Supplementary Material

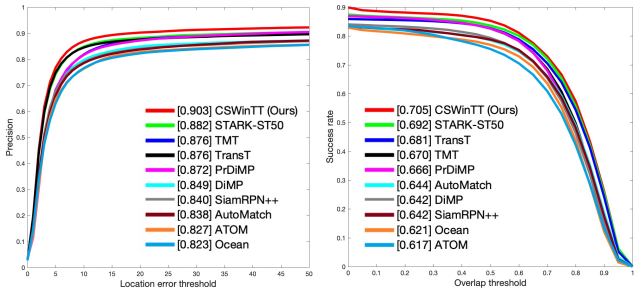
In the supplementary material, we provide additional detailed results and visual comparisons. In Section A, we report the detailed experimental results on UAV123 and LaSOT datasets, and supplement results for OTB [9] dataset. We then provide some visual comparison results of our CSWinTT with the state-of-the-art trackers in Section B. Finally, we show attribute-based evaluation results on the LaSOT dataset in Section C.

A. More Detailed Results

Here, we supplement with experimental result for the OTB [9] benchmarks, and provide more detailed results on UAV123 [7] and LaSOT [5] benchmarks. OTB [9] is a commonly used tracking dataset by the visual object tracking community including 100 sequences that are categorized according to 11 attributes. However, since the OTB dataset was proposed relatively early, most of the algorithms are now less distinguishable on this dataset (e.g., most algorithms are close to or above 90% in terms of Precision).

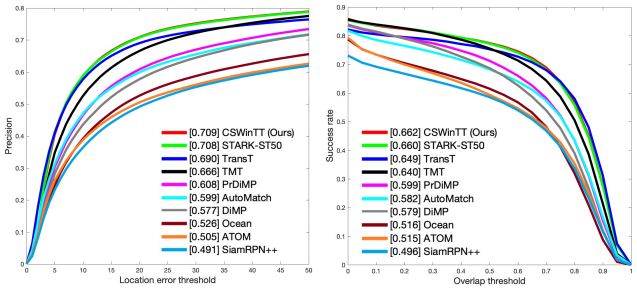
We use the precision plot and success plot for evaluating the trackers. For precision, we calculate the Euclidean distance of the centers between an estimated bounding box and the ground truth. The precision plot shows the percentage of the estimated bounding box in each frame whose center distance is less than a given threshold. Precision reflects the performance of location measurement of tracking algorithms but does not consider the target size. Success determines whether a tracker has been tracking a target successfully by calculating the size of the overlapped area, which is the IoU (intersection over union) between ground-truth and estimated box. Success plot shows the percentage of the estimated box whose overlap score is larger than a given threshold, and we adopt the Area Under the Curve (AUC) to rank trackers. We refer readers to [9] for more details about the metrics.

We compare with the most recent transformer trackers and some other representative algorithms, including STARK-ST50 [10], TransT [2], TMT [8], PrDiMP [4], DiMP [1], AutoMatch [11], Ocean [12], ATOM [3], and SiamRPN++ [6]. The precision plot and success plot on UAV123, LaSOT and OTB datasets are shown in Figure 1, Figure 2 and Figure 3 respectively. Our CSWinTT tracker achieves the best performance on both UAV123 and LaSOT datasets, and gets a relatively good result on the OTB dataset. On the OTB dataset, our approach gets a score of 88.4% in Precision and 68.0% in Success, this result does not exceed all tracking algorithms, but it performs very well in the transformer-based trackers such as STARK-ST50 [10] and TransT [2]. On UAV123, our approach achieves an absolute gain of 2.1% precision score and 1.3% success score over the previous best method STARK-ST50 [10].



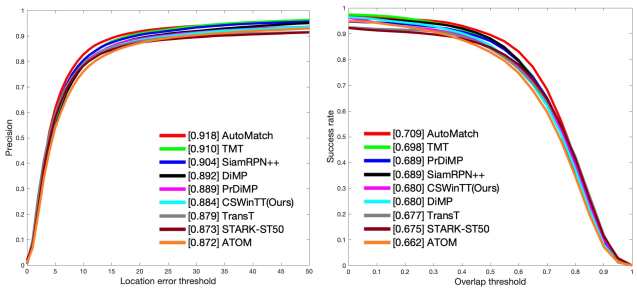
(a) Precision plots (b) Success plots

Figure 1. Precision and success plots on UAV123 [7].



(a) Precision plots (b) Success plots

Figure 2. Precision and success plots on LaSOT [5] test set.



(a) Precision plots (b) Success plots

Figure 3. Precision and success plots on OTB [9].

On the challenging LaSOT dataset, our approach obtains a 70.9% precision score and 66.2% success score, which significantly outperforms all previous state-of-the-art trackers.

B. Visual Comparisons

In Figure 4, we show exemplary visual comparisons between our approach and four state-of-the-art trackers on five challenging sequences. The sequences either contain motion blur, scale variation, occlusions, and similar objects from the challenging LaSOT [5] benchmark. The sequences in Figure 4 from top to bottom are 'bottle-1', 'goldfish-10', 'guitar-10', 'skateboard-8', and 'zebra-16'. It is clear from the comparison that, benefiting from the multi-scale window strategy which has better discrimination ability when

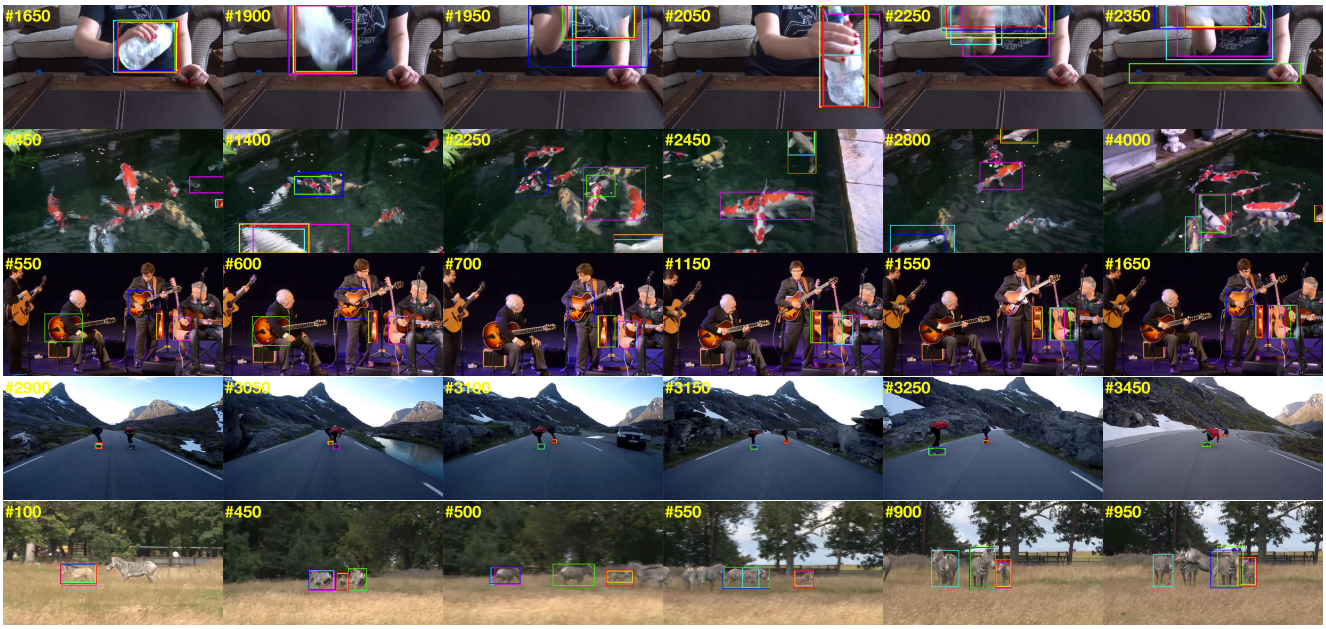


Figure 4. A visual comparison in special situation of our CSWinTT with other state-of-the-art trackers, i.e., STARK-ST50 [10], TransT [2], TrDiMP [8] and PrDiMP [4]. The ground-truth box (red line) and estimated boxes of each tracker is marked by lines with different colors as shown at the bottom. Each sequence covers one specific situation, including motion blur (first row), scale variation(second row), occlusions (third row) and similar objects (fourth and fifth rows) from the LaSOT [5] test set (from top to bottom are 'bottle-1', 'goldfish-10', 'guitar-10', 'skateboard-8', 'zebra-16').

the target scale changes greatly, our approach can locate the target more accurately when motion blur ('bottle-1', first row) and scale variation ('goldfish-10', second row) have happened. And with the cyclic shifting window attention which ensures the integrity of the tracking object and brings greater accuracy by expanding window samples, our tracker can better discriminate targets from complex backgrounds in the situation of occlusions ('guitar-10', third row) and similar objects ('skateboard-8', fourth row and 'zebra-16', fifth row).

C. Attribute Analysis

The attribute-based evaluation in terms of success on LaSOT [5] benchmark shown in Figure 5. Compared to other transformer trackers such as START-ST50 [10], TransT [2] and TrDiMP [8], our approach show remarkable results w.r.t. Aspect Ration Change, Camera Motion, Fast Motion, Low Resolution, Out-of-View, Partial Occlusion, Rotation, Scale Variation, and Viewpoint Change. This is because the proposed multi-scale cyclic shifting window attention that maintains the integrity of the object is very effective, compared with others our method performs more robustly when the appearance of the target changes greatly. Especially, in the situation of scale variation and out-of-view port, our tracker exhibits remarkable improvement and obtains AUC scores of 70.3% and 63.2%, significantly outperforms the second-best trackers by 3.9% and 1.3% respectively.

References

[1] Goutam Bhat, Martin Danelljan, Luc Van Gool, and Radu Timofte. Learning discriminative model prediction for tracking. In *Proceedings of the ICCV*, pages 6182–6191. IEEE, October 2019. 1

[2] Xin Chen, Bin Yan, Jiawen Zhu, Dong Wang, Xiaoyun Yang, and Huchuan Lu. Transformer tracking. In *Proceedings of the CVPR*, pages 8126–8135, 2021. 1, 2

[3] Martin Danelljan, Goutam Bhat, Fahad Shahbaz Khan, and Michael Felsberg. Atom: Accurate tracking by overlap maximization. In *Proceedings of the CVPR*, pages 4660–4669. IEEE, June 2019. 1

[4] Martin Danelljan, Luc Van Gool, and Radu Timofte. Probabilistic regression for visual tracking. In *Proceedings of the CVPR*, pages 7183–7192, 2020. 1, 2

[5] 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 CVPR*. IEEE, June 2019. 1, 2, 3

[6] 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 CVPR*, pages 4282–4291. IEEE, June 2019. 1

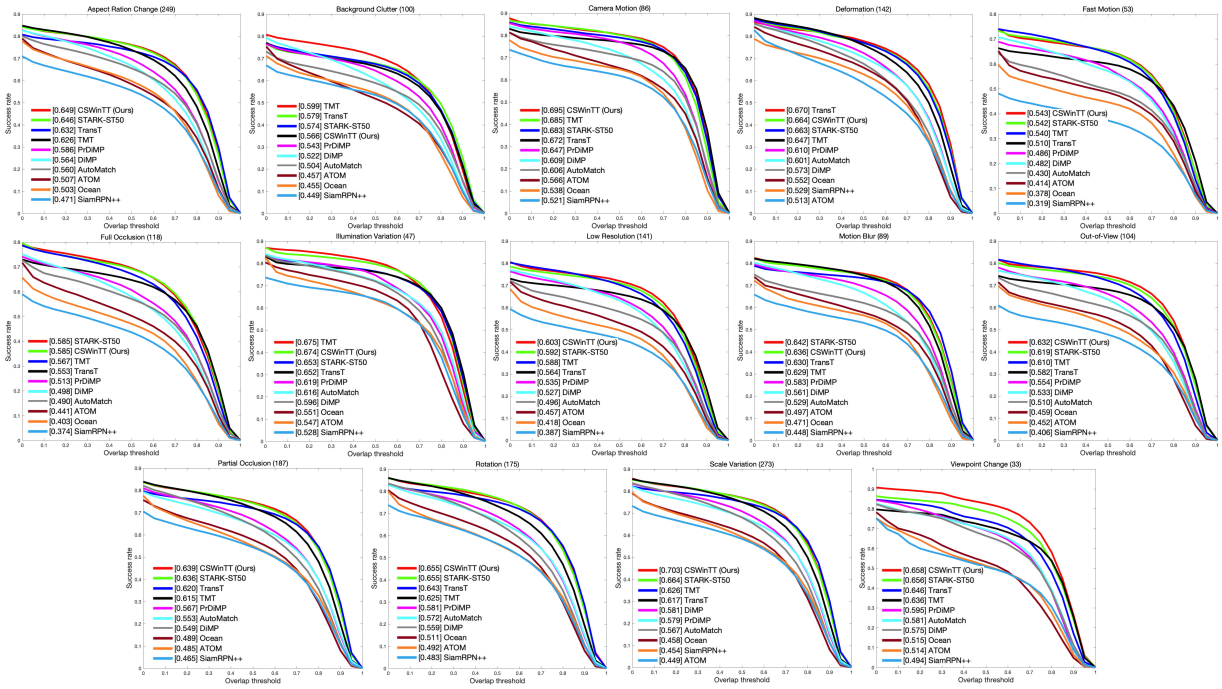


Figure 5. Attribute-based evaluation on the LaSOT [5] test set. The legend shows the AUC scores of the success plots.

[7] Matthias Mueller, Neil Smith, and Bernard Ghanem. A benchmark and simulator for uav tracking. In *Proceedings of the ECCV*, pages 445–461. Springer, 2016. 1

[8] Ning Wang, Wengang Zhou, Jie Wang, and Houqiang Li. Transformer meets tracker: Exploiting temporal context for robust visual tracking. In *Proceedings of the CVPR*, pages 1571–1580, 2021. 1, 2

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

[10] Bin Yan, Houwen Peng, Jianlong Fu, Dong Wang, and Huchuan Lu. Learning spatio-temporal transformer for visual tracking. In *Proceedings of the ICCV*, 2021. 1, 2

[11] Zhipeng Zhang, Yihao Liu, Xiao Wang, Bing Li, and Weiming Hu. Learn to match: Automatic matching network design for visual tracking. In *Proceedings of the ICCV*, pages 13339–13348, 2021. 1

[12] Zhipeng Zhang, Houwen Peng, Jianlong Fu, Bing Li, and Weiming Hu. Ocean: Object-aware anchor-free tracking. In *Proceedings of the ECCV*, pages 771–787, 2020. 1