.816 (↑**1.3%**) |
| | Success | 0.588 | 0.608 (↑**2.0%**) |
| OTB-2015 [53] | Precision | 0.756 | 0.777 (↑**2.1%**) |
| | Success | 0.565 | 0.582 (↑**1.7%**) |

## 5. Conclusion

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 B. Huang, X. Li, Q. Zhou, L. Chen, J. Liang, J. Wang and anonymous volunteers for their help in constructing LaSOT. This work is supported in part by the China National Key Research and Development Plan (Grant No. 2016YFB1001200), and in part by US NSF Grants 1618398, 1407156 and 1350521, and Yong Xu thanks the supports by National Nature Science Foundation of China (U1611461 and 61672241), the Cultivation Project of Major Basic Research of NSF-Guangdong Province (2016A030308013).

## References


[1] Boris Babenko, Ming-Hsuan Yang, and Serge Belongie. Visual tracking with online multiple instance learning. In *CVPR*, 2009. 6

[2] Chenglong Bao, Yi Wu, Haibin Ling, and Hui Ji. Real time robust l1 tracker using accelerated proximal gradient approach. In *CVPR*, 2012. 6

[3] Luca Bertinetto, Jack Valmadre, Stuart Golodetz, Ondrej Miksik, and Philip HS Torr. Staple: Complementary learners for real-time tracking. In *CVPR*, 2016. 6

[4] Luca Bertinetto, Jack Valmadre, Joao F Henriques, Andrea Vedaldi, and Philip HS Torr. Fully-convolutional siamese networks for object tracking. In *ECCVW*, 2016. 6, 8

[5] Jongwon Choi, Hyung Jin Chang, Tobias Fischer, Sangdoo Yun, Kyuewang Lee, Jiyeoup Jeong, Yiannis Demiris, and Jin Young Choi. Context-aware deep feature compression for high-speed visual tracking. In *CVPR*, 2018. 6

[6] Jongwon Choi, Hyung Jin Chang, Jiyeoup Jeong, Yiannis Demiris, and Jin Young Choi. Visual tracking using attention-modulated disintegration and integration. In *CVPR*, 2016. 6

[7] Martin Danelljan, Goutam Bhat, Fahad Shahbaz Khan, and Michael Felsberg. Eco: Efficient convolution operators for tracking. In *CVPR*, 2017. 6

[8] Martin Danelljan, Gustav Häger, Fahad Khan, and Michael Felsberg. Accurate scale estimation for robust visual tracking. In *BMVC*, 2014. 6

[9] Martin Danelljan, Gustav Häger, Fahad Shahbaz Khan, and Michael Felsberg. Discriminative scale space tracking. *TPAMI*, 39(8):1561–1575, 2017. 6

[10] Martin Danelljan, Gustav Hager, Fahad Shahbaz Khan, and Michael Felsberg. Learning spatially regularized correlation filters for visual tracking. In *ICCV*, 2015. 6

[11] Martin Danelljan, Fahad Shahbaz Khan, Michael Felsberg, and Joost Van de Weijer. Adaptive color attributes for real-time visual tracking. In *CVPR*, 2014. 6

[12] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 4

[13] Heng Fan and Haibin Ling. Parallel tracking and verifying: A framework for real-time and high accuracy visual tracking. In *ICCV*, 2017. 6

[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 *ICCV*, 2017. 2, 3

[15] Hamed Kiani Galoogahi, Ashton Fagg, and Simon Lucey. Learning background-aware correlation filters for visual tracking. In *ICCV*, 2017. 6

[16] Qing Guo, Wei Feng, Ce Zhou, Rui Huang, Liang Wan, and Song Wang. Learning dynamic siamese network for visual object tracking. In *ICCV*, 2017. 6

[17] Sam Hare, Amir Saffari, and Philip H. S. Torr. Struck: Structured output tracking with kernels. In *ICCV*, 2011. 6

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 1

[19] João F Henriques, Rui Caseiro, Pedro Martins, and Jorge Batista. Exploiting the circulant structure of tracking-by-detection with kernels. In *ECCV*, 2012. 6

[20] João F Henriques, Rui Caseiro, Pedro Martins, and Jorge Batista. High-speed tracking with kernelized correlation filters. *TPAMI*, 37(3):583–596, 2015. 6

[21] Ronghang Hu, Huazhe Xu, Marcus Rohrbach, Jiashi Feng, Kate Saenko, and Trevor Darrell. Natural language object retrieval. In *CVPR*, 2016. 2

[22] Lianghua Huang, Xin Zhao, and Kaiqi Huang. Got-10k: A large high-diversity benchmark for generic object tracking in the wild. *arXiv:1810.11981*, 2018. 2, 3

[23] Xu Jia, Huchuan Lu, and Ming-Hsuan Yang. Visual tracking via adaptive structural local sparse appearance model. In *CVPR*, 2012. 6

[24] Zdenek Kalal, Krystian Mikolajczyk, and Jiri Matas. Tracking-learning-detection. *TPAMI*, 34(7):1409–1422, 2012. 6

[25] Matej Kristan, Jiri Matas, Aleš Leonardis, Tomáš Vojíř, Roman Pflugfelder, Gustavo Fernandez, Georg Nebehay, Fatih Porikli, and Luka Čehovin. A novel performance evaluation methodology for single-target trackers. *TPAMI*, 38(11):2137–2155, 2016. 2, 3

[26] Matej Kristan et al. The visual object tracking vot2014 challenge results. In *ECCVW*, 2014. 1, 2

[27] Matej Kristan et al. The visual object tracking vot2017 challenge results. In *ICCVW*, 2017. 1, 2

[28] Annan Li, Min Lin, Yi Wu, Ming-Hsuan Yang, and Shuicheng Yan. Nus-pro: A new visual tracking challenge. *TPAMI*, 38(2):335–349, 2016. 1, 2, 3

[29] Feng Li, Cheng Tian, Wangmeng Zuo, Lei Zhang, and Ming-Hsuan Yang. Learning spatial-temporal regularized correlation filters for visual tracking. In *CVPR*, 2018. 6

[30] Peixia Li, Dong Wang, Lijun Wang, and Huchuan Lu. Deep visual tracking: Review and experimental comparison. *PR*, 76:323–338, 2018. 2

[31] Shuang Li, Tong Xiao, Hongsheng Li, Bolei Zhou, Dayu Yue, and Xiaogang Wang. Person search with natural language description. In *CVPR*, 2017. 2

[32] Xi Li, Weiming Hu, Chunhua Shen, Zhongfei Zhang, Anthony Dick, and Anton Van Den Hengel. A survey of appearance models in visual object tracking. *ACM TIST*, 4(4):58, 2013. 1, 2

[33] Yang Li and Jianke Zhu. A scale adaptive kernel correlation filter tracker with feature integration. In *ECCVW*, 2014. 6

[34] Zhenyang Li, Ran Tao, Efstratios Gavves, Cees GM Snoek, Arnold WM Smeulders, et al. Tracking by natural language specification. In *CVPR*, 2017. 2

[35] Pengpeng Liang, Erik Blasch, and Haibin Ling. Encoding color information for visual tracking: Algorithms and benchmark. *TIP*, 24(12):5630–5644, 2015. 1, 2, 3, 5

[36] Alan Lukezic, Tomas Vojir, Luka Cehovin Zajc, Jiri Matas, and Matej Kristan. Discriminative correlation filter with channel and spatial reliability. In *CVPR*, 2017. 6

[37] Chao Ma, Jia-Bin Huang, Xiaokang Yang, and Ming-Hsuan Yang. Hierarchical convolutional features for visual tracking. In *ICCV*, 2015. 6

[38] Chao Ma, Xiaokang Yang, Chongyang Zhang, and Ming-Hsuan Yang. Long-term correlation tracking. In *CVPR*, 2015. 6

[39] Matthias Mueller, Neil Smith, and Bernard Ghanem. A benchmark and simulator for uav tracking. In *ECCV*, 2016. 1, 2, 3

[40] Matthias Mueller, Neil Smith, and Bernard Ghanem. Context-aware correlation filter tracking. In *CVPR*, 2017. 6

[41] Matthias Müller, Adel Bibi, Silvio Giancola, Salman Al-Subaihi, and Bernard Ghanem. Trackingnet: A large-scale dataset and benchmark for object tracking in the wild. In *ECCV*, 2018. 2, 3, 6

[42] Hyeonseob Nam and Bohyung Han. Learning multi-domain convolutional neural networks for visual tracking. In *CVPR*, 2016. 6, 8

[43] Esteban Real, Jonathon Shlens, Stefano Mazzocchi, Xin Pan, and Vincent Vanhoucke. Youtube-boundingboxes: A large high-precision human-annotated data set for object detection in video. In *CVPR*, 2017. 1, 3

[44] David A Ross, Jongwoo Lim, Ruei-Sung Lin, and Ming-Hsuan Yang. Incremental learning for robust visual tracking. *IJCV*, 77(1-3):125–141, 2008. 6

[45] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *IJCV*, 115(3):211–252, 2015. 1, 8

[46] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *ICLR*, 2015. 1

[47] Arnold WM Smeulders, Dung M Chu, Rita Cucchiara, Simone Calderara, Afshin Dehghan, and Mubarak Shah. Visual tracking: An experimental survey. *TPAMI*, 36(7):1442–1468, 2014. 1, 2, 3

[48] Yibing Song, Chao Ma, Xiaohe Wu, Lijun Gong, Linchao Bao, Wangmeng Zuo, Chunhua Shen, Rynson Lau, and Ming-Hsuan Yang. Vital: Visual tracking via adversarial learning. In *CVPR*, 2018. 6

[49] Ran Tao, Efstratios Gavves, and Arnold WM Smeulders. Siamese instance search for tracking. In *CVPR*, 2016. 6

[50] Jack Valmadre, Luca Bertinetto, João Henriques, Andrea Vedaldi, and Philip HS Torr. End-to-end representation learning for correlation filter based tracking. In *CVPR*, 2017. 6

[51] Jack Valmadre, Luca Bertinetto, João F Henriques, Ran Tao, Andrea Vedaldi, Arnold Smeulders, Philip Torr, and Efstratios Gavves. Long-term tracking in the wild: A benchmark. In *ECCV*, 2018. 2, 3

[52] Yi Wu, Jongwoo Lim, and Ming-Hsuan Yang. Online object tracking: A benchmark. In *CVPR*, 2013. 1, 2, 3, 8

[53] Yi Wu, Jongwoo Lim, and Ming-Hsuan Yang. Object tracking benchmark. *TPAMI*, 37(9):1834–1848, 2015. 1, 2, 3, 5, 6, 8

[54] Alper Yilmaz, Omar Javed, and Mubarak Shah. Object tracking: A survey. *ACM CSUR*, 38(4):13, 2006. 1, 2

[55] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? In *NIPS*, 2014. 1

[56] Jianming Zhang, Shugao Ma, and Stan Sclaroff. Meem: robust tracking via multiple experts using entropy minimization. In *ECCV*, 2014. 6

[57] Kaihua Zhang, Lei Zhang, Qingshan Liu, David Zhang, and Ming-Hsuan Yang. Fast visual tracking via dense spatio-temporal context learning. In *ECCV*, 2014. 6

[58] Kaihua Zhang, Lei Zhang, and Ming-Hsuan Yang. Real-time compressive tracking. In *ECCV*, 2012. 6

[59] Yunhua Zhang, Lijun Wang, Jinqing Qi, Dong Wang, Mengyang Feng, and Huchuan Lu. Structured siamese network for real-time visual tracking. In *ECCV*, 2018. 6