Towards More Flexible and Accurate Object Tracking with Natural Language: Algorithms and Benchmark
— Supplementary Material —

Xiao Wang[1]*, Xiujun Shu[2,1]*, Zhipeng Zhang[3], Bo Jiang[4], Yaowei Wang[1], Yonghong Tian[1,5], Feng Wu[1,6]
[1]Peng Cheng Laboratory, Shenzhen, China
[2]School of Electronic and Computer Engineering, Peking University, Shenzhen, China
[3]NLPR, Institute of Automation, Chinese Academy of Sciences
[4]School of Computer Science and Technology, Anhui University, Hefei, China
[5]Department of Computer Science and Technology, Peking University, Beijing, China
[6]University of Science and Technology of China, Hefei, China
https://sites.google.com/view/langtrackbenchmark/

1. The TNL2K Benchmark

1.1. Motivation and Protocols

Motivation: Directly extending existing datasets like GOT-10k [19] is an intuitive and good idea for this task, but GOT-10k contains few videos with special properties as mentioned in Fig. 1 in our paper. Also, its videos are all short-term which can't reflect performance gain of re-detection with language. As for LaSOT [14], many of its language annotations can not point out target object clearly, as shown in Fig. 1. Thus, LaSOT is not suitable for tracking-by-language only. Similar views can also be found in GTI [59]. Therefore, we build the TNL2K (from video collection, dense bbox and language annotation, to diverse baseline construction) to better reflect the characteristics (see below) of tracking by natural language. The target of this work is not to construct the largest tracking dataset, but to build the first benchmark specifically designed for tracking-by-language task. Compared with GOT-10k and LaSOT, the data collection of TNL2K is a compromise between length and quantity.

Protocol: When collecting the videos, we attempt to search the target object is *severely occluded in the first frame*, with *significant appearance variation* (e.g., cloth changing for human), *can only be located with reasoning*, which correspond to Fig. 1 in our paper. Also, we collect videos from other thermal tracking datasets and annotate language descriptions only to check the robustness to certain challenging factors like domain adaptation, modality switch, etc.

*The first two authors contribute equally to this work. Yaowei Wang is the corresponding author. Email: {wangx03, shuxj, wangyw, tianyh}@pcl.ac.cn, zhangzhipeng2017@ia.ac.cn, jiangbo@ahu.edu.cn, fengwu@ustc.edu.cn.

1.2. Why add Attribute Modality Switch (MS) ?

In the proposed TNL2K dataset, we design a new attribute termed Modality Switch (MS) for object tracking. This is mainly motivated by the fact that the RGB cameras work well in the daytime but nearly ineffective at night, meanwhile, the thermal cameras work well in the night time. If we track a target for an extremely long-term (e.g., several days or weeks), collaboration between RGB and thermal cameras are needed. Therefore, the connections between the two modalities need to be set up. Similar views can be found in cross-modality person re-identification [50, 51]. There are still no works on object tracking try to build such connections and they usually study these two cameras separately (i.e., RGB tracking [14, 47, 52], Thermal Tracking [33]) or in an integrated approach (i.e., RGB-T tracking [28]). In this work, we propose the modality switch and attempt to encourage researches on such cross-modality object tracking.

1.3. Highlights of TNL2K Dataset

Generally speaking, our proposed benchmark TNL2K have the following features as shown in Table 1:

- **TNL2K is the first benchmark specifically designed for tracking-by-natural language.** Different from regular tracking benchmarks like OTB, GOT10k, and TrackingNet, we provide both language annotation and dense bounding box annotation for each video sequence which will be a good platform for natural language-related tracking. Different from the recently released long-term tracking dataset LaSOT which also

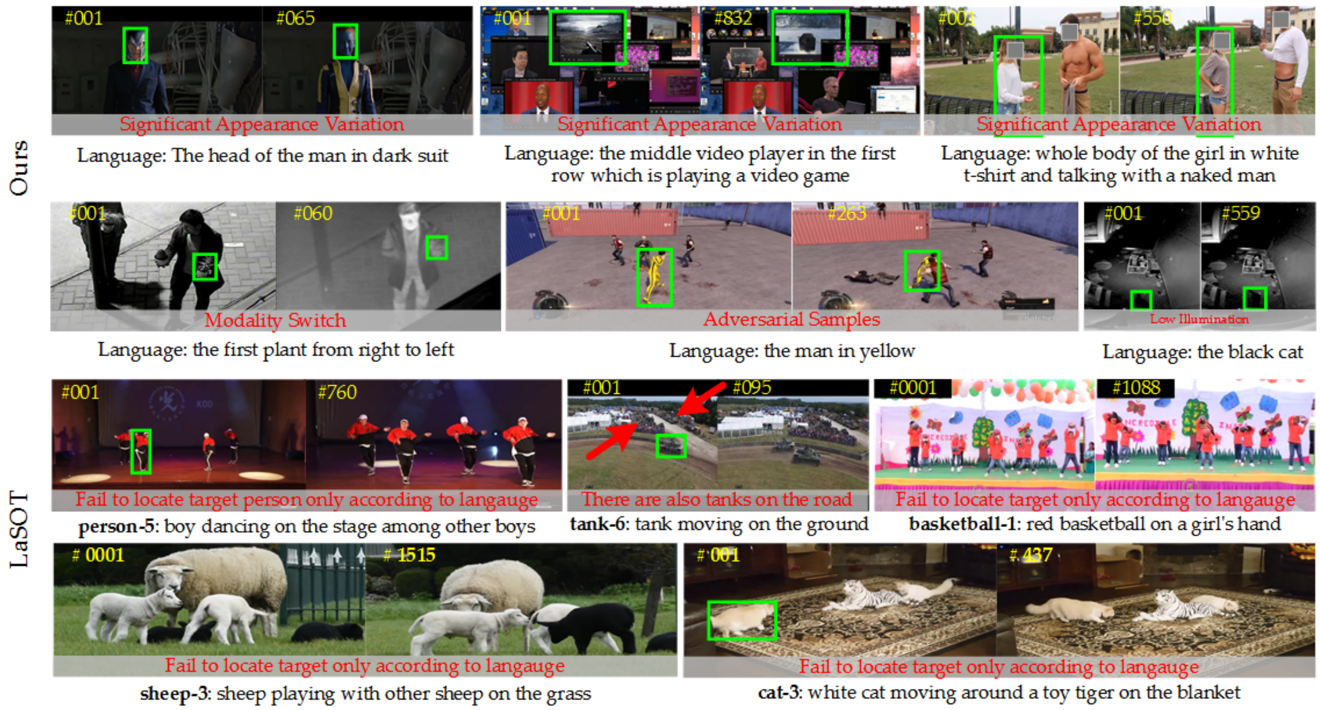

Figure 1. Comparison between our proposed TNL2K dataset and existing LaSOT dataset. Best viewed by zooming in.

Table 1. Comparison of current datasets for object tracking. # denotes the number of corresponding item. Lang-A and Lang-I denotes the dataset can be used for language assisted and initialized tracking task. SAV denotes the dataset contains many videos with significant appearance variation. Adv means the dataset contains adversarial samples (i.e., malicious attacks). DA is short for domain adaptation.

Datasets	#Videos	#Min	#Mean	#Max	#Total	#FR	#Attributes	Aim	Absent	Lang-A	Lang-I	SAV	Adv	DA
OTB50 [52]	51	71	578	3,872	29K	30 fps	11	Eval						
OTB100 [53]	100	71	590	3,872	59K	30 fps	11	Eval						
TC-128 [32]	128	71	429	3,872	55K	30 fps	11	Eval						
VOT-2017 [24]	60	41	356	1,500	21K	30 fps	-	Eval						
NUS-PRO [25]	365	146	371	5040	135K	30 fps	-	Eval						
UAV123 [35]	123	109	915	3085	113K	30 fps	12	Eval						
UAV20L [35]	20	1717	2934	5527	59K	30 fps	12	Eval						
NfS [23]	100	169	3830	20665	383K	240 fps	9	Eval						
TrackingNet [36]	30,643	-	480	-	14.43M	30 fps	15	Train/Eval						
OxUvA [41]	366	900	4260	37440	1.55M	30 fps	6	Train/Eval						
GOT-10k [19]	10,000	29	149	1,418	1.5M	10 fps	6	Train/Eval	✓					
LaSOT [14]	1,400	1000	2506	11397	3.52M	30 fps	14	Train/Eval	✓	✓				
TNL2K (Ours)	2,000	21	622	18488	1.24M	30 fps	17	Train/Eval	✓	✓	✓	✓	✓	✓

provides language annotation, their annotation only describes the attribute of target object, but ignores the spatial position. Therefore, this benchmark can be only used for the task of *tracking by joint language and bbox*. Our language annotations not only embody the attribute, category, shape, properties, and structural relationship with other objects, therefore, our dataset can also be used for the task of *tracking by natural language only*. Some video sequences and corresponding annotations are provided in Figure 1 to give an intuitive understanding of the difference between our TNL2K and LaSOT.

- **TNL2K is the first benchmark to provides videos with actively introduced adversarial samples** which will be beneficial for the development of adversarial training for tracking.

- **TNL2K is the first benchmark to provides videos with significant appearance variation**, such as *cloth/face changing*. We believe our benchmark will greatly boost related research on abrupt appearance variation based tracking.

- **TNL2K provides a heterogeneous dataset** that contains RGB video, Thermal video [1], Cartoon, and Syn-

[1] There are 518 videos totally borrowed from existing RGB-T dataset [28] and infrared tracking dataset [33].

thetic data (i.e., videos from games). It can be used for the study of domain adaptation, e.g., train the tracker on RGB data and test it on Thermal videos.

- **TNL2K provides three kinds of baseline methods for future works to compare**, including Tracking-by-BBox, Tracking-by-Language, Tracking-by-Joint-BBox-Language.

2. The Proposed Method

2.1. YOLO Loss and BCE Loss Functions

In the training phase, we use the YOLO loss function for the optimization of the visual grounding module by following [58]. This loss is first proposed in YOLOv3 [39] which attempt to predict the five quantities of each anchor box by shifting its *center, width, height,* and the *confidence* on this shifted box. To better use it for visual grounding, the following two changes are modified by Yang et al.: 1). recalibrate its anchor boxes; 2). change its sigmoid layer to a softmax function. Due to the object detection is designed for output multiple locations, while visual grounding only needs to predict one bbox which best fit the language description. Therefore, the sigmoid function in YOLOv3 is replaced by softmax function. The cross-entropy is used for the measurement of confidence scores, and the regions with maximum IoU with ground truth are labeled as 1, other regions are set as 0. More details can be found in [39, 58]. For the training of TANet, we adopt Binary Cross-Entropy (BCE) loss to measure the distance between the ground truth mask and the prediction.

2.2. Details of Evaluated Trackers

In this section, we provide the details of evaluated BBox-based trackers on our TNL2K dataset. As shown in Table 2, the publication, feature representation, update or not, need pre-train or not, search scheme, tracking efficiency, and results (Precision Plot and Success Plot) on the TNL2K are all reported. These tracking algorithms are ranked according to the results.

2.3. Introduction to TANet

Inspired by [46, 48, 49], we introduce the TANet for the global search to replace the Grounding module [58] in the setting of *tracking-by-joint language and BBox*, termed Ours-II. Generally speaking, the TANet is inspired by semantic segmentation, which takes the target object and video frames as input and output an attention map using a decoder network. The estimated attention maps can highlight the possible search regions from a global view. Therefore, it can be seen as a kind of global search scheme and can be integrated with the baseline tracker and our proposed AdaSwitcher module for robust and accurate tracking. Our experimental results also demonstrate that we can attain good performance on three used datasets, i.e., the OTB-Lang [31], LaSOT [14], and TNL2K. This will be a strong baseline method for future works to compare on the language guided visual tracking. The implementation of our all networks will be released for other researchers to follow.

3. Experimental Results

3.1. Attribute Analysis

As shown in Figure 2, we provide experimental results of all the defined 17 attributes of our TNL2K dataset. Generally speaking, we can find that the SiamRCNN [42] achieves the best performance on most of the attributes, like *Scale Variation, Rotation, Background Clutter, Partial Occlusion, Adversarial Samples, Deformation, Fast Motion, Out-of-view, Motion Blur, Aspect Ration Change, Illumination Variation, Camera Motion*, and *Viewpoint Change*. Meanwhile, the SuperDiMP [4], LTMU [8], PrDiMP [11] and KYS [15] also attains good performance on these attributes, and the KYS also achieves top-1 results on the *Low Resolution*. These results all demonstrate the strong performance of Siamese network based trackers with the help of pre-training and joint local and global search scheme. Interestingly, we can also find that on the attribute *Thermal Crossover* which are all thermal videos, the MDNet [37] which is an online learned tracker attain the best results. Even the Staple and SRDCF are better than most of the other Siamese trackers, such as SiamKPN, SiamCAR, SiamRPN++, SiamRCNN, KYS, etc. The huge contrast demonstrates that online learning is very important for the tracker which is trained on one domain and tested on another domain (for example, the tracker trained on RGB videos and tested on Thermal videos).

3.2. Efficiency Analysis

In this work, two baseline methods are proposed for the *natural language initialized tracking* (Our-I) and *natural language guided tracking* (Our-II). For Our-I, the overall running efficiency is 24.39 FPS on the OTB-Lang, tested on a laptop with Intel Core I7, RTX2070. For Our-II, the overall efficiency on the OTB-Lang is 12.44 FPS.

3.3. More Visualization

In this section, more visualization on the tracking results is given to better understand our proposed method. As shown in Figure 3, 20 video sequences from OTB-Lang are selected to demonstrate the results of the visual grounding module. From the first three rows, we can find that the grounding module can locate the target object accurately when the background is relatively clean. Also, it works well in some challenge videos, like *car*, and *human head*. For the fourth row, the grounding is not accurate enough for track-

Table 2. Summary of evaluated trackers on TNL2K dataset.

Index	Tracker	Publication	Feature	Update	Pre-train	Search Scheme	FPS	Results
001	SiamRCNN [42]	CVPR-2020	ResNet-101	✗	✓	Local + Global	5@GPU	0.528|0.523
002	SuperDiMP [4]	ICCV-2019	ResNet-50	✓	✓	Local	40@GPU	0.484|0.492
003	LTMU [8]	CVPR-2020	ResNet-50	✓	✓	Local	13@GPU	0.473|0.485
004	PrDiMP50 [11]	CVPR-2020	ResNet-50	✓	✓	Local	30@GPU	0.459|0.470
005	KYS [15]	ECCV-2020	ResNet-50	✓	✓	Local + Global	20@GPU	0.435|0.449
006	DiMP50 [4]	ICCV-2019	ResNet-50	✓	✓	Local	40@GPU	0.434|0.447
007	TACT [6]	ACCV-2020	ResNet-50	✗	✓	Local + Global	42@GPU	0.422|0.438
008	SiamBAN [5]	CVPR-2020	ResNet-50	✗	✓	Local	40@GPU	0.417|0.410
009	SiamRPN++ [26]	CVPR-2019	ResNet-50	✗	✓	Local	35@GPU	0.412|0.413
010	CLNet [13]	ECCV-2020	ResNet-50	✗	✓	Local	45@GPU	0.411|0.408
011	D3S [34]	CVPR-2020	ResNet-50	✗	✓	Local	25@GPU	0.393|0.388
012	ATOM [10]	CVPR-2019	ResNet-50	✗	✓	Local	30@GPU	0.392|0.401
013	SiamKPN [30]	arXiv-2020	ResNet-50	✗	✓	Local	24@GPU	0.389|0.352
014	GlobalTrack [20]	AAAI-2020	ResNet-50	✗	✓	Global	6@GPU	0.386|0.405
015	SiamCAR [16]	CVPR-2020	ResNet-50	✗	✓	Local	52@GPU	0.384|0.353
016	DeepMTA [46]	TCSVT-2021	ResNet-50	✓	✓	Local + Global	12@CPU	0.381|0.385
017	SiamMask [45]	CVPR-2019	ResNet-50	✗	✓	Local	55@GPU	0.380|0.383
018	Ocean [62]	ECCV-2020	ResNet-50	✗	✓	Local	58@GPU	0.377|0.384
019	MDNet [37]	CVPR-2016	CNN-3	✓	✓	Local	1@GPU	0.371|0.384
020	SiamFC++ [54]	AAAI-2020	GoolgeNet	✓	✓	Local	90@GPU	0.369|0.386
021	VITAL [40]	CVPR-2018	CNN-3	✓	✓	Local	1.5@GPU	0.353|0.366
022	Meta-Tracker [38]	ECCV-2018	CNN-3	✓	✓	Local	1@GPU	0.333|0.339
023	SiamDW [61]	CVPR-2019	Res22W	✗	✓	Local	150@GPU	0.326|0.323
024	RT-MDNet [22]	ECCV-2018	CNN-3	✓	✓	Local	46@GPU	0.322|0.308
025	SPLT [55]	ICCV-2019	ResNet-50	✗	✓	Local + Global	25@GPU	0.321|0.337
026	GradNet [29]	ICCV-2019	CNN-5	✓	✓	Local	80@GPU	0.318|0.317
027	ECO [9]	CVPR-2017	VGG	✓	✗	Local	8@CPU	0.317|0.326
028	MemTracking [56]	ECCV-2018	CNN-5	✓	✓	Local	50@GPU	0.305|0.304
029	MAML [43]	CVPR-2020	ResNet-50	✗	✓	Local	40@GPU	0.295|0.284
030	DaSiamRPN [63]	ECCV-2018	ResNet-50	✗	✓	Local	110@GPU	0.288|0.329
031	FCOT [7]	arXiv-2020	ResNet-50	✗	✓	Local	45@GPU	0.288|0.320
032	SiamFC [3]	ECCVW-2016	CNN-5	✗	✓	Local	58@GPU	0.286|0.295
033	SiamRPN [27]	CVPR-2018	ResNet-50	✗	✓	Local	160@GPU	0.281|0.300
034	ADNet [60]	CVPR-2017	CNN-3	✓	✓	Local	3@GPU	0.278|0.285
035	UDT [44]	CVPR-2019	CNN-5	✗	✓	Local	70@GPU	0.271|0.266
036	Staple [2]	CVPR-2016	HOG	✓	✗	Local	80@CPU	0.270|0.270
037	SRDCF [12]	ICCV-2015	HOG	✓	✗	Local	6@CPU	0.269|0.265
038	GOTURN [17]	ECCV-2016	CaffeNet-5	✗	✓	Local	100@GPU	0.205|0.198
039	RTAA [21]	ECCV-2018	ResNet-50	✗	✓	Local	2.2@GPU	0.193|0.217
040	KCF [18]	TPAMI-2015	HOG	✓	✗	Local	172@CPU	0.153|0.200
041	VisGround [58]	ICCV-2019	DarkNet-53	✗	✓	Global	147@GPU	0.143|0.159
042	ROAM [57]	CVPR-2020	ResNet-50	✗	✓	Local	13@GPU	0.108|0.157
043	MIL [1]	CVPR-2009	HOG	✓	✗	Local	25@CPU	0.063|0.042

ing, including the central location and scale. We can find that the performance of visual grounding is needed to be further improved for more accurate tracking. More experimental results of our proposed baseline and other trackers can be found in Figure 4.

Acknowledgement

This work is jointly supported by Key-Area Research and Development Program of Guangdong Province 2019B010155002, Postdoctoral Innovative Talent Support Program BX20200174, China Postdoctoral Science Foundation Funded Project 2020M682828, National Natural Science Foundation of China (61976002, 61825101), National Key Research and Development Program of China 2020AAA0106800.

We appreciate all the anonymous reviewers and AC for their effects on improving the quality of this work. We also thank all the staff for the annotation of TNL2K, including Rui Yang, Bin Li, Xiao Yuan, Wenjing Wang, Weiwen Wu, Zihuan Huang, and Zixing Xu, et al.

References

[1] Boris Babenko, Ming-Hsuan Yang, and Serge Belongie. Visual tracking with online multiple instance learning. In *2009 IEEE Conference on computer vision and Pattern Recognition*, pages 983–990. IEEE, 2009.

[2] Luca Bertinetto, Jack Valmadre, Stuart Golodetz, Ondrej Miksik, and Philip HS Torr. Staple: Complementary learners for real-time tracking. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1401–1409, 2016.

[3] Luca Bertinetto, Jack Valmadre, Joao F Henriques, Andrea Vedaldi, and Philip HS Torr. Fully-convolutional siamese networks for object tracking. In *European conference on computer vision*, pages 850–865. Springer, 2016.

[4] Goutam Bhat, Martin Danelljan, Luc Van Gool, and Radu Timofte. Learning discriminative model prediction for tracking. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 6182–6191, 2019.

[5] Zedu Chen, Bineng Zhong, Guorong Li, Shengping Zhang, and Rongrong Ji. Siamese box adaptive network for visual tracking. In *Proceedings of the IEEE/CVF Conference*

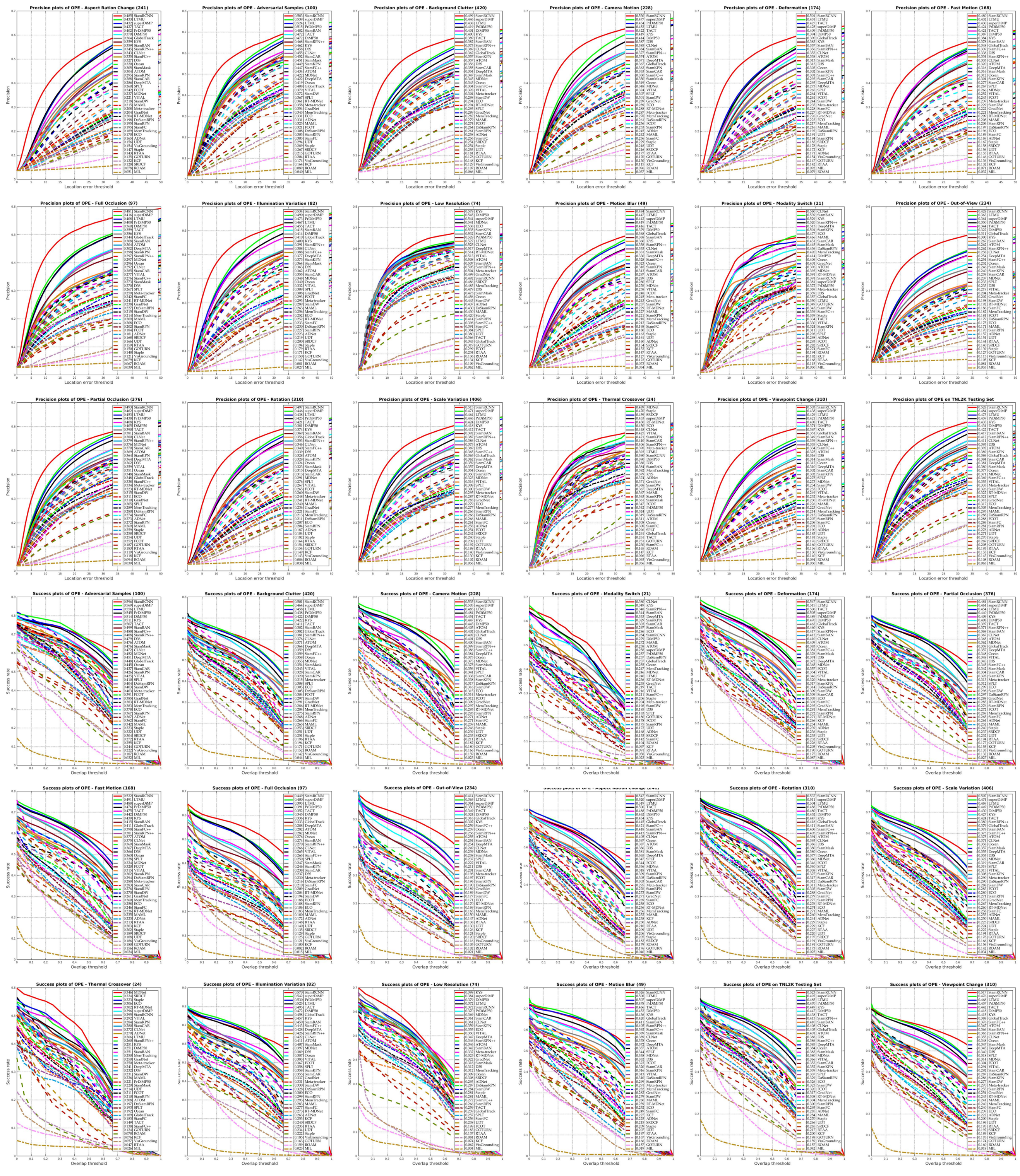

Figure 2. Tracking results under each challenging factors on TNL2K dataset (Tracking-by-BBox). Best viewed by zooming in.

on Computer Vision and Pattern Recognition, pages 6668–6677, 2020.

[6] Janghoon Choi, Junseok Kwon, and Kyoung Mu Lee. Visual tracking by tridentalign and context embedding. In *Proceed-*

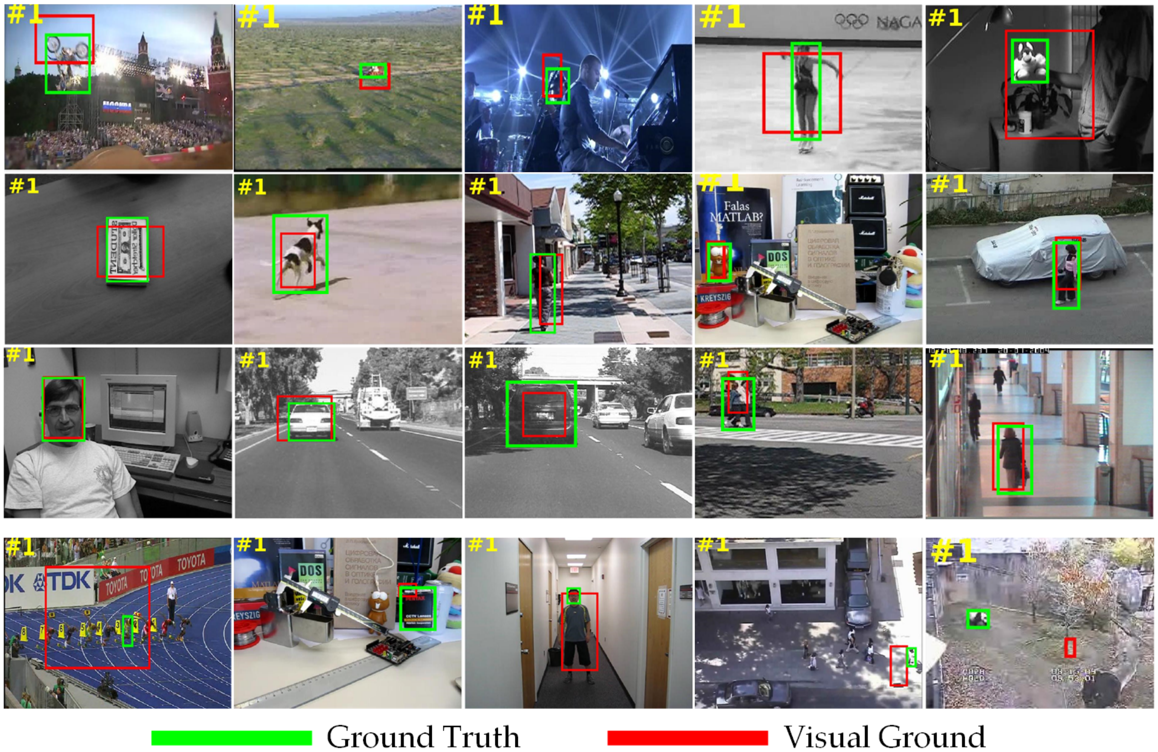

Figure 3. Results of the first frame of visual grounding module.

Figure 4. Tracking results of our method and other state-of-the-art tracking algorithms.

ings of the Asian Conference on Computer Vision, 2020.

[7] Yutao Cui, Cheng Jiang, Limin Wang, and Gangshan Wu. Fully convolutional online tracking. *arXiv preprint arXiv:2004.07109*, 2020.

[8] Kenan Dai, Yunhua Zhang, Dong Wang, Jianhua Li, Huchuan Lu, and Xiaoyun Yang. High-performance long-term tracking with meta-updater. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6298–6307, 2020.

[9] Martin Danelljan, Goutam Bhat, Fahad Shahbaz Khan, and Michael Felsberg. Eco: Efficient convolution operators for tracking. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

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

[11] Martin Danelljan, Luc Van Gool, and Radu Timofte. Probabilistic regression for visual tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7183–7192, 2020.

[12] Martin Danelljan, Gustav Hager, Fahad Shahbaz Khan, and Michael Felsberg. Learning spatially regularized correlation filters for visual tracking. In *Proceedings of the IEEE international conference on computer vision*, pages 4310–4318, 2015.

[13] Xingping Dong, Jianbing Shen, Ling Shao, and Fatih Porikli. Clnet: A compact latent network for fast adjusting siamese trackers. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

[14] Heng Fan, Liting Lin, Fan Yang, Peng Chu, Ge Deng, Sijia Yu, Hexin Bai, Yong Xu, Chunyuan Liao, and Haibin Ling. Lasot: A high-quality benchmark for large-scale single object tracking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5374–5383, 2019.

[15] Bhat Goutam, Danelljan Martin, Van Gool Luc, and Timofte Radu. Know your surroundings: Exploiting scene information for object tracking. In *Proceedings of the European Conference on Computer Vision*, 2020.

[16] Dongyan Guo, Jun Wang, Ying Cui, Zhenhua Wang, and Shengyong Chen. Siamcar: Siamese fully convolutional classification and regression for visual tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6269–6277, 2020.

[17] David Held, Sebastian Thrun, and Silvio Savarese. Learning to track at 100 fps with deep regression networks. In *European Conference on Computer Vision*, pages 749–765. Springer, 2016.

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

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

[20] Lianghua Huang, Xin Zhao, and Kaiqi Huang. Globaltrack: A simple and strong baseline for long-term tracking. *AAAI*, 2020.

[21] Shuai Jia, Chao Ma, Yibing Song, and Xiaokang Yang. Robust tracking against adversarial attacks. In *European Conference on Computer Vision*, pages 69–84. Springer, 2020.

[22] Ilchae Jung, Jeany Son, Mooyeol Baek, and Bohyung Han. Real-time mdnet. In *The European Conference on Computer Vision (ECCV)*, September 2018.

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

[24] Matej Kristan, Jiri Matas, Aleš Leonardis, Tomas Vojir, Roman Pflugfelder, Gustavo Fernandez, Georg Nebehay, Fatih Porikli, and Luka Čehovin. A novel performance evaluation methodology for single-target trackers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(11):2137–2155, Nov 2016.

[25] A Li, M Lin, Y Wu, MH Yang, and S Yan. NUS-PRO: A New Visual Tracking Challenge. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(2):335–349, 2016.

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

[27] Bo Li, Junjie Yan, Wei Wu, Zheng Zhu, and Xiaolin Hu. High performance visual tracking with siamese region proposal network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8971–8980, 2018.

[28] Chenglong Li, Xinyan Liang, Yijuan Lu, Nan Zhao, and Jin Tang. Rgb-t object tracking: benchmark and baseline. *Pattern Recognition*, 96:106977, 2019.

[29] Peixia Li, Boyu Chen, Wanli Ouyang, Dong Wang, Xiaoyun Yang, and Huchuan Lu. Gradnet: Gradient-guided network for visual object tracking. In *The IEEE International Conference on Computer Vision (ICCV)*, October 2019.

[30] Qiang Li, Zekui Qin, Wenbo Zhang, and Wen Zheng. Siamese keypoint prediction network for visual object tracking. *arXiv preprint arXiv:2006.04078*, 2020.

[31] Zhenyang Li, Ran Tao, Efstratios Gavves, Cees GM Snoek, and Arnold WM Smeulders. Tracking by natural language specification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6495–6503, 2017.

[32] Pengpeng Liang, Erik Blasch, and Haibin Ling. Encoding color information for visual tracking: Algorithms and benchmark. *IEEE Transactions on Image Processing*, 24(12):5630–5644, 2015.

[33] Qiao Liu, Zhenyu He, Xin Li, and Yuan Zheng. Ptb-tir: A thermal infrared pedestrian tracking benchmark. *IEEE Transactions on Multimedia*, 22(3):666–675, 2019.

[34] Alan Lukezic, Jiri Matas, and Matej Kristan. D3s-a discriminative single shot segmentation tracker. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7133–7142, 2020.

[35] Matthias Mueller, Neil Smith, and Bernard Ghanem. A benchmark and simulator for uav tracking. In *European conference on computer vision*, pages 445–461. Springer, 2016.

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

[37] Hyeonseob Nam and Bohyung Han. Learning multi-domain convolutional neural networks for visual tracking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4293–4302, 2016.

[38] Eunbyung Park and Alexander C. Berg. Meta-tracker: Fast and robust online adaptation for visual object trackers. In *The European Conference on Computer Vision (ECCV)*, September 2018.

[39] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018.

[40] Yibing Song, Chao Ma, Xiaohe Wu, Lijun Gong, Linchao Bao, Wangmeng Zuo, Chunhua Shen, Rynson W.H. Lau, and Ming-Hsuan Yang. Vital: Visual tracking via adversarial learning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

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

[42] Paul Voigtlaender, Jonathon Luiten, Philip HS Torr, and Bastian Leibe. Siam r-cnn: Visual tracking by re-detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6578–6588, 2020.

[43] Guangting Wang, Chong Luo, Xiaoyan Sun, Zhiwei Xiong, and Wenjun Zeng. Tracking by instance detection: A meta-learning approach. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6288–6297, 2020.

[44] Ning Wang, Yibing Song, Chao Ma, Wengang Zhou, Wei Liu, and Houqiang Li. Unsupervised deep tracking. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1308–1317, 2019.

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

[46] Xiao Wang, Zhe Chen, Jin Tang, Bin Luo, Yaowei Wang, Yonghong Tian, and Feng Wu. Dynamic attention guided multi-trajectory analysis for single object tracking. *IEEE Transactions on Circuits and Systems for Video Technology*, 2021.

[47] Xiao Wang, Chenglong Li, Bin Luo, and Jin Tang. Sint++: Robust visual tracking via adversarial positive instance generation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4864–4873, 2018.

[48] Xiao Wang, Chenglong Li, Rui Yang, Tianzhu Zhang, Jin Tang, and Bin Luo. Describe and attend to track: Learning natural language guided structural representation and visual attention for object tracking. *arXiv preprint arXiv:1811.10014*, 2018.

[49] Xiao Wang, Rui Yang, Tao Sun, and Bin Luo. Learning target-aware attention for robust tracking with conditional adversarial network. In *30TH British Machine Vision Conference (BMVC)*, page 131, 2019.

[50] Ancong Wu, Wei-Shi Zheng, Shaogang Gong, and Jianhuang Lai. Rgb-ir person re-identification by cross-modality similarity preservation. *International journal of computer vision*, pages 1–21, 2020.

[51] Ancong Wu, Wei-Shi Zheng, Hong-Xing Yu, Shaogang Gong, and Jianhuang Lai. Rgb-infrared cross-modality person re-identification. In *Proceedings of the IEEE international conference on computer vision*, pages 5380–5389, 2017.

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

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

[54] Yinda Xu, Zeyu Wang, Zuoxin Li, Ye Yuan, and Gang Yu. Siamfc++: Towards robust and accurate visual tracking with target estimation guidelines. In *AAAI*, pages 12549–12556, 2020.

[55] Bin Yan, Haojie Zhao, Dong Wang, Huchuan Lu, and Xiaoyun Yang. 'skimming-perusal'tracking: A framework for real-time and robust long-term tracking. *ICCV*, 2019.

[56] Tianyu Yang and Antoni B. Chan. Learning dynamic memory networks for object tracking. In *The European Conference on Computer Vision (ECCV)*, September 2018.

[57] Tianyu Yang, Pengfei Xu, Runbo Hu, Hua Chai, and Antoni B Chan. Roam: Recurrently optimizing tracking model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6718–6727, 2020.

[58] Zhengyuan Yang, Boqing Gong, Liwei Wang, Wenbing Huang, Dong Yu, and Jiebo Luo. A fast and accurate one-stage approach to visual grounding. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4683–4693, 2019.

[59] Zhengyuan Yang, Tushar Kumar, Tianlang Chen, Jingsong Su, and Jiebo Luo. Grounding-tracking-integration. *IEEE Transactions on Circuits and Systems for Video Technology*, 2020.

[60] Sangdoo Yun, Jongwon Choi, Youngjoon Yoo, Kimin Yun, and Young Choi Jin. Action-decision networks for visual tracking with deep reinforcement learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1349–1358, 2017.

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

[62] Zhipeng Zhang, Houwen Peng, Jianlong Fu, Bing Li, and Weiming Hu. Ocean: Object-aware anchor-free tracking. In *European Conference on Computer Vision*, volume 12366, pages 771–787. Springer, 2020.

[63] Zheng Zhu, Qiang Wang, Bo Li, Wei Wu, Junjie Yan, and Weiming Hu. Distractor-aware siamese networks for visual object tracking. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 101–117, 2018.