Supplementary Material

Bin Yu[1,2], Ming Tang[2], Linyu Zheng[1,2], Guibo Zhu[1,2], Jinqiao Wang[1,2,3],
Hao Feng[4], Xuetao Feng[4], Hanqing Lu[1,2]
[1]School of Artificial Intelligence, UCAS, China
[2]National Laboratory of Pattern Recognition, Institute of Automation, CAS, China
[3]ObjectEye Inc. [4]Alibaba Group
{bin.yu,tangm,linyu.zheng,gbzhu,jqwang,luhq}@nlpr.ia.ac.cn
{yuanning.fh,xuetao.fxt}@alibaba-inc.com

A. Discussion About Concurrent Work

So far, there are three concurrent single-object visual tracking methods based on Transformers, *i.e.*, TrDiMP [3], TransT [1], STARK [4]. Though these methods and ours all use transformers, the difference lies clearly in them. 1) TransT focuses on feature fusion between the target object and the search patches. 2) TrSiam is based on DiMP and improves the appearance features with transformers. 3) STARK is similar to our Concatenation baseline. However, Less background is contained in template images in STARK and the prediction head and training loss in STARK are different from those in our approach. 4) Different from them, DTT provides a novel online discriminative tracking pipeline. It models the foreground and background information of training frames with encoders and provides the discriminative feature embeddings for the decoders. Besides, the transformer architecture we use is pure, straightforward and efficient.

B. Detailed Results About Hyperparameters

Learning Rate We set the updating rate $\gamma = 0.01$ following the common settings in previous discriminative trackers [2, 5]. In our experiments on GOT-10k, the performance of DTT is not sensitive to γ when $0.001 < \gamma < 0.02$ (AUC ranges from 0.623 to 0.634). Using larger or smaller γ will cause worse results of DTT (*e.g.*, 0.226 when $\gamma = 1$ and 0.605 when $\gamma = 0$) due to the serious error accumulation and the lack of generalization ability, respectively.

Layers of Encoders/Decoders Setting only one layer of decoder and encoder in DTT obtains AUC of 0.615 and over 70 FPS on GOT-10k. Setting 4 layers can improve the results slightly but cause lower FPS (around 30). Thus we selected 2 layers at last for the balance of efficiency and accuracy.

Elliptic Weight The elliptic weight in Sec. 3.4 provides an improvement of 0.5% in our experiments compared with rectangular one.

References

[1] Xin Chen, Bin Yan, Jiawen Zhu, Dong Wang, Xiaoyun Yang, and Huchuan Lu. Transformer tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8126–8135, 2021. 1

[2] Martin Danelljan, Goutam Bhat, Fahad Shahbaz Khan, and Michael Felsberg. Eco: Efficient convolution operators for tracking. In *CVPR*, pages 6638–6646, 2017. 1

[3] Ning Wang, Wengang Zhou, Jie Wang, and Houqiang Li. Transformer meets tracker: Exploiting temporal context for robust visual tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1571–1580, 2021. 1

[4] Bin Yan, Houwen Peng, Jianlong Fu, Dong Wang, and Huchuan Lu. Learning spatio-temporal transformer for visual tracking. *arXiv preprint arXiv:2103.17154*, 2021. 1

[5] Linyu Zheng, Ming Tang, Yingying Chen, Jinqiao Wang, and Hanqing Lu. Fast-deepkcf without boundary effect. In *ICCV*, October 2019. 1