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

Xiao Wang\textsuperscript{1,*}, Xiujun Shu\textsuperscript{2,1,*}, Zhipeng Zhang\textsuperscript{3}, Bo Jiang\textsuperscript{4}, Yaowei Wang\textsuperscript{1}, Yonghong Tian\textsuperscript{1,5}, Feng Wu\textsuperscript{1,6}  
\textsuperscript{1}Peng Cheng Laboratory, Shenzhen, China  
\textsuperscript{2}School of Electronic and Computer Engineering, Peking University, Shenzhen, China  
\textsuperscript{3}NLPR, Institute of Automation, Chinese Academy of Sciences  
\textsuperscript{4}School of Computer Science and Technology, Anhui University, Hefei, China  
\textsuperscript{5}Department of Computer Science and Technology, Peking University, Beijing, China  
\textsuperscript{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

\textbf{Motivation}: Directly extending existing datasets like GOT-10k \cite{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 \cite{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 \cite{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.

\textbf{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.

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 \cite{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 \cite{14, 47, 52}, Thermal Tracking \cite{33}) or in an integrated approach (i.e., RGB-T tracking \cite{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

---

\textsuperscript{*}The first two authors contribute equally to this work. Yaowei Wang is the corresponding author. Email: \{wang03, shuxj, wangyw, tianyh\}@pcl.ac.cn, zhangzhipeng2017@ia.ac.cn, jiangbo@ahu.edu.cn, fengwu@ustc.edu.cn.
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, Cartoon, and Syn-

---

1There are 518 videos totally borrowed from existing RGB-T dataset [28] and infrared tracking dataset [33].
thctic 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 Ratio 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-
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 the staff for the annotation of TNL2K, including Rui Yang, Bin Li, Xiao Yuan, Wenjing Wang, Weiwen Wu, Zhuan Huang, and Zixing Xu, et al.

References


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

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.


