Online Resource 3

June 2, 2020

1 Experiments

1.1 Datasets and Settings

Lingual OTB2013 and OTB2015. We borrow the same dataset used in [9] and part of videos for training and the rest for testing the results. In the purpose of comparison study with other trackers, we employ the OTB 2013 and OTB 2015 [14] for evaluation. OTB 2013 contains 50 sequences of videos with two targets annotation for *Jogging* sequence, while OTB 2015 has 99 with two targets annotation for *Skating* and Jogging, and we annotates each one of them. Therefore, we get lingual OTB 2013 with 50 sequences and OTB 2015 with 99 sequences. The benchmarks consider the average per-frame success rate at different thresholds. Trackers are then compared regarding area under the curve(AUC) of success rates for one pass evaluation(OPE).

Algorithm (Precision/Success)	**SiamFC**	**SiamRPN**	**DaSiamRPN**	**SiamMask**	**MDNet**
Lingual OTB2013	0.809/0.607	0.884/0.658	0.888/0.656	0.841/0.643	0.948/0.708
Lingual OTB2015	0.771/0.582	0.851/0.637	0.880/0.658	0.840/0.647	0.909/0.678
Speed (FPS)	86	160	160	55	1
Algorithm (Precision/Success)	**ATOM**	**DiMP**	**PrDiMP**	**SiamRPN++**	**Ours**
Lingual OTB2013	0.875/0.859	0.909/0.691	0.909/0.703	0.918/0.691	***0.896/0.715***
Lingual OTB2015	0.879/0.667	0.899/0.686	0.903/0.699	0.915/0.696	***0.911/0.722***
Speed (FPS)	30	43	30	35	8

Table 1: Comparison study on OTB 2013 and OTB 2015

LaSOT [3]. As one of the largest dataset in object tracking, the dataset consists of 1400 sequences with 70 categories and more than 3.5M frames in total, and provides not only bounding boxes for object but also the natural language annotation corresponding to the target. Therefore we choose one video from one category mixed with OTB to train the language-guided module. And different parts of LaSOT is utilized for evaluation.

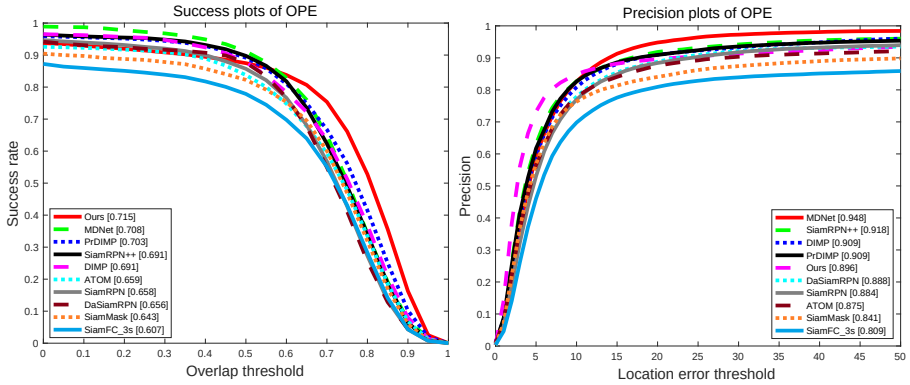

Figure 1: Comparision study on Lingual OTB 2013.Left is the success plots and right is the precision plots.

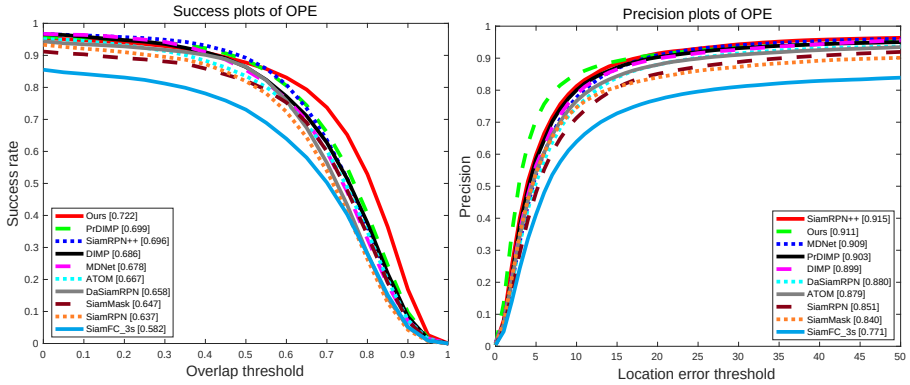

Figure 2: Comparision study on Lingual OTB 2015.Left is the success plots and right is the precision plots.

ReferIt. This dataset is proposed in [6] for object localization and segmentation by natural language experssion, which contains about 20,000 images and 130,525 expressions. Our language model has been pre-trained by this dataset.

1.2 Experimental Results

We test our Langtrack model on 3 popular tracking datasets to demonstrate the effectiveness. As the most popular dataset for evaluation, OTB use success rate and precision rate for evaluation criterion. Based on evaluation tools from OTB benchmark [13], we conduct several experiments to test and evaluate the results of our model and compare them with recent trackers like SiamFC [1], SiamRPN [7], DaSiamRPN [15], MDNet [10], ATOM [2],SiamRPN++ [8],SiamMask [12], DiMP, PrDiMP and Lang-tracker [9]. Our algorithm achieves better success rate than other trackers on OTB 2013 and 2015. Lang-tracker [9] is the first study to do tracking task with language guidance,

which achieves 0.578 success rate on OTB 2013, while we have a much better result. Our method outperforms other trackers on success rate, though the precision rate is not the best on OTB 2013 and 2015 as we show in Fig.1.2. Experiments show that our model has been more robust under the circumstances of complex scenarios according to the success rate and our language-guided module has been enabled the model more efficiently to resist the distractors. IoU-based refinement process help to improve the precision on OTB 2015 compare with other mainstream trackers. In addition, as the recent popular tracker, SiamMask achieves the state-of-the-art performance in terms of tracking accuracy. We employ it as our visual tracking module and also take it as the benchmark for our comparison. The results as we see in Fig.2 show our model improve the tracking accuracy (precision) on lingual OTB 2015 from 0.840 to 0.911, and success rate from 0.647 to 0.722, which justify the effectiveness of our method. We analyze the speed of different trackers on Table.1 and our tracker runs at 8fps. Though it's not real-time compared with other trackers, we think it strikes well balance among the speed, stability and accuracy. The reason that our model runs slower than SiamMask is that our language module to process the language description consumes more computing resources and leads to the latency.

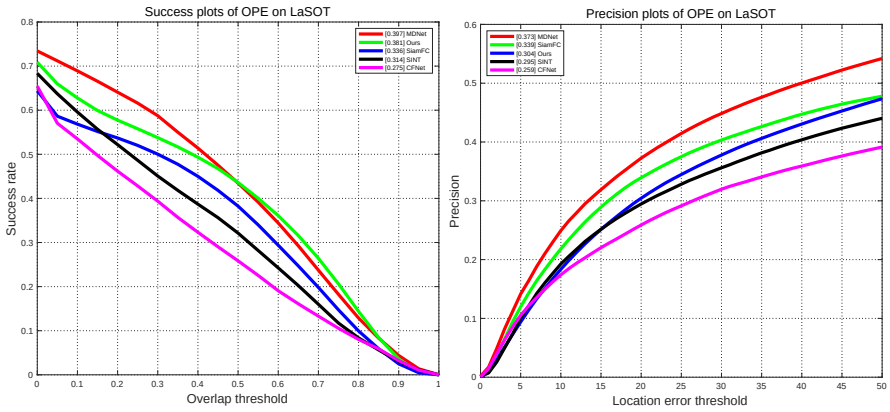

Figure 3: Comparative study on LaSOT.

We also test our model on LaSOT dataset, the only single object tracking dataset with language annotations. As one of the largest tracking dataset, LaSOT provides two protocols to evaluate the trackers. Protocal 1 is to evaluate all the 1400 sequences while 2 to evaluate the 280 sequences of test split. We employ the protocol 2 to evaluate OPE of our LangTrack model. The result in Fig.3 and Table 2 shows that our model achieves the competitive performance, especially on success plot, though it is not the best tracker according to the comparative study. We attribute this to lack of enough training carried out on high-resolution frames. Therefore, its performance on LaSOT which contains high-resolution videos seems to be less outstanding than the one on Lingual OTB.

	Precision plot	Success plot	Speed(fps)
CFNet	0.259	0.275	75
SINT	0.295	0.314	2
SiamFC	0.339	0.336	86
MDNet	0.373	0.397	1
Ours	***0.304***	***0.381***	***8***

Table 2: Evaluation on LaSOT testing set

1.3 Ablation Studies

Different experiments have been carried out to test the effect of various modules and methods on lingual OTB2015. The result has been shown in Table 3.

Algrithm	**Precision**	**Success**
SiamMask	0.840	0.647
No Language	0.701	0.529
No Optical Flow	0.906	0.721
No IoU optimization	0.901	0.709
Ours	0.911	0.722

Table 3: Ablation Study on lingual OTB2015

Language guided Module. We take the language module away, and the precision and success rate of results drop to 0.705 and 0.511 on OTB2013, 0.701 and 0.529 on OTB2015, it performs even worse than the siamMask as shows in Fig.4. It fully validates that the language module plays important role to alleviate the influence of distractors during tracking and improve the discriminative ability of the model. The reason about why the result is below the SiamMask is that the optical flow location confidence is not always stable, the performance drifting leads to worsen the results. It can not be considered as the only location score but the assisting factor with visual grounding factor.

Trackers	**Precision**	**Success**
Ours(Box+Lang)	***0.911***	***0.722***
Qi [4](Box+Lang)	0.79	0.61
QI [5](Box+Lang)	0.73	0.67
Li [9](Box+Lang)	0.72	0.55
Ours(Lang Only)	***0.813***	***0.651***
Qi [4](Lang Only)	0.78	0.54
QI [5](Lang Only)	0.56	0.54
Li [9](Lang Only)	0.29	0.25

Table 4: Comparative Study on Lingual OTB 2015 among Natural Language Trackers

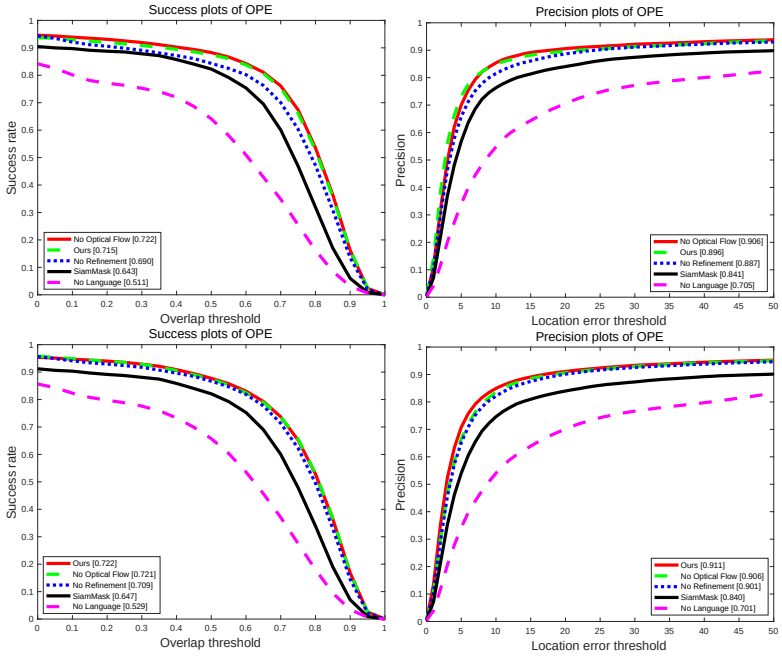

Figure 4: Ablation study on Lingual OTB 2013 and 2015.Left is the success plots and right is the precision plots. No Optical Flow refers there is no Temporal Supervision Module. No Refinement refers there is no Optimization-based Bbox Refinement block. No language means there is no Language-Guided module.

Temporal Supervision Module. According to our experiments, the optical flow branch contributes to the slight improvement on OTB2015, while not better on OTB2013. We consider that this module proves to be more effective in more complex environment than the simple one. Therefore, we won't use it as the only supervisor along with visual tracking module, it should be utilized with our language guided module.

IoU-guided optimization. We take LangTrack as the benchmark with 0.722/0.911 on OTB2015 and 0.715/0.896 on OTB2013.The result show a little drop to 0.709/0.901 on OTB2015 and 0.690/0.887 on OTB2013 respectively when we fail IoU-guided optimization function. It demonstrates the bounding box refinement procedure can improve the accuracy of the tracking and the proposals are not always the best.

1.4 Further Analysis

[11] addresses weakness of IoU in the case of non-overlapping bounding boxes by introducing a generalized version of IoU formula. We experiment to employ GIoU to synthesize the results of visual, lingual and optical flow bounding boxes in our tracking task. According to our test, the

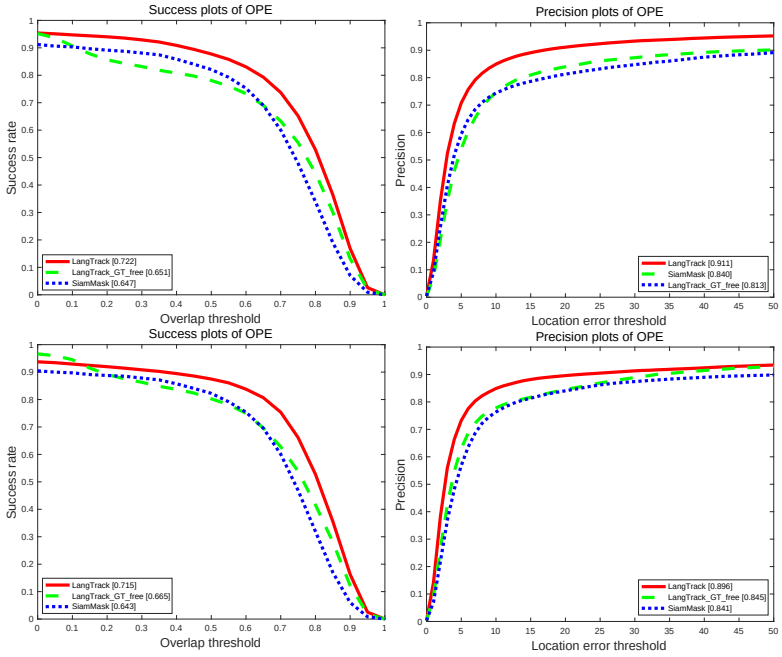

Figure 5: Comparative performance when initializing the LangTracker without a bounding box. We use the language module to detect and initialize the target. Left is the result on Lingual OTB 2015 while the right is on Lingual OTB 2013 comparing with original LangTrack and SiamMask. GTfree refers to initializing tracker without a Ground-Truth bounding box.

result show the model using GIoU to substitute for IoU outperform the SiamMask tracker though with 0.688 and 0.860 for success and precision rate , it performs worse than the original model using IoU. It demonstrates that distance of bounding boxes from different branches makes no contribution to supervise the tracking when there is no overlapping between them. We can consider it as the supervision disfunction and just depend on visual feature to finish the tracking.

In addition, we find the initial position of the target is already given before tracking in standard object tracking task and tracker can track the calibrated target directly. However,it is not always the case if we take the practical application into consideration, the target is not easily given but needs to be detected by algorithm based on the task or user's intention. Therefore we challenge our model by abandoning initializing target of the Ground Truth in the first frame, and instead using the language guided model to identify the visual target based on refering expression, then the tracker can track the subsequenial frames. We test the challenging task on lingual OTB 2013 and 2015 as showed in Fig.5, and we find our model achieve better performance than SiamMask, though worse than the one which initializing the target by a bounding box in the first frame. A comparative study has also been carried out among our method, the lastest NL trackers [4] [5] and the best prior attempt work [9]. The result has been illustrated on Table 4.

References

[1] Luca Bertinetto, Jack Valmadre, João F. Henriques, Andrea Vedaldi, and Philip H. S. Torr. Fully-convolutional siamese networks for object tracking. 2016.

[2] Martin Danelljan, Goutam Bhat, Fahad Shahbaz Khan, and Michael Felsberg. Atom: Accurate tracking by overlap maximization. 2018.

[3] Heng Fan, Liting Lin, Fan Yang, Peng Chu, and Haibin Ling. Lasot: A high-quality benchmark for large-scale single object tracking. 2018.

[4] Qi Feng, Vitaly Ablavsky, Qinxun Bai, Guorong Li, and Stan Sclaroff. Tell me what to track. 07 2019.

[5] Qi Feng, Qinxun Bai Vitaly Ablasvsky, and Stan Sclaroff. Robust visual object tracking with natural language region proposal network. In *ArXiv:1912.02048v1[cs.CV]*.

[6] Sahar Kazemzadeh, Mark Matten Vicente Ordonez, and Tamara L. Berg. Referitgame: Referring to objects in photographs of natural scene. In *Empirical Methods in Natural Language Processing (EMNLP)*, 2014.

[7] Bo Li, Wei Wu Juejie Yan, and Xiaolin Hu Zhu Zheng. High performance visual tracking with siamese region proposal network. *CVPR*, 2018.

[8] Bo Li, Wei Wu, Qiang Wang, Fangyi Zhang, and Junjie Yan. Siamrpn++: Evolution of siamese visual tracking with very deep networks. 2018.

[9] Zhenyang Li, Tao Ran, Efstratios Gavves, Cees G. M. Snoek, and Arnold W. M. Smeulders. Tracking by natural language specification. In *IEEE Conference on Computer Vision & Pattern Recognition*, 2017.

[10] Hyeonseob Nam and Bohyung Han. Learning multi-domain convolutional neural networks for visual tracking. 2015.

[11] Hamid Rezatofighi, JunYoung Gwak Nathan Tsoi, Ian Reid Amir Sadeghian, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. 2019.

[12] Qiang Wang, Li Zhang, Luca Bertinetto, Weiming Hu, and Philip H. S. Torr. Fast online object tracking and segmentation: A unifying approach. 2018.

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

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

[15] Zheng Zhu, Qiang Wang, Bo Li, Wei Wu, Junjie Yan, and Weiming Hu. Distractor-aware siamese networks for visual object tracking. 2018.