LSOTB-TIR: A Large-Scale High-Diversity Thermal Infrared Object Tracking Benchmark

Qiao Liu
Xin Li
Harbin Institute of Technology, Shenzhen
liuqiao.hit@gmail.com

Zhenyu He*
Harbin Institute of Technology, Shenzhen
Peng Cheng Laboratory
zhenyuhe@hit.edu.cn

Chenglong Li
Anhui University
chenglongli@ahu.edu.cn

Jun Li
Zikun Zhou
Di Yuan
Harbin Institute of Technology, Shenzhen

Jing Li
Kai Yang
Nana Fan
Harbin Institute of Technology, Shenzhen

Feng Zheng
Southern University of Science and Technology
zfeng02@gmail.com

ABSTRACT

In this paper, we present a Large-Scale and high-diversity general Thermal InfraRed (TIR) Object Tracking Benchmark, called LSOTB-TIR, which consists of an evaluation dataset and a training dataset with a total of 1,400 TIR sequences and more than 600K frames. We annotate the bounding box of objects in every frame of all sequences and generate over 730K bounding boxes in total. To the best of our knowledge, LSOTB-TIR is the largest and most diverse TIR object tracking benchmark to date. To evaluate a tracker on different attributes, we define 4 scenario attributes and 12 challenge attributes in the evaluation dataset. By releasing LSOTB-TIR, we encourage the community to develop deep learning based TIR trackers and evaluate them fairly and comprehensively. We evaluate and analyze more than 30 trackers on LSOTB-TIR to provide a series of baselines, and the results show that deep trackers achieve promising performance. Furthermore, we re-train several representative deep trackers on LSOTB-TIR, and their results demonstrate that the proposed training dataset significantly improves the performance of deep TIR trackers. Codes and dataset are available at https://github.com/QiaoLiuHit/LSOTB-TIR.

CCS CONCEPTS

• **Computing methodologies** → **Tracking**; **Image representations**.

KEYWORDS

thermal infrared dataset, thermal infrared object tracking, deep representation learning

*Corresponding author.

MM '20, October 12–16, 2020, Seattle, WA, USA

ACM ISBN 978-1-4503-7988-5/20/10…$15.00
https://doi.org/10.1145/3394171.3413922

ACM Reference Format:
Qiao Liu, Xin Li, Zhenyu He, Chenglong Li, Jun Li, Zikun Zhou, Di Yuan, Jing Li, Kai Yang, Nana Fan, and Feng Zheng. 2020. LSOTB-TIR: A Large-Scale High-Diversity Thermal Infrared Object Tracking Benchmark. In *Proceedings of the 28th ACM International Conference on Multimedia (MM '20), October 12–16, 2020, Seattle, WA, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3394171.3413922

1 INTRODUCTION

TIR object tracking is an important task in artificial intelligence. Given an initial position of a TIR object in the first frame, TIR object tracking is to locate the object in the rest of the frames of a sequence. With the popularization of civilian thermal imaging devices, TIR object tracking receives more and more attention as a crucial intelligent vision technology. It is widely used in video surveillance, maritime rescue, and driver assistance at night [18] since it can track the object in total darkness. In the past several years, some TIR object tracking methods [20, 23, 34, 39, 60, 61] are proposed. Despite much progress, TIR object tracking faces many unsolved problems, such as distractor, intensity variation, and thermal crossover [37].

Evaluating a tracker fairly and comprehensively on a benchmark is crucial to the development of TIR object tracking. However, currently widely used TIR object tracking benchmarks, e.g., LTIR [3], VOT-TIR16 [17], and PTB-TIR [37] suffer from the following drawbacks that make them less effective in conducting a fair and comprehensive evaluation. First, their scale is too small to make an effective evaluation because a tracker can easily overfit to a small dataset using the parameter fine-tuning. Second, they have too few kinds of objects, *e.g.*, PTB-TIR only contains pedestrian objects, which cannot provide an evaluation on general TIR objects. Third, they only have a few tracking scenarios and challenges, which does not meet the requirements of real-world applications. Therefore, it is imperative to build a larger and more diverse TIR object tracking benchmark.

Recently, motivated by the success of deep learning in most visual tasks, several attempts [20, 34, 38, 39] incorporate deep feature models for TIR object tracking and achieve some success. However, the used deep feature models are learned from RGB images, and we find by experiments that these RGB based deep feature models

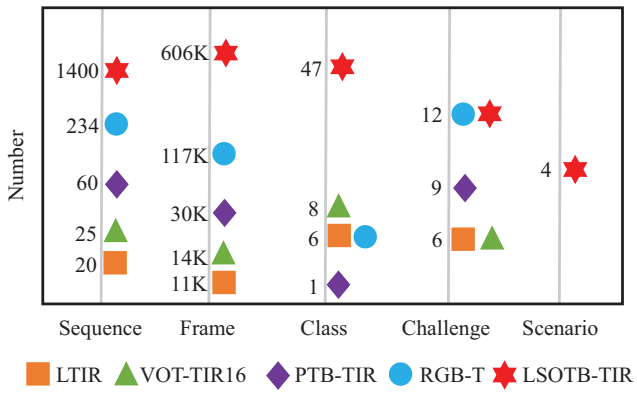


Figure 1: Comparison of currently widely used TIR object tracking benchmarks, including LTIR [3], VOT-TIR16 [17], PTB-TIR [37], RGB-T [31], and LSOTB-TIR. Among these benchmarks, LSOTB-TIR is the largest and most diverse. Noting that only LSOTB-TIR contains the scenario attribute.

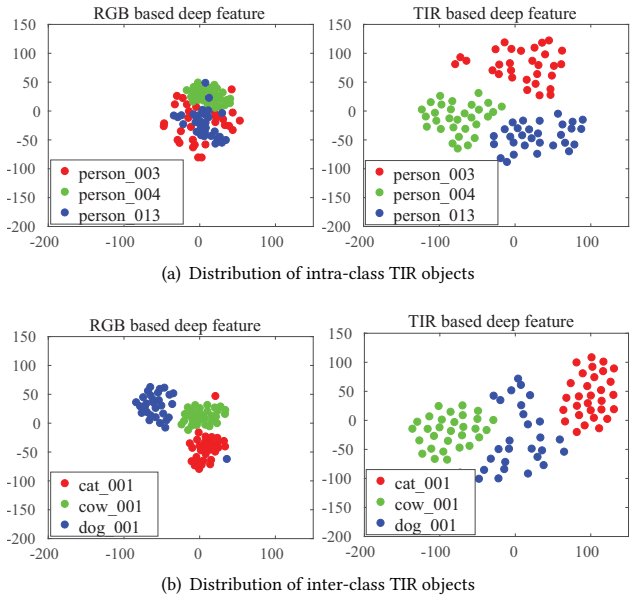


Figure 2: Comparison of the RGB based deep feature and the TIR based deep feature using the t-SNE visualized method [42]. The RGB and TIR based deep features are extracted from the backbone network of two CFNets [52], which are trained on an RGB dataset (VID [47]) and LSOTB-TIR, respectively. We randomly choose 30 objects from each sequence in LSOTB-TIR. (a) All objects belong to the person class but come from different sequences. We can see that the TIR based deep feature can recognize the differences between intra-class TIR objects, which is curial to distinguish distractors in TIR tracking. (b) All objects belong to different classes. The TIR based deep feature can separate inter-class objects more effectively than the RGB based deep feature. Noting that each point in the figure denotes an object.

are less effective in representing TIR objects, as shown in Fig. 2. Different from RGB images, TIR images do not have color information and lack texture features, hence it is crucial to use fine-grained features, such as local contour and structure, to distinguish objects in TIR images. Furthermore, we find by experiments that learning TIR-specific deep features for representing further promotes the performance of TIR object tracking. However, the lacking of a large-scale TIR dataset for deep model training hinders the advantage of the deep leaning in TIR object tracking.

To address the above-mentioned issues, we develop a large-scale and high-diversity TIR object tracking benchmark, called LSOTB-TIR, which consists of an evaluation dataset and a training dataset with a total of 1,400 TIR sequences and more than 600K frames. We annotate the bounding box of objects in every frame of all sequences and generate more than 730K bounding boxes in total. We carefully select 120 sequences with 22 object classes and more than 82K frames as the evaluation dataset, which is larger and more diverse than existing TIR datasets. To understand the strengths and weaknesses of a tracker on specific attributes, we define 4 scenario attributes and 12 challenge attributes for attribute-based evaluation. The training dataset contains 1,280 sequences with 47 object classes and over 650K bounding boxes, which is used for learning TIR-specific deep features. In addition to the benchmark, we evaluate more than 30 trackers on LSOTB-TIR, and provide a detailed analysis. The results show that deep trackers achieve promising performance. Moreover, we re-train several representative deep trackers using the proposed training dataset, and their results on three benchmarks demonstrate that the proposed TIR training dataset significantly improves the performance of deep TIR trackers.

The contributions of this paper are three-fold:

- We propose a large-scale TIR object tracking benchmark, LSOTB-TIR, with high-quality annotations using a self-designed semi-automatic label tool. LSOTB-TIR is currently the largest and most diverse TIR object tracking benchmark, which consists of the training and evaluation datasets with the richest object classes, scenarios, and challenges.
- We conduct extensive evaluation experiments with more than 30 trackers on LSOTB-TIR and provide a series of comparative analyses. The results show that deep trackers achieve promising results and have the potential to obtain better performance.
- We re-train several representative deep trackers on LSOTB-TIR, and their results on three benchmarks demonstrate that the proposed training dataset significantly improves the performance of deep TIR trackers.

2 RELATED WORK

2.1 TIR trackers

In the past decade, some TIR object tracking algorithms have been proposed to handle various challenges. These algorithms can be

Table 1: Comparison of the proposed LSOTB-TIR benchmark with other TIR object tracking benchmarks.

Benchmarks	Num. of sequences	Max frames	Min frames	Mean frames	Total frames	Frame rates	Object classes	Num. of challenges	Scenario attributes	Training dataset
OSU [12]	6	2,031	601	1,424	8K	30 fps	1	n/a	✗	✗
PDT-ATV [45]	8	775	77	486	4K	20 fps	3	n/a	✗	✗
BU-TIV [58]	16	26,760	150	3,750	60K	30 fps	5	n/a	✗	✗
LTIR [3]	20	1,451	71	563	11K	30 fps	6	6	✗	✗
VOT-TIR16 [17]	25	1,451	71	555	14K	30 fps	8	6	✗	✗
PTB-TIR [37]	60	1,451	50	502	30K	30 fps	1	9	✗	✗
RGB-T [31]	234	4,000	45	500	117K	30 fps	6	12	✗	✗
LSOTB-TIR (train.)	1,280	3,056	47	410	524K	30 fps	47	1	✓	✓
LSOTB-TIR (eval.)	120	2,110	105	684	82K	30 fps	22	12		

roughly divided into two categories: conventional TIR trackers and deep TIR trackers.

Conventional TIR trackers. Conventional TIR object tracking methods usually combine a conventional machine learning method with a hand-crafted feature for handling various challenges. To adapt the appearance variation of the object, Venkataraman *et al.* [53] propose to online learn a robust intensity histogram based appearance model using adaptive Kalman filtering, while TBOOST [24] maintains a dynamic MOSSE filter [7] set using a continuously switching mechanism according to appearance variation. Demir *et al.* [13] use a part-based matching method, which integrates the co-difference feature of multiple parts to overcome the partial deformation. To obtain more effective representations, DSLT [60] combines gradient histograms with motion features and then is used in a Structural Support Vector Machine (SSVM) [26] for TIR object tracking. Observing that TIR images do not have color information and lack sharp edges, Berg *et al.* [4] propose a distribution field representation [48] based matching algorithm for TIR object tracking. Despite much progress, these trackers are limited by the hand-crafted feature representation.

Deep TIR trackers. Inspired by the success of deep learning in visual tracking, several works introduce the Convolution Neural Network (CNN) to improve the performance of TIR trackers. These methods can be roughly divided into two categories, deep feature based TIR trackers and matching based deep TIR trackers. Deep feature based TIR trackers often use a pre-trained CNN for feature extraction and then integrates the deep feature into conventional tracking frameworks. For example, Gundogdu *et al.* [23] train a classification network on a small TIR dataset to extract the deep feature of the TIR object and then combine it with the DSST tracker [10] for TIR object tracking. MCFTS [39] combines multiple convolutional features of VGGNet [49] with the Correlation Filters (CFs) [27] to construct an ensemble TIR tracker. Gao *et al.* [20] combine deep appearance features [49] and deep motion features [22] with SSVM for TIR object tracking. ECO-stir [61] trains a Siamese network on synthetic TIR images to extract TIR features and then integrates them into the ECO [9] tracker. Li *et al.* [33] propose a mask sparse representation deep appearance model with the particle filter framework for TIR object tracking. Matching based deep TIR trackers cast the tracking as a matching problem and usually off-line train a matching network for online tracking. For example, Li *et al.* [34] train a spatial variation aware matching network by introducing a spatial attention mechanism for TIR object tracking. Liu *et al.* [38] propose a multi-level similarity based matching network using a semantic similarity module and a complementary structural similarity module for TIR object tracking. However, most of these deep models are learned from RGB images, which do not learn specific patterns of TIR images and hence are less effective in representing TIR objects.

2.2 TIR object tracking benchmarks

To evaluate TIR trackers, there are several widely used benchmarks [3, 12, 16, 17, 31, 37, 45, 58]. In the following, we introduce these benchmarks briefly.

OSU. OSU [12] is a TIR and RGB image fusion dataset, which can be used for TIR object tracking. This dataset contains only pedestrian objects and all 6 videos are captured by a low-resolution TIR camera in a static background.

PDT-ATV. PDT-ATV [45] is a simulative aerial TIR object tracking and detection dataset, which contains 8 sequences captured from a low frame rate and low-resolution TIR camera. The dataset does not have attribute labels and the tracking objects are dim and small.

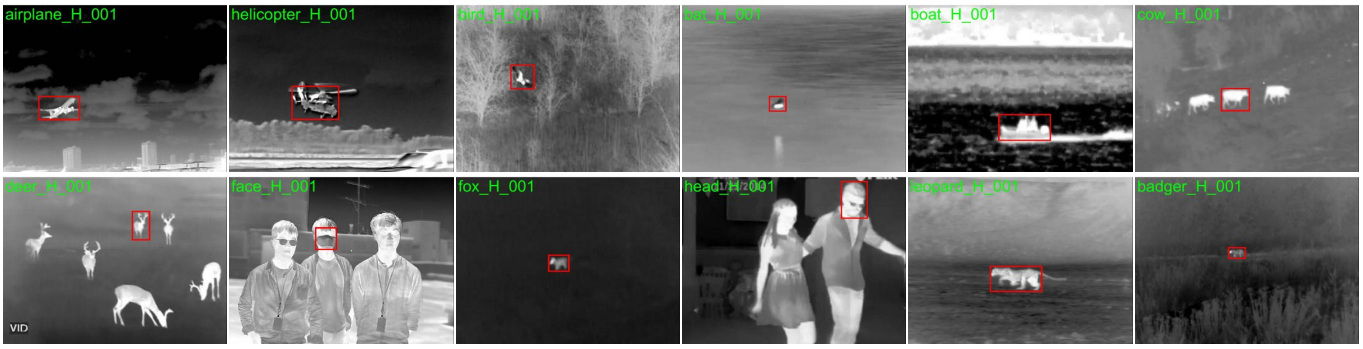
BU-TIV. BU-TIV [58] is used for several TIR visual tasks, including object tracking, counting, and group motion estimation. This dataset contains 16 sequences and more than 60K frames with a high resolution.

LTIR. LTIR [3] is the first standard TIR object tracking benchmark which contains 20 sequences with 6 object classes and an evaluation toolkit. This benchmark is adopted by a TIR object tracking competition, VOT-TIR15 [16].

VOT-TIR16. VOT-TIR16 [17] is extended from VOT-TIR15. It contains 25 sequences with 8 object classes and is more challenging than VOT-TIR15. It has 6 challenge subsets that can be used to evaluate a tracker on specific attributes.

PTB-TIR. PTB-TIR [37] focuses on TIR pedestrian tracking, which contains 60 sequences and 9 attribute challenges. The dataset is collected from a variety of devices but lacks the division of different scenario attributes.

RGB-T. RGB-T [31] is a multi-modal tracking benchmark that contains RGB and TIR videos on the same scene simultaneously. The TIR videos are captured from a single low-resolution TIR camera and can also be used for TIR object tracking.



(b) Hand-held scenario subset with 35 sequences and 16 object classes.

(c) Drone-mounted scenario subset with 25 sequences and 8 object classes.

(d) Vehicle-mounted scenario subset with 20 sequences and 4 object classes.

(e) Video surveillance scenario subset with 40 sequences and 8 object classes.

Figure 3: Examples of the proposed evaluation subset of LSOTB-TIR. It contains 4 non-overlapped scenario subsets with a total of 120 sequences and 22 object classes.

Table 2: Definition of 4 scenarios and 12 challenges on LSOTB-TIR.

Scenario	Definition	Scenario	Definition
VS	The videos come from a surveillance camera.	DS	The videos are captured from a drone-mounted camera.
HH	The videos are shotted from a hand-held camera.	VM	The videos come from a vehicle-mounted camera.
Challenge	**Definition**	**Challenge**	**Definition**
TC	Two same intensity targets cross each other.	IV	The target intensity is changed during tracking.
DIS	Existing the intra-class object near the target.	BC	The background has a similar appearance to the target.
DEF	The target is deformable during tracking.	OCC	The target is partly or fully occluded during tracking.
OV	The target partly or fully leaves the image.	SC	The ratio of the target size is out of the range [0.5, 2].
FM	The target moves more than 20 pixels.	MB	The target is blurry due to the target or camera motion.
LR	The target size is lower than 800 pixels.	ARV	The target aspect ratio is out of the range [0.5, 2].

Although these benchmarks are widely used, they suffer from several problems, such as the small-scale, limited object classes, scenarios and challenges, and lack of training dataset. To solve these issues, we present a large-scale and high-diversity TIR object tracking benchmark, LSOTB-TIR, consisting of an evaluation dataset and

a training dataset. The evaluation dataset contains 120 sequences with more than 82K frames, 22 classes, 4 scenario subsets, and 12 challenge subsets, which is more diverse than these benchmarks. The training dataset contains 1,280 sequences with more than 650K bounding boxes and 47 object classes. Table 1 compares the proposed benchmark with existing TIR object tracking benchmarks. More comparisons with existing RGB tracking benchmarks are shown in the **supplementary material**.

3 PROPOSED LSOTB-TIR BENCHMARK

In this section, we describe details of the proposed TIR object tracking benchmark, LSOTB-TIR. We first introduce TIR videos collection and processing in Section 3.1 and then we show how to annotate the sequence in Section 3.2. Finally, we define attributes of a sequence in Section 3.3.

3.1 Data collection and processing

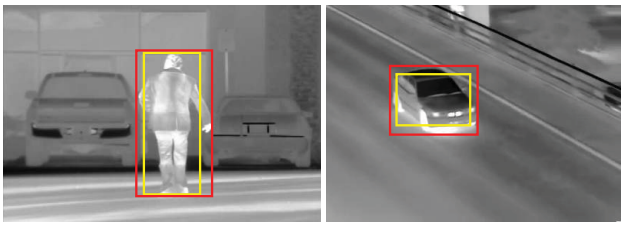
Our goal is to provide a large-scale and high-diversity general TIR object tracking benchmark with the real-world scenarios and challenges. To this end, we determine to track 5 kinds of moving objects of interest (*i.e.*, person, animal, vehicle, aircraft, and boat) in 4 kinds of scenarios (*i.e.*, hand-held, drone-mounted, vehicle-mounted, and video surveillance, as shown in Table 2). Unlike RGB object tracking, which is interested in arbitrary objects, TIR object tracking is usually interested in objects with prominent thermal radiation as the 5 categories mentioned above. The selected 5 categories of objects cover most of the targets of interest of TIR object tracking in the civilian field.

After determining the object classes and scenarios, we first search TIR videos on the Youtube website. Unfortunately, unlike RGB videos, TIR videos are limited on Youtube. We try our best to obtain 600 TIR videos and each of them is within ten minutes. Since we use TIR videos with the white-hot style for tracking, we filter out some videos of other palette styles (*e.g.*, iron, rainbow, and cyan). We then convert the rest videos into image sequences and choose the fragments according to the following principles. First, the object must be active, which is caused by the movement of itself or the camera. Second, the time that the object is fully occluded or out of image range does not exceed one second. Third, the length of the fragment does not exceed 4K frames. In addition, we choose 150 TIR sequences from existing datasets, such as BU-TIV [58], and RGB-T [31]. Finally, we get 1,400 sequences that contain more than 600K frames and 47 object sub-classes of interest.

After getting all the sequences, we split them into a training dataset and an evaluation dataset. We first choose 200 sequences, each of them contains at least one tracking challenging factor, for evaluation. Then, we evaluate all the trackers using these sequences and then select the most difficult 120 sequences as the final evaluation subset according to the difficulty degree of each sequence, which is computed from the average success score of all the evaluated trackers.

3.2 High-quality annotation

We first decide an object of each frame contains 4 kinds of local annotation information, including object class, position, occlusion, and identity. We use a straight minimum bounding box to record the position of an object. When an object is occluded above 50% or out of the image above 50%, we define this occlusion as true. This attribute can be used to exclude the obvious noise and is useful for training deep models.



(a) Deformation (b) Scale variation

Figure 4: Examples of the bounding box adjustment. The yellow bounding boxes are generated by using a tracking algorithm and the red bounding boxes are manually adjusted.

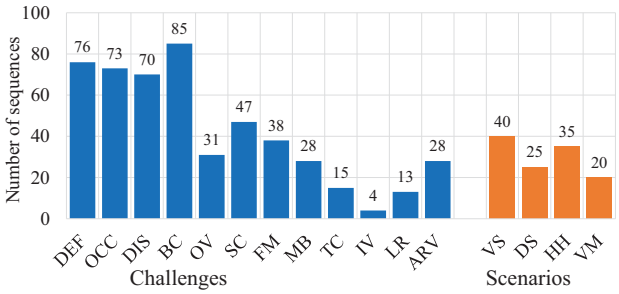


Figure 5: Attributes distribution of the evaluation subet of LSOTB-TIR.

Considering labeling is a time-consuming and labor-intensive task, we design an auxiliary label tool (see the **supplementary material**) based on the ECO-HC [9] tracker. This label tool helps us generate a bounding box of the object in every frame semi-automatically. When the tracker is set to track the object in a short time (*e.g.*, within 10 frames), the generated bounding boxes are accurate in most situations. However, when the object undergoes drastic appearance variation or scale change in a short time, the generated bounding boxes of the label tool are not quite accurate. For these bounding boxes, we adjust them manually, as shown in Fig. 4. We suggest that the label tool makes the annotation more accurate, smoother, and faster than the annotation in each frame manually.

To complete the annotation, we assemble an annotation team comprised of 8 Ph.D. students with careful training. To ensure the quality of the annotations, we verify the annotations frame by frame twice. Eventually, we get 1,400 carefully annotated sequences. Some annotated sequences of the evaluation subset are shown in Fig. 3, and some annotated sequences of the training subset are shown in the **supplementary material**.

Table 3: Comparison of tracking results of more than 30 trackers on LSOTB-TIR. We rank these trackers according to their success score. The property of a tracker includes feature representation (*e.g.*, Deep: deep feature, HoG: histogram of gradient, Cova: covariance feature, CN: color name, Raw: raw pixel), search strategy (*e.g.*, DS: dense search, RS: random search, PF: particle filter), category (*e.g.*, D: discriminative, G: generative), and venue.

Tracker	Performance				Property			
	Success	Precision	Norm. Precision	Speed	Representation	Search	Category	Venue
ECO-TIR (Ours)	① 0.631	① 0.768	① 0.695	18 fps	Deep	DS	D	–
ECO-stir [61]	② 0.616	② 0.750	0.672	13 fps	Deep	DS	D	TIP19
ECO [9]	③ 0.609	0.739	0.670	18 fps	Deep	DS	D	CVPR17
SiamRPN++ [30]	0.604	0.711	0.651	24 fps	Deep	DS	D	CVPR19
MDNet [44]	0.601	② 0.750	② 0.686	1 fps	Deep	RS	D	CVPR16
VITAL [51]	0.597	③ 0.749	③ 0.682	3 fps	Deep	RS	D	CVPR18
ATOM [8]	0.595	0.729	0.647	20 fps	Deep	RS	D	CVPR19
TADT [35]	0.587	0.710	0.635	40 fps	Deep	DS	D	CVPR19
SiamMask [56]	0.579	0.705	0.637	44 fps	Deep	DS	D	CVPR19
ECO-HC [9]	0.561	0.690	0.627	27 fps	HoG	DS	D	CVPR17
SiamFC-TIR (Ours)	0.554	0.700	0.626	45 fps	Deep	DS	D	–
BACF [28]	0.535	0.648	0.591	26 fps	HoG	DS	D	ICCV17
SRDCF [11]	0.529	0.642	0.574	11 fps	HoG	DS	D	ICCV15
UDT [54]	0.523	0.629	0.575	35 fps	Deep	DS	D	CVPR19
MCCT [55]	0.522	0.634	0.574	27 fps	HoG&CN	DS	D	CVPR18
SiamFC [6]	0.517	0.651	0.587	③ 45 fps	Deep	DS	D	ECCVW16
SiamFC-tri [14]	0.513	0.649	0.583	40 fps	Deep	DS	D	ECCV18
CREST [50]	0.504	0.597	0.544	2 fps	Deep	DS	D	ICCV17
Staple [5]	0.492	0.606	0.548	12 fps	HoG&CN	DS	D	CVPR16
MCFTS [39]	0.479	0.635	0.546	4 fps	Deep	DS	D	KBS17
CFNet-TIR (Ours)	0.478	0.580	0.540	24 fps	Deep	DS	D	–
DSST [10]	0.477	0.555	0.505	② 50 fps	HoG	DS	D	BMVC14
MLSSNet [38]	0.459	0.596	0.549	25 fps	Deep	DS	D	arXiv19
CFNet [52]	0.416	0.519	0.481	24 fps	Deep	DS	D	CVPR17
HSSNet [34]	0.409	0.515	0.488	15 fps	Deep	DS	D	KBS19
HCF [40]	0.404	0.536	0.485	14 fps	Deep	DS	D	ICCV15
HDT [46]	0.403	0.538	0.478	6 fps	Deep	DS	D	CVPR16
TGPR [19]	0.403	0.514	0.495	1 fps	Cova	DS	D	ECCV14
RPT [36]	0.388	0.475	0.427	6 fps	HoG	PF	D	CVPR15
Struck [26]	0.384	0.477	0.432	17 fps	Haar	DS	D	TPAMI15
DSiam [25]	0.380	0.451	0.393	12 fps	Deep	DS	D	ICCV17
L1APG [2]	0.371	0.446	0.424	2 fps	Raw	PF	G	CVPR12
LCT [41]	0.364	0.471	0.430	27 fps	HoG&Raw	DS	D	CVPR15
ASLA [59]	0.338	0.429	0.393	3 fps	Raw	PF	G	CVPR12
KCF [27]	0.321	0.418	0.385	① 272 fps	HoG	DS	D	TPAMI15
MIL [1]	0.309	0.378	0.336	19 fps	Haar	DS	D	CVPR09

3.3 Attribute definition

In addition to the local attribute of each frame, we define two kinds of global attributes of a sequence in the evaluation dataset, namely scenario and challenge. The corresponding attribute subsets can be used to further evaluate a tracker on specific attributes. For the scenario attribute, we define 4 scenarios according to the TIR camera platform, including video surveillance, drone-mounted, hand-held, and vehicle-mounted, as shown in Table 2. These scenario subsets can help us understand the strengths and weaknesses of a tracker on specific application scenarios. For the challenge attribute, we define 12 challenges according to the real-world challenging factors in TIR videos. For example, Thermal Crossover (TC) is defined as that two TIR objects with the same intensity cross each other and then lose their contour partly or fully. Distractor (DIS) is defined as that the background near the target exists intra-class objects, which disturbs the tracker to recognize the tracking target. This challenge is a frequent and serious problem in TIR object tracking. Intensity Variation (IV) is defined as that the intensity of the target is changed due to its temperature variation or the brightness variation of the TIR camera. This challenge is an unfrequent issue since the temperature of the target is stable in a short time. Some other challenges, such as Background Clutter (BC), Deformation (DEF), Occlusion (OCC), Out of View (OV), Scale Change (SC), Fast Motion (FM), Motion Blur (MB), Low Resolution (LR), and Aspect Ratio Variation (ARV), are defined in Table 2, and the distribution of each challenge is shown in Fig. 5.

4 EXPERIMENTS

4.1 Evaluation criteria

We use two widely used evaluation criteria in visual tracking, *i.e.*, Center Location Error (CLE) and Overlap Ratio (OR), as the base metrics [57]. Base on these two metrics, precision, normalized precision, and success under One Pass Evaluation (OPE) are computed to measure the overall performance of a tracker.

Precision. CLE is the Euclidean distance between the center location of the predicted position and the ground-truth. Precision denotes the percentage of the successful frame whose CLE is within a given threshold (*e.g.*, 20 pixels).

Normalized precision. Since the precision is sensitive to the resolution of the image and the size of the bounding box, we normalize the precision over the size of the ground-truth bounding box as that in TrackingNet [43] and LaSOT [15]. We then use the Area Under Curve (AUC) of the normalized precision between 0 and 0.5 to rank the trackers.

Success. OR is the overlap rate between the predicted bounding box and the ground-truth. Success denotes the percentage of the successful frame whose OR is larger than a given threshold. We use a dynamic threshold [0 1], and the corresponding AUC is used to rank the tracking algorithms.

Speed. We use the average frame rate of a tracker on the dataset as the speed metric. We run all the trackers on the same PCs with an i7 4.0GHZ CPU and a GTX-1080 GPU.

4.2 Overall performance evaluation

Evaluated trackers. We choose publicly available 33 TIR and RGB tracking methods for evaluation. These methods include sparse trackers, such as L1APG [2], and ASLA [59]; correlation filter trackers, such as KCF [27], DSST [10], SRDCF [11], BACF [28], Staple [5], ECO-HC [9], and MCCT [55]; other hand-crafted feature based trackers, including MIL [1], TGPR [19], LCT [41], Struck [26], and RPT [36]; deep feature based correlation filter trackers, such as HDT [46], ECO [9], HCF [40], DeepSTRCF [32], MCFTS [39], CREST [50], and ECO-stir [61]; matching based deep trackers, including SiamFC [6], CFNet [52], DSiam [25], SiamFC-tri [14], TADT [35], SiamRPN++ [30], SiamMask [56], UDT [54], HSSNet [34], and MLSSNet [38]; classification based deep trackers, such as MDNet [44], VITAL [51], and ATOM [8]. We do not change the parameters of these trackers provided by the authors in the experiment. Furthermore, we re-train several deep trackers on the proposed TIR training dataset for evaluation, such as ECO-TIR, SiamFC-TIR, and CFNet-TIR. We use the backbone network of SiamFC-TIR as the feature extractor in the ECO-TIR tracker.

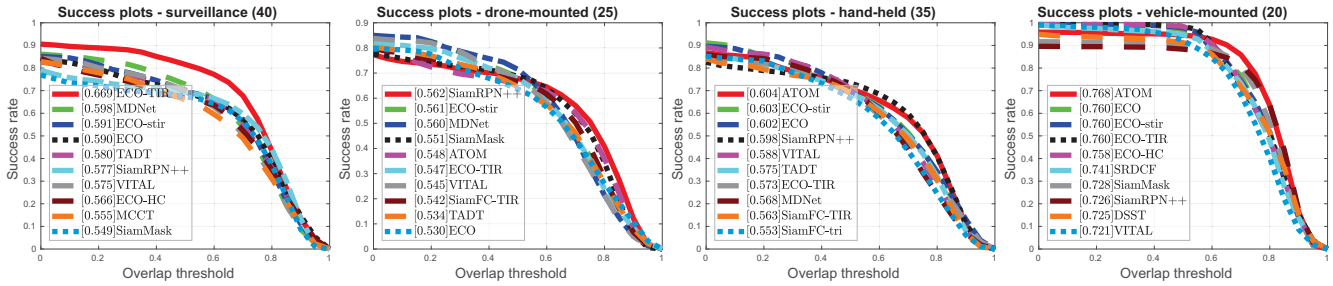
Results and analysis. Table 3 shows the overall performance and property of all the evaluated trackers. Almost top 10 trackers are the deep feature based methods. This shows that the deep feature is superior to the hand-crafted feature and deep trackers achieve promising performance in TIR object tracking. ECO-TIR obtains the best success score (0.631) and precision score (0.768). While ECO-stir [61] using the synthetic TIR based deep feature obtains the second-best success score (0.616) and precision score (0.750). Compared with ECO [9] which obtains the third-best success score (0.609) using the pre-trained RGB based deep feature, ECO-TIR and ECO-stir gain the success score by 2.2% and 0.7%, respectively. This shows that the TIR based deep feature is superior to the RGB based deep feature in TIR object tracking.

Matching based deep trackers, such as SiamRPN++ [30], TADT [35], SiamMask [56], and SiamFC [6], achieve comparable performance while running at a real-time speed. These trackers are usually offline trained to learn a matching network from a large-scale RGB dataset end-to-end. Their favorable results demonstrate that the RGB based deep feature models of these trackers can represent TIR objects since there are some common patterns between RGB objects and TIR objects. However, we argue that the RGB based deep feature has less discriminative capacity in representing TIR objects because the RGB based deep feature often tends to focus on the texture feature [21]. Unlike RGB images, TIR images do not have color information and lack rich texture features. We suggest that the contour and structure features are critical for recognizing TIR objects. For example, compared with SiamFC [6], our trained SiamFC-TIR achieves a 3.7% success gain and a 3.9% normalized precision gain. This demonstrates that the learned TIR based deep feature has better discriminative capacity than the RGB based deep feature for distinguishing TIR objects.

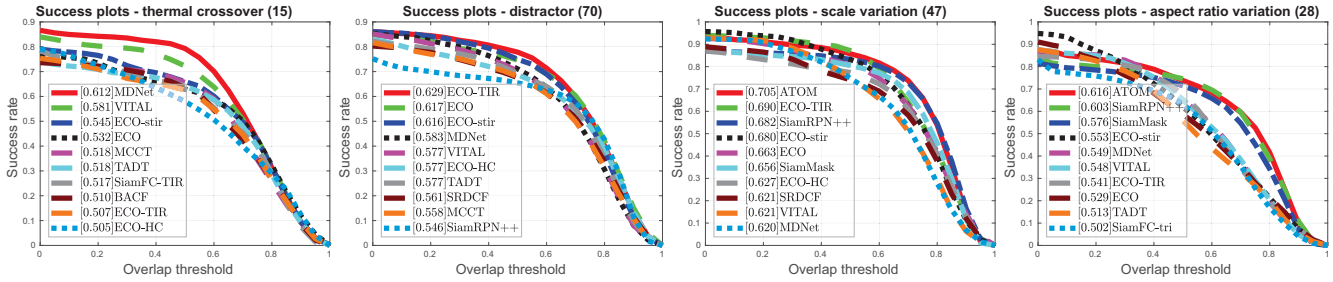
Classification based deep trackers, including MDNet [44], VITAL [51], and ATOM [8], obtain favorable precision, which comes from the powerful discriminative capacity of the online learned binary classifier using the positive and negative samples of the tracked target. This is important for a tracker to adapt the appearance variation of the tracked target. Unfortunately, online training severely hampers their speed and easily leads to the over-fitting problem. However, for TIR trackers, we argue that online training is critical for more robust tracking. Because the online training can obtain a more powerful classifier to recognize subtle differences between the tracked object and distractors.

4.3 Attribute-based evaluation

To understand the strengths and weaknesses of a tracker on special attributes, we evaluate all the trackers on the defined attribute subsets. Fig. 6(a) shows the success scores of the top 10 trackers on the 4 scenario attribute subsets. We notice that the ranking of these trackers is quite different between different scenario subsets. For example, ECO-TIR is higher than MDNet [44] by 7.1% on the surveillance subset, while it is lower than MDNet by 1.3% on the drone subset. This shows that a tracker can not perform well on all the scenarios. Furthermore, we find that the difficulty of the vehicle-mounted scenario subset is the smallest and all the top 10 trackers achieve good performance. The major reason is that vehicle-mounted scenario subset contains fewer challenges and the limited tracked object classes. These scenario subsets can help us develop specific-scenario based trackers to meet the requirements of real-world applications. Fig. 6(b) shows the success scores of the top 10 trackers on 4 challenge attribute subsets. MDNet gets the best success score (0.612) on the thermal crossover subset, which is higher than ECO-TIR by 10.5%. While ECO-TIR achieves the best success scores (0.629 and 0.621) on the distractor and background clutter subsets, which are higher than MDNet by 4.6% and 2.7% respectively. This shows that a tracker can not handle all the challenges. We attribute the good performance of ECO-TIR to the learned TIR based deep feature. ATOM [8] obtains the best success



(a) Scenario attribute subset evaluation.



(b) Challenge attribute subset evaluation.

Figure 6: Attribute-based evaluation results. Only the top 10 trackers are shown for clarity.

Table 4: Comparison of 4 deep trackers trained on an RGB dataset and the proposed TIR training dataset, respectively. "-TIR" denotes this tracker is trained on the proposed TIR training dataset.

Tracker	VOT-TIR16 [17]	PTB-TIR [37]		LSOTB-TIR (Ours)	
	EAO	Pre.	Suc.	Pre.	Suc.
SiamFC [6]	0.225	0.623	0.480	0.651	0.517
SiamFC-TIR	**0.250**	**0.758**	**0.566**	**0.700**	**0.554**
CFNet [52]	0.254	0.629	0.449	0.519	0.416
CFNet-TIR	**0.289**	**0.726**	**0.530**	**0.580**	**0.478**
HSSNet [34]	0.262	0.689	0.468	0.515	0.409
HSSNet-TIR	**0.271**	**0.723**	**0.490**	**0.566**	**0.435**
ECO [9]	0.267	0.838	0.633	0.739	0.609
ECO-TIR	**0.290**	**0.858**	**0.650**	**0.768**	**0.631**

score on the scale variation and aspect ratio variation subsets. This is because ATOM equips an overlap prediction network that can obtain a more accurate bounding box of the target. More attribute-based results are shown in the **supplementary material**.

4.4 Training dataset validation

To validate that the proposed TIR training dataset can boost the performance of deep TIR trackers, we re-train 4 representative deep trackers using the proposed TIR training dataset. Then, we compare them with the original trackers on 3 TIR object tracking benchmarks, including VOT-TIR16 [17], PTB-TIR [37], and LSOTB-TIR. Table 4 shows that all the re-trained deep trackers achieve better performance. Despite CFNet [52] trained on a larger RGB dataset (VID [47]), which is 4 times larger than the proposed TIR training dataset, CFNet-TIR obtains an 8.1% success gain on PTB-TIR and a 3.5% Expected Average Overlap (EAO [29]) gain on VOT-TIR16. While SiamFC-TIR achieves a 13.5% precision gain on PTB-TIR and a 3.7% success gain on LSOTB-TIR. ECO-TIR not only gets the best performance on LSOTB-TIR but also obtains, compared with ECO, a 2.3% EAO gain on VOT-TIR16 and a 1.7% success gain on PTB-TIR. These results demonstrate that the proposed TIR training dataset significantly improves the performance of deep TIR trackers.

5 CONCLUSIONS

In this paper, we propose a large-scale and high-diversity general thermal infrared object tracking benchmark with high-quality annotations, called LSOTB-TIR. To the best of our knowledge, LSOTB-TIR is the largest and most diverse TIR object tracking benchmark to date. By releasing LSOTB-TIR, we help the community develop deep learning based TIR trackers and evaluate them fairly and comprehensively. We conduct extensive evaluation experiments with more than 30 trackers on LSOTB-TIR. The results show that the deep trackers achieve promising performance. Furthermore, we re-train several deep trackers using the proposed training dataset, and their results demonstrate that the proposed training dataset significantly boosts the performance of deep TIR trackers. We suggest that learning TIR-specific deep feature for improving TIR object tracking is one of the main ways in the future.

6 ACKNOWLEDGMENT

This study is supported by the National Natural Science Foundation of China (Grant No.61672183), by the Shenzhen Research Council (Grant No.JCYJ20170413104552269 46, JCYJ20170815113552036).

REFERENCES

[1] B. Babenko, Ming Hsuan Yang, and S. Belongie. 2009. Visual tracking with online Multiple Instance Learning. In *IEEE Conference on Computer Vision and Pattern Recognition*.

[2] Chenglong Bao, Yi Wu, Haibin Ling, and Hui Ji. 2012. Real time robust L1 tracker using accelerated proximal gradient approach. In *IEEE Conference on Computer Vision and Pattern Recognition*.

[3] A. Berg, J. Ahlberg, and M. Felsberg. 2015. A thermal Object Tracking benchmark. In *IEEE International Conference on Advanced Video and Signal Based Surveillance*.

[4] Amanda Berg, Jorgen Ahlberg, and Michael Felsberg. 2016. Channel coded distribution field tracking for thermal infrared imagery. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*.

[5] Luca Bertinetto, Jack Valmadre, Stuart Golodetz, Ondrej Miksik, and Philip HS Torr. 2016. Staple: Complementary learners for real-time tracking. In *IEEE Conference on Computer Vision and Pattern Recognition*.

[6] Luca Bertinetto, Jack Valmadre, Joao F Henriques, Andrea Vedaldi, and Philip HS Torr. 2016. Fully-Convolutional Siamese Networks for Object Tracking. In *European Conference on Computer Vision Workshops*.

[7] David S Bolme, J Ross Beveridge, Bruce A Draper, and Yui Man Lui. 2010. Visual object tracking using adaptive correlation filters. In *IEEE Conference on Computer Vision and Pattern Recognition*.

[8] Martin Danelljan, Goutam Bhat, Fahad Shahbaz Khan, and Michael Felsberg. 2019. Atom: Accurate tracking by overlap maximization. In *IEEE Conference on Computer Vision and Pattern Recognition*.

[9] Martin Danelljan, Goutam Bhat, Fahad Shahbaz Khan, and Michael Felsberg. 2017. ECO: efficient convolution operators for tracking. In *IEEE Conference on Computer Vision and Pattern Recognition*.

[10] Martin Danelljan, Gustav Häger, Fahad Khan, and Michael Felsberg. 2014. Accurate scale estimation for robust visual tracking. In *British Machine Vision Conference*.

[11] Martin Danelljan, Gustav Hager, Fahad Shahbaz Khan, and Michael Felsberg. 2015. Learning spatially regularized correlation filters for visual tracking. In *IEEE International Conference on Computer Vision*.

[12] James W Davis and Vinay Sharma. 2007. Background-subtraction using contour-based fusion of thermal and visible imagery. *Computer Vision and Image Understanding* 106, 2 (2007), 162–182.

[13] Huseyin Seckin Demir and Omer Faruk Adil. 2018. Part-Based Co-Difference Object Tracking Algorithm for Infrared Videos. In *International Conference on Image Processing*.

[14] Xingping Dong and Jianbing Shen. 2018. Triplet Loss in Siamese Network for Object Tracking. In *European Conference on Computer Vision*.

[15] Heng Fan, Liting Lin, Fan Yang, Peng Chu, Ge Deng, Sijia Yu, Hexin Bai, Yong Xu, Chunyuan Liao, and Haibin Ling. 2019. Lasot: A high-quality benchmark for large-scale single object tracking. In *IEEE Conference on Computer Vision and Pattern Recognition*.

[16] Michael Felsberg, Amanda Berg, Gustav Hager, Jorgen Ahlberg, et al. 2015. The thermal infrared visual object tracking VOT-TIR2015 challenge results. In *IEEE International Conference on Computer Vision Workshops*.

[17] Michael Felsberg, Matej Kristan, Jiři Matas, Aleš Leonardis, et al. 2016. The Thermal Infrared Visual Object Tracking VOT-TIR2016 Challenge Results. In *European Conference on Computer Vision Workshops*.

[18] Rikke Gade and Thomas B Moeslund. 2014. Thermal cameras and applications: a survey. *Machine vision and applications* 25, 1 (2014), 245–262.

[19] Jin Gao, Haibin Ling, Weiming Hu, and Junliang Xing. 2014. Transfer learning based visual tracking with gaussian processes regression. In *European Conference on Computer Vision*.

[20] Peng Gao, Yipeng Ma, Ke Song, Chao Li, Fei Wang, and Liyi Xiao. 2018. Large margin structured convolution operator for thermal infrared object tracking. In *International Conference on Pattern Recognition*.

[21] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. 2098. ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *International Conference on Learning Representations*.

[22] Georgia Gkioxari and Jitendra Malik. 2015. Finding action tubes. In *IEEE Conference on Computer Vision and Pattern Recognition*.

[23] Erhan Gundogdu, Aykut Koc, Berkan Solmaz, et al. 2016. Evaluation of feature channels for correlation-filter-based visual object tracking in infrared spectrum. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*.

[24] Erhan Gundogdu, Huseyin Ozkan, H. Seckin Demir, et al. 2015. Comparison of infrared and visible imagery for object tracking: Toward trackers with superior IR performance. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*.

[25] Qing Guo, Wei Feng, Ce Zhou, et al. 2017. Learning dynamic siamese network for visual object tracking. In *IEEE International Conference on Computer Vision*.

[26] Sam Hare, Stuart Golodetz, Amir Saffari, Vibhav Vineet, et al. 2015. Struck: Structured output tracking with kernels. *IEEE Transactions of Pattern Analysis and Machine Intelligence* 38, 10 (2015), 2096–2109.

[27] João F Henriques, Rui Caseiro, Pedro Martins, and Jorge Batista. 2015. High-speed tracking with kernelized correlation filters. *IEEE Transactions of Pattern Analysis and Machine Intelligence* 37, 3 (2015), 583–596.

[28] Hamed Kiani Galoogahi, Ashton Fagg, and Simon Lucey. 2017. Learning background-aware correlation filters for visual tracking. In *IEEE International Conference on Computer Vision*.

[29] M. Kristan, A. Leonardis, J. Matas, M. Felsberg, et al. 2017. The Visual Object Tracking VOT2017 Challenge Results. In *IEEE International Conference on Computer Vision Workshops*.

[30] Bo Li, Wei Wu, Qiang Wang, Fangyi Zhang, Junliang Xing, and Junjie Yan. 2019. Siamrpn++: Evolution of siamese visual tracking with very deep networks. In *IEEE Conference on Computer Vision and Pattern Recognition*.

[31] Chenglong Li, Xinyan Liang, Yijuan Lu, Nan Zhao, and Jin Tang. 2019. RGB-T object tracking: benchmark and baseline. *Pattern Recognition* 96 (2019), 106977.

[32] Feng Li, Cheng Tian, Wangmeng Zuo, Lei Zhang, and Ming-Hsuan Yang. 2018. Learning spatial-temporal regularized correlation filters for visual tracking. In *IEEE Conference on Computer Vision and Pattern Recognition*.

[33] Meihui Li, Lingbing Peng, Yingpin Chen, et al. 2019. Mask Sparse Representation Based on Semantic Features for Thermal Infrared Target Tracking. *Remote Sensing* 11, 17 (2019), 1967.

[34] Xin Li, Qiao Liu, Nana Fan, et al. 2019. Hierarchical spatial-aware Siamese network for thermal infrared object tracking. *Knowledge-Based Systems* 166 (2019), 71–81.

[35] Xin Li, Chao Ma, Baoyuan Wu, Zhenyu He, and Ming-Hsuan Yang. 2019. Target-Aware Deep Tracking. In *IEEE Conference on Computer Vision and Pattern Recognition*.

[36] Yang Li, Jianke Zhu, and Steven CH Hoi. 2015. Reliable patch trackers: Robust visual tracking by exploiting reliable patches. In *IEEE Conference on Computer Vision and Pattern Recognition*.

[37] Qiao Liu, Zhenyu He, Xin Li, and Yuan Zheng. 2019. PTB-TIR: A Thermal Infrared Pedestrian Tracking Benchmark. *IEEE Transaction on Multimedia* 22, 3 (2019), 666–675.

[38] Qiao Liu, Xin Li, Zhenyu He, Nana Fan, Di Yuan, and Hongpeng Wang. 2020. Learning Deep Multi-Level Similarity for Thermal Infrared Object Tracking. *IEEE Transaction on Multimedia* (2020). https://doi.org/10.1109/TMM.2020.3008028

[39] Qiao Liu, Xiaohuan Lu, Zhenyu He, et al. 2017. Deep convolutional neural networks for thermal infrared object tracking. *Knowledge-Based Systems* 134 (2017), 189–198.

[40] Chao Ma, Jia-Bin Huang, Xiaokang Yang, and Ming-Hsuan Yang. 2015. Hierarchical convolutional features for visual tracking. In *IEEE International Conference on Computer Vision*.

[41] Chao Ma, Xiaokang Yang, Chongyang Zhang, and Ming Hsuan Yang. 2015. Long-term correlation tracking. In *IEEE Conference on Computer Vision and Pattern Recognition*.

[42] Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of machine learning research* 9, Nov (2008), 2579–2605.

[43] Matthias Muller, Adel Bibi, Silvio Giancola, Salman Alsubaihi, and Bernard Ghanem. 2018. Trackingnet: A large-scale dataset and benchmark for object tracking in the wild. In *European Conference on Computer Vision*.

[44] Hyeonseob Nam and Bohyung Han. 2016. Learning multi-domain convolutional neural networks for visual tracking. In *IEEE Conference on Computer Vision and Pattern Recognition*.

[45] Jan Portmann, Simon Lynen, Margarita Chli, and Roland Siegwart. 2014. People detection and tracking from aerial thermal views. In *IEEE Robotics and Automation Society*.

[46] Yuankai Qi, Shengping Zhang, Lei Qin, et al. 2016. Hedged Deep Tracking. In *IEEE Conference on Computer Vision and Pattern Recognition*.

[47] Olga Russakovsky, Jia Deng, Hao Su, et al. 2015. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision* 115, 3 (2015), 211–252.

[48] Laura Sevilla-Lara and Erik Learned-Miller. 2012. Distribution fields for tracking. In *IEEE Conference on Computer Vision and Pattern Recognition*.

[49] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).

[50] Yibing Song, Chao Ma, Lijun Gong, et al. 2017. Crest: Convolutional residual learning for visual tracking. In *IEEE International Conference on Computer Vision*.

[51] Yibing Song, Chao Ma, Xiaohe Wu, Lijun Gong, et al. 2018. Vital: Visual tracking via adversarial learning. In *IEEE Conference on Computer Vision and Pattern Recognition*.

[52] Jack Valmadre, Luca Bertinetto, João Henriques, Andrea Vedaldi, and Philip HS Torr. 2017. End-to-end representation learning for correlation filter based tracking. In *IEEE Conference on Computer Vision and Pattern Recognition*.

[53] Vijay Venkataraman, Guoliang Fan, Joseph P Havlicek, et al. 2012. Adaptive kalman filtering for histogram-based appearance learning in infrared imagery. *IEEE Transactions on Image Processing* 21, 11 (2012), 4622–4635.

[54] Ning Wang, Yibing Song, Chao Ma, Wengang Zhou, Wei Liu, and Houqiang Li. 2019. Unsupervised Deep Tracking. In *IEEE Conference on Computer Vision and Pattern Recognition*.

[55] Ning Wang, Wengang Zhou, Qi Tian, Richang Hong, Meng Wang, and Houqiang Li. 2018. Multi-cue correlation filters for robust visual tracking. In *IEEE Conference on Computer Vision and Pattern Recognition*.

[56] Qiang Wang, Li Zhang, Luca Bertinetto, Weiming Hu, and Philip HS Torr. 2019. Fast online object tracking and segmentation: A unifying approach. In *IEEE Conference on Computer Vision and Pattern Recognition*.

[57] Yi Wu, Jongwoo Lim, and Ming-Hsuan Yang. 2013. Online object tracking: A benchmark. In *IEEE Conference on Computer Vision and Pattern Recognition*.

[58] Zheng Wu, Nathan Fuller, Diane Theriault, and Margrit Betke. 2014. A thermal infrared video benchmark for visual analysis. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*.

[59] Jia Xu, Huchuan Lu, and Ming Hsuan Yang. 2012. Visual tracking via adaptive structural local sparse appearance model. In *IEEE Conference on Computer Vision and Pattern Recognition*.

[60] Xianguo Yu, Qifeng Yu, Yang Shang, and Hongliang Zhang. 2017. Dense structural learning for infrared object tracking at 200+ frames per second. *Pattern Recognition Letter* 100 (2017), 152–159.

[61] Lichao Zhang, Abel Gonzalez-Garcia, Joost van de Weijer, Martin Danelljan, and Fahad Shahbaz Khan. 2019. Synthetic data generation for end-to-end thermal infrared tracking. *IEEE Transactions on Image Processing* 28, 4 (2019), 1837–1850.