MixFormer: End-to-End Tracking with Iterative Mixed Attention Supplementary Material

Yutao Cui   Cheng Jiang   Limin Wang ✉   Gangshan Wu
State Key Laboratory for Novel Software Technology, Nanjing University, China
{cuiyutao,mg1933027}@smail.nju.edu.cn   {lmwang,gswu}@nju.edu.cn

1. Introduction

In this appendix, we first provide more results and analysis on OTB100 [2] and LaSOT [1] datasets. Then we give more visualization results of the attention weights on LaSOT. Finally, we provide more training details.

2. More Results

OTB-100. OTB100 [2] is a commonly used benchmark, which evaluates performance on Precision and AUC scores. Figure. 1 presents results of our trackers on both two metrics on OTB-100 benchmark. MixFormer-L reaches competitive performance w.r.t. state-of-the-art trackers, surpassing the transformer tracker TransT by 1.3% on AUC score. Besides, MixFormer-L is slightly higher than MixFormer.

LaSOT. LaSOT [1] has 280 videos in its test set. We evaluate our MixFormer on the test set to validate its long-term capability. To give a further analysis, we provide Success plot and Precision plot for LaSOT in Fig. 2. It proves that improvement is due to both higher accuracy and robustness.

3. More Visualization Results

In this section, we provide more visualization results of attention weights on *car-2* of LaSOT test dataset in Fig. 3. From the example, we can arrive at the same conclusion with section 4.3. Besides, from the last two lines, we infer that the features of last two blocks tend to adapt to the bounding box prediction head.

4. Training Details

We propose a 320x320 search region plus two 128x128 input images to make a fair comparison with prevailing trackers (e.g., Siamese-based trackers, STARK and TransT). Generally, we use 8 Tesla V100 GPUSs to train MixFormer with batch size of 32. MixFormer can also be trained on 8 2080Ti GPUs having only 11GB memory, with batch size of 8 per GPU. We use CvT21 and CvT24-W as the pretrained model for MixFormer and MixFormer-L respectively. We apply gradient clip strategy with the clip normalization rate of 0.1. For training stage-1 of MixFormer (i.e., MixFormer without SPM), we use GIoU loss and L_1 loss, with the weights of 2.0 and 5.0 respectively. Besides, the Batch Normalization layers of MixFormer backbone are frozen during the whole training process. For SPM training process, the backbone and corner-based localization head are frozen and the batch size is 32. SPM is trained for 40 epochs with the learning rate decays at 30 epochs.

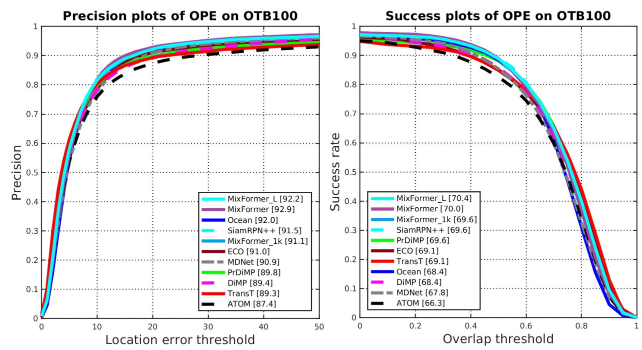

Figure 1. State-of-the-art comparison on the OTB100 dataset. Best viewed with zooming in.

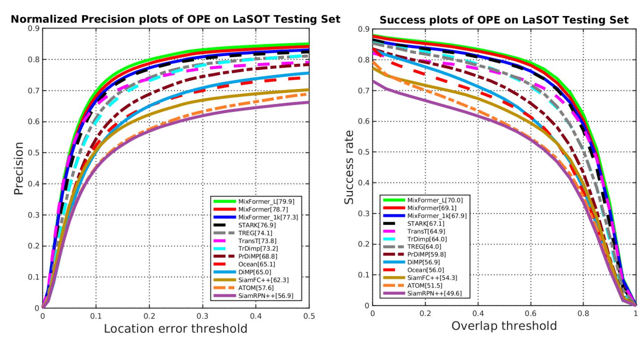

Figure 2. State-of-the-art comparison on the LaSOT dataset.

✉: Corresponding author.

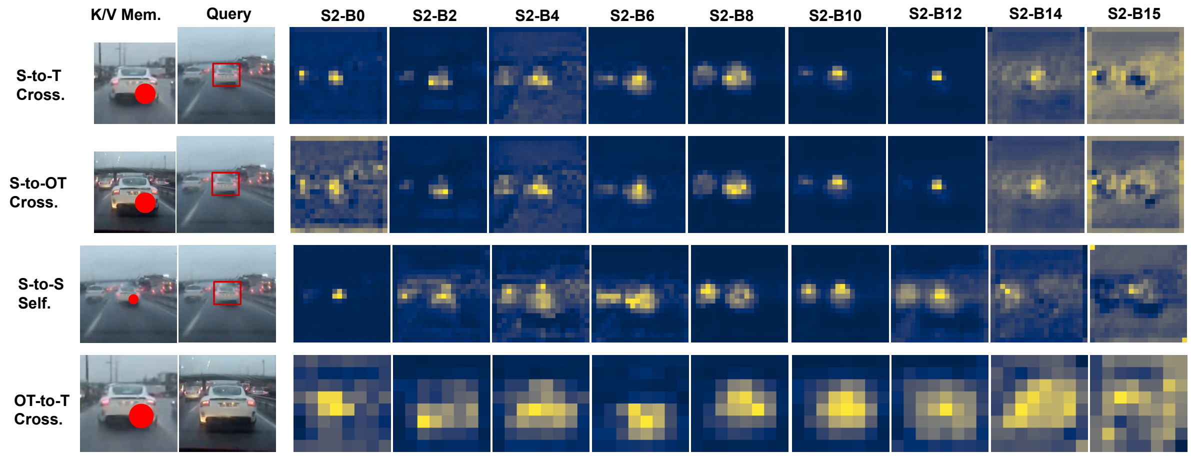

Figure 3. Visualization results of different attention weights on *car-2* of LaSOT. **S-to-t** is search-to-template cross attention, **S-to-OT** is search-to-online-template cross attention, **S-to-S** is self attention of search region and **OT-to-T** is online-template-to-template cross attention. **S**i-**B**j represents for Stage-i and Block-j of MixFormer. Best viewed with zooming in.

References

[1] 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 *CVPR*, 2019. 1

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