Supplementary Material for Towards Sequence-Level Training for Visual Tracking

Minji Kim[1*] Seungkwan Lee[3,4*] Jungseul Ok[3] Bohyung Han[1,2] Minsu Cho[3]

[1]ECE & [2]IPAI, Seoul National University
[3]Pohang University of Science and Technology (POSTECH)
[4]Deeping Source Inc.
https://github.com/byminji/SLTtrack

A Training Details

We use Adam optimizer for SiamRPN++, SiamAttn, and TrDiMP, and AdamW optimizer for TransT. SiamRPN++ and SiamAttn are trained for 20 epochs with 10000 videos per epoch, while the learning rate starts from 10^{-5} and exponentially decays to 10^{-6}. Following the original papers, we use 0.1 times smaller learning rate for backbone layers for SiamRPN++ and 0.05 times smaller for SiamAttn, respectively. TransT is trained for 120 epochs with 1000 videos per epoch, and the learning rate starts from 0.1 times the original model setup in the paper and decreases by a factor of 10 after 100 epochs. TrDiMP is trained for 40 epochs with 5000 videos per epoch, and the learning rate starts from 0.04 times the initial learning rate from the original model and halves every 8 epochs. After the pre-training stage, the statistics of batch normalization layers are fixed and not updated during the RL fine-tuning stage.

B Evaluation on Additional Benchmarks

Table 1: Experimental results on NFS, UAV123, TNL2K, VOT2018, and VOT2020 baseline analysis. Test-time hyper-parameters are tuned only in VOT.

Method		NFS	UAV123	TNL2K	VOT2018			VOT2020		
		AUC	AUC	AUC	EAO	AO*	AUC*	EAO	A	R
SiamRPN++	Base	50.5	59.3	38.8	**39.7**	45.2	44.9	24.3	**45.5**	65.5
	+SLT	**56.5**	**61.2**	**44.1**	34.3	**48.8**	**48.6**	**25.4**	45.3	**70.8**
TrDiMP	Base	65.0	64.8	49.8	44.8	53.6	53.2	28.2	46.8	74.7
	+SLT	**65.6**	**66.3**	**50.7**	**45.0**	**54.1**	**53.6**	**28.9**	**47.0**	**76.2**
TransT	Base	65.3	66.6	53.5	30.0	51.0	50.6	**29.3**	**47.7**	75.3
	+SLT	**66.2**	**68.6**	**55.0**	**30.6**	**52.6**	**52.3**	**29.3**	46.7	**76.0**

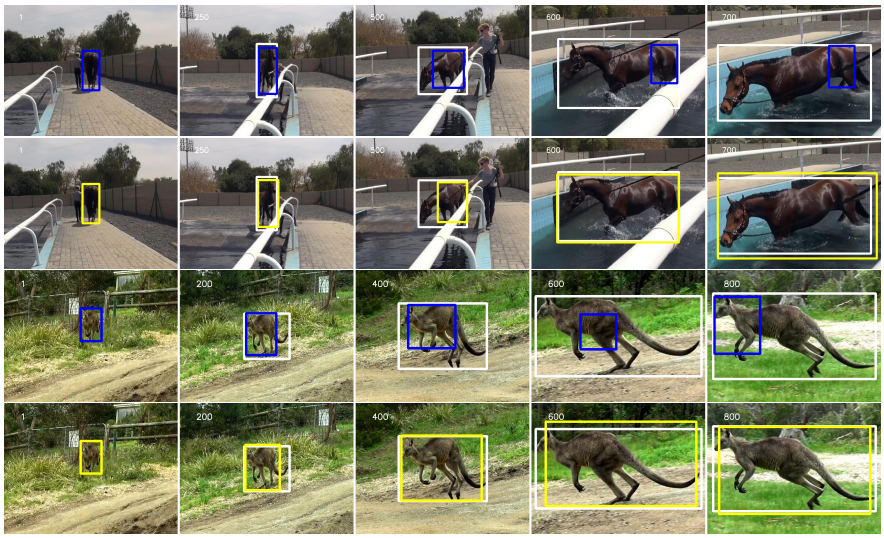
Table 1 summarizes the evaluation results of our method on additional benchmarks, where SLT consistently improves the AUC scores on NFS [1], UAV123 [5], and TNL2K [6], while strengthening the robustness (R) on VOT2020 [3]. Note that the re-initialization policy of VOT evaluation does not match with the reward system of SLT, which is designed for one-pass evaluation. Therefore, we also compare AO and AUC on VOT2018 [2],

* These authors contributed equally to this work.

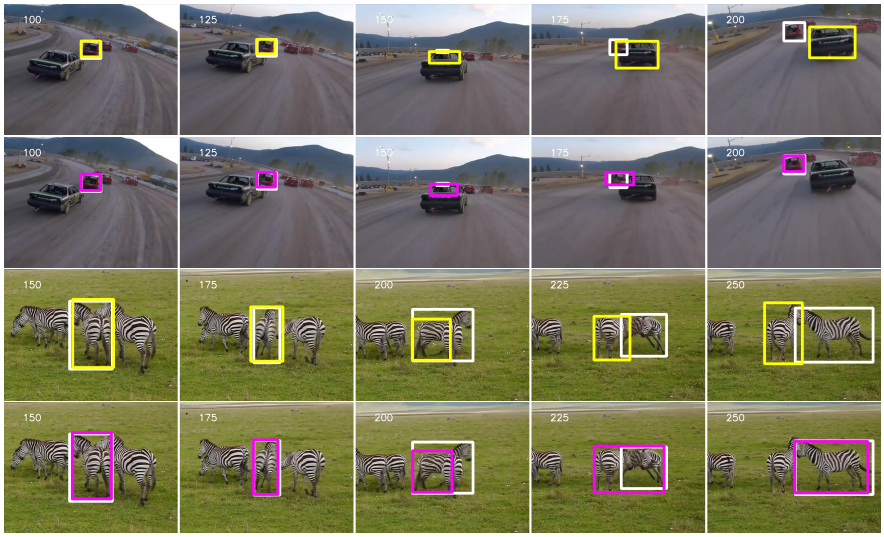
showing that SLT further benefits when the test-time metric and the reward system are aligned. Because VOT2020 does not provide the bounding box annotation, we cannot report AO and AUC.

C Qualitative Results

Section 4.4 of the main paper describes the benefits of sequence-level sampling (SS) and sequence-level objective (SO). To support our argument about the effect of these two components, we present qualitative results in Fig. 1, which visualizes the bounding-boxes corresponding to ground truth (white) and results from baseline (blue), baseline+SS (yellow), and baseline+SS+SO (magenta). For this analysis, we adopt SiamRPN++ [4] as the baseline tracker. As discussed in the main paper, SS makes trackers more robust to appearance updates given by scale changes, aspect ratio variations, and rotation. In Fig. 1a, there are two videos whose target objects change their appearance significantly. The tracker trained with SS successfully adapts to appearance variations, while the baseline tracker fails to capture the entire bodies of the target objects. Moreover, SO alleviates the drift issue in trackers in some challenging situations such as occlusion and background clutter. To qualitatively validate the properties, we present the target trajectories of the baseline+SS and baseline+SS+SO trackers in Fig. 1b, which includes the videos with such challenging attributes.



(a) Baseline *vs.* Baseline+SS



(b) Baseline+SS *vs.* Baseline+SS+SO

Fig. 1: Qualitative results. white: ground truth, blue: baseline, yellow: tracker trained with SS, magenta: tracker trained with SS and SO.

References

1. Kiani Galoogahi, H., Fagg, A., Huang, C., Ramanan, D., Lucey, S.: Need for speed: A benchmark for higher frame rate object tracking. In: ICCV (2017) 1
2. Kristan, M., Leonardis, A., Matas, J., Felsberg, M., Pflugfelder, R., Cehovin Zajc, L., Vojir, T., Bhat, G., Lukezic, A., Eldesokey, A., et al.: The sixth visual object tracking VOT2018 challenge results. In: ECCV Workshops (2018) 1
3. Kristan, M., Leonardis, A., Matas, J., Felsberg, M., Pflugfelder, R., Kämäräinen, J.K., Danelljan, M., Zajc, L.Č., Lukežič, A., Drbohlav, O., et al.: The eighth visual object tracking vot2020 challenge results. In: ECCV (2020) 1
4. Li, B., Wu, W., Wang, Q., Zhang, F., Junliang Xing, J.Y.: Siamrpn++: Evolution of siamese visual tracking with very deep networks. In: CVPR (2019) 2
5. Mueller, M., Smith, N., Ghanem, B.: A benchmark and simulator for uav tracking. In: ECCV (2016) 1
6. Wang, X., Shu, X., Zhang, Z., Jiang, B., Wang, Y., Tian, Y., Wu, F.: Towards more flexible and accurate object tracking with natural language: Algorithms and benchmark. In: CVPR (2021) 1