AiATrack: Attention in Attention for Transformer Visual Tracking (Supplementary Material)

The supplementary material provides additional details about the experiments and analyses of the proposed method.

1 Additional Experiment Details

1.1 Target Prediction

To make the tracking procedure in an end-to-end manner without tedious post-processing, we adopt the anchor-free prediction head proposed in [20], which outputs the probability maps $P_{tl}(x, y)$ and $P_{br}(x, y)$ for the top-left and the bottom-right bounding box corners. The coordinates $\widehat{x}_{tl}$, $\widehat{y}_{tl}$, $\widehat{x}_{br}$, $\widehat{y}_{br}$ of the predicted bounding box are then obtained by

$$\widehat{x}_{tl} = \sum_{y=0}^{H}\sum_{x=0}^{W} x \cdot P_{tl}(x, y),\ \widehat{y}_{tl} = \sum_{y=0}^{H}\sum_{x=0}^{W} y \cdot P_{tl}(x, y) \tag{1}$$

$$\widehat{x}_{br} = \sum_{y=0}^{H}\sum_{x=0}^{W} x \cdot P_{br}(x, y),\ \widehat{y}_{br} = \sum_{y=0}^{H}\sum_{x=0}^{W} y \cdot P_{br}(x, y) \tag{2}$$

1.2 Training Objective

With the predicted bounding box $\widehat{b}$ and predicted IoU $\widehat{i}$, the whole network is jointly trained by minimizing prediction errors. The bounding box prediction loss is defined as the combination of GIoU loss [16] and L1 loss. Together with the IoU prediction loss, the loss function can be written as

$$L = \lambda_{giou} L_{giou}(b, \widehat{b}) + \lambda_{l1} \|b - \widehat{b}\|_1 + \lambda_{mse}(i - \widehat{i})^2 \tag{3}$$

where b and i represent the ground truths of bounding box and IoU respectively and λ_{giou}, λ_{l1}, λ_{mse} are the trade-off weights.

1.3 Training Strategy

Similar to previous works [3,1,2,18,20], we utilize the training splits of LaSOT [5], TrackingNet [15], GOT-10k [7], and COCO [12] for offline training. As for the COCO image dataset, we apply data augmentation to generate synthetic video clips of diverse classes. During training, we randomly sample the search frame and reference frames such that the index of the search frame is larger than the indexes of reference frames. For training efficiency, we only sample one frame as the short-term reference. We also apply random affine transformations to jitter the sizes and locations of the short-term reference frame and search frame to simulate real tracking scenarios and avoid the influence of absolute position bias

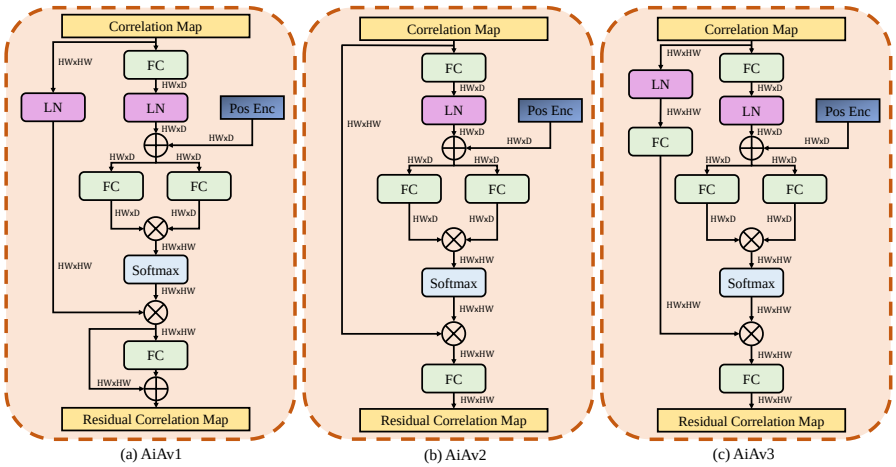

Fig. 1. Detailed illustration of the differences between different structures of the AiA module. ⊗ denotes matrix multiplication and ⊕ denotes element-wise addition. The numbers beside arrows are feature dimensions which do not include the batch size. Matrix transpose operations are omitted for brevity.

Structure	Modification			LaSOT [5]			$\text{LaSOT}_{\text{Ext}}$ [4]		
	LN	LT	IC	AUC	P_{Norm}	P	AUC	P_{Norm}	P
AiAv1	✓		✓	68.7	79.3	73.7	46.8	54.4	54.2
AiAv2				68.8	79.3	73.6	46.7	54.5	53.8
AiAv3	✓	✓		**69.2**	**79.6**	**74.3**	**48.4**	**56.6**	**56.2**

Table 1. Study about the different structures of the AiA module. **LN** denotes applying layer normalization to the value. **LT** denotes applying linear transformation to the value. **IC** denotes using identical connection after the correlation aggregation.

caused by padding [8,11,23]. The network is trained with the AdamW optimizer [13]. The learning rate is 1e-5 for the network backbone and 1e-4 for the other components. It decays by a factor of 10 during training. The parameters of the first convolutional layer and the first stage in the ResNet-50 [6] backbone are fixed during training.

1.4 Different Structures of the AiA Module

Besides variant (h) in the paper, we also explore other structures of the AiA module, where the following components are studied: (1) Layer normalization applied to the value. (2) Linear transformation applied to the value. (3) Identical connection after the correlation aggregation. To evaluate their effect, we design two other structures of the AiA module, *i.e.* AiAv2 and AiAv3. The differences between these structures are shown in Fig. 1. Note that AiAv1 is the structure we implement in AiATrack and AiAv3 is a typical self-attention structure in the vanilla Transformer [17].

Tracker	Alpha-Refine [21]	OceanPlus [22]	DualTFR [19]	STARK-ST50 [20]	AiATrack (Ours)
EAO	0.482	0.491	0.528	0.505	**0.530**
Accuracy	0.754	0.685	0.755	0.759	**0.764**
Robustness	0.777	**0.842**	0.836	0.817	0.827

Table 2. State-of-the-art comparison on VOT2020.

From the results in Tab. 1, we can observe that the layer normalization and the identical connection are not key components in our AiA module. Applying linear transformation to the value can further improve the performance, but we remove it for the trade-off between performance and computational cost. Besides the observations above, all the experimental results validate the effectiveness of correlation refinement in the conventional attention mechanism with an extra attention module.

1.5 Results on VOT

Different from previous reset-based evaluation protocol [10], VOT2020 [9] proposes a new anchor-based evaluation protocol which is more reasonable. The same as STARK [20] and DualTFR [19], we use Alpha-Refine [21] to generate masks for evaluation since the ground truths of VOT2020 are annotated by the segmentation masks. The overall performance is ranked by the Expected Average Overlap (EAO). As shown in Tab. 2, our tracker exhibits very competitive performance, outperforming STARK with a margin of 5% in terms of EAO.

2 Additional Visualization Results

2.1 Attribute Analysis

We also provide detailed attribute analysis on LaSOT [5]. Fig. 2 shows that our tracker has an encouraging performance in various kinds of scenarios like background clutter, camera motion, and deformation. The results suggest the great potential of the proposed method when dealing with challenging scenarios.

2.2 Qualitative Comparisons

To qualitatively compare our tracker with the state-of-the-art trackers, we visualize our tracking results with two recent representative trackers: KeepTrack [14] and STARK [20]. Fig. 3 shows the tracking outputs for these trackers on some challenging video examples.

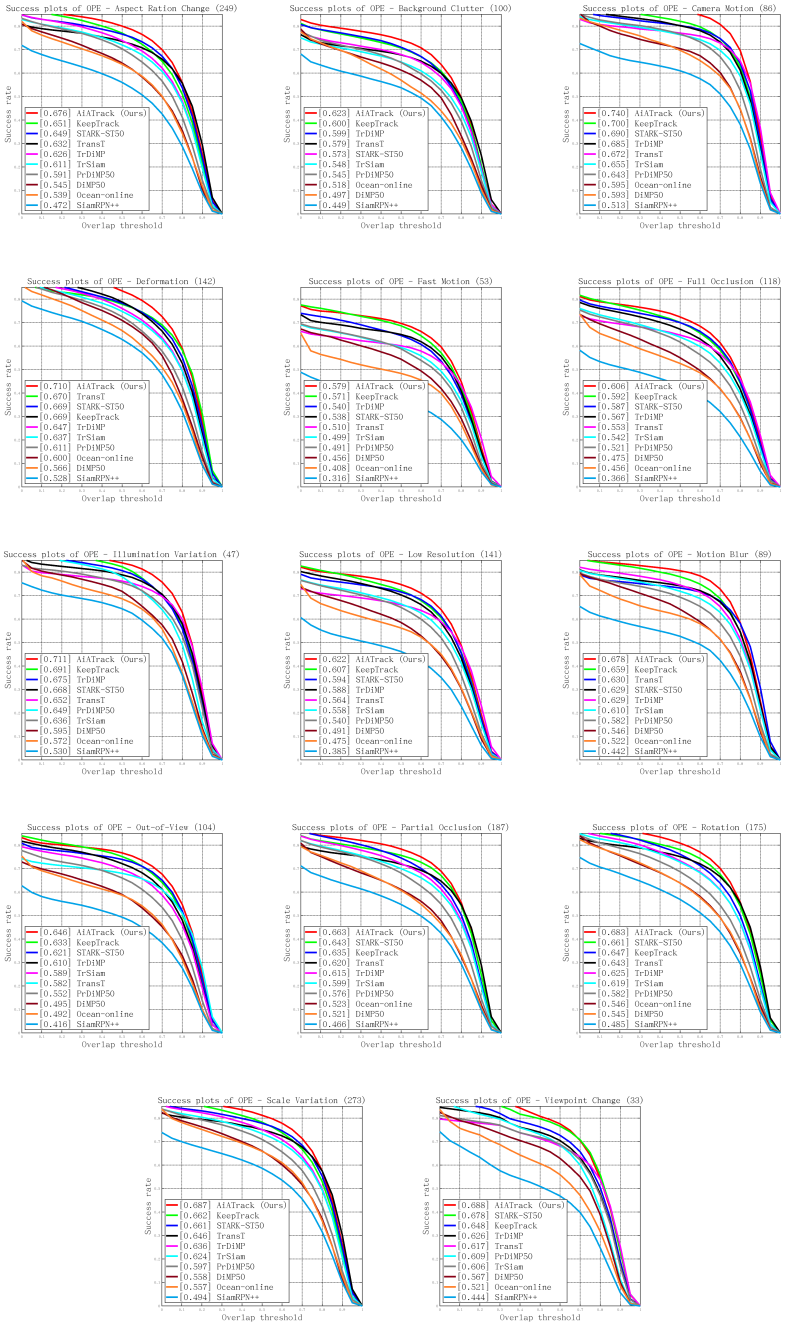

Fig. 2. Attribute analysis on LaSOT. AUC scores are showed in the legend.

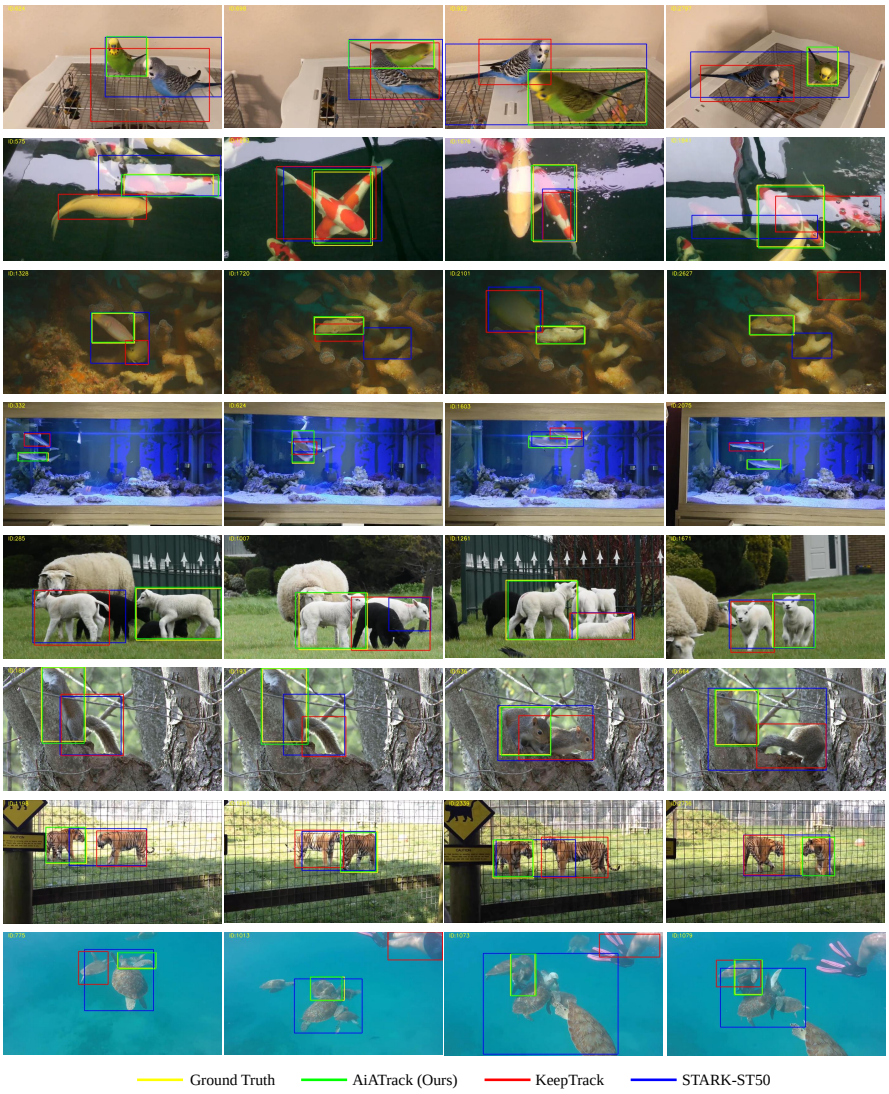

Fig. 3. Qualitative comparisons with two representative state-of-the-art trackers on 8 challenging sequences: *bird-17*, *goldfish-8*, *sepia-13*, *shark-2*, *sheep-3*, *squirrel-8*, *tiger-4*, *turtle-8*. Frame indexes are given on the top-left of each figure.

References

1. Bhat, G., Danelljan, M., Gool, L.V., Timofte, R.: Learning discriminative model prediction for tracking. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 6182–6191 (2019) 1
2. Chen, X., Yan, B., Zhu, J., Wang, D., Yang, X., Lu, H.: Transformer tracking. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8126–8135 (2021) 1
3. Danelljan, M., Bhat, G., Khan, F.S., Felsberg, M.: Atom: Accurate tracking by overlap maximization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4660–4669 (2019) 1
4. Fan, H., Bai, H., Lin, L., Yang, F., Chu, P., Deng, G., Yu, S., Huang, M., Liu, J., Xu, Y., et al.: Lasot: A high-quality large-scale single object tracking benchmark. International Journal of Computer Vision **129**(2), 439–461 (2021) 2
5. Fan, H., Lin, L., Yang, F., Chu, P., Deng, G., Yu, S., Bai, H., Xu, Y., Liao, C., Ling, H.: Lasot: A high-quality benchmark for large-scale single object tracking. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5374–5383 (2019) 1, 2, 3
6. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016) 2
7. Huang, L., Zhao, X., Huang, K.: Got-10k: A large high-diversity benchmark for generic object tracking in the wild. IEEE Transactions on Pattern Analysis and Machine Intelligence **43**(5), 1562–1577 (2019) 1
8. Islam, M.A., Jia, S., Bruce, N.D.: How much position information do convolutional neural networks encode? arXiv preprint arXiv:2001.08248 (2020) 2
9. 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: European Conference on Computer Vision. pp. 547–601. Springer (2020) 3
10. Kristan, M., Matas, J., Leonardis, A., Felsberg, M., Pflugfelder, R., Kamarainen, J.K., ˇCehovin Zajc, L., Drbohlav, O., Lukezic, A., Berg, A., et al.: The seventh visual object tracking vot2019 challenge results. In: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops. pp. 0–0 (2019) 3
11. Li, B., Wu, W., Wang, Q., Zhang, F., Xing, J., Yan, J.: Siamrpn++: Evolution of siamese visual tracking with very deep networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4282–4291 (2019) 2
12. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: European conference on computer vision. pp. 740–755. Springer (2014) 1
13. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017) 2
14. Mayer, C., Danelljan, M., Paudel, D.P., Van Gool, L.: Learning target candidate association to keep track of what not to track. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 13444–13454 (2021) 3
15. Muller, M., Bibi, A., Giancola, S., Alsubaihi, S., Ghanem, B.: Trackingnet: A large-scale dataset and benchmark for object tracking in the wild. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 300–317 (2018) 1
16. Rezatofighi, H., Tsoi, N., Gwak, J., Sadeghian, A., Reid, I., Savarese, S.: Generalized intersection over union: A metric and a loss for bounding box regression. In:

Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 658–666 (2019) 1
17. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017) 2
18. Wang, N., Zhou, W., Wang, J., Li, H.: Transformer meets tracker: Exploiting temporal context for robust visual tracking. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1571–1580 (2021) 1
19. Xie, F., Wang, C., Wang, G., Yang, W., Zeng, W.: Learning tracking representations via dual-branch fully transformer networks. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2688–2697 (2021) 3
20. Yan, B., Peng, H., Fu, J., Wang, D., Lu, H.: Learning spatio-temporal transformer for visual tracking. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 10448–10457 (2021) 1, 3
21. Yan, B., Zhang, X., Wang, D., Lu, H., Yang, X.: Alpha-refine: Boosting tracking performance by precise bounding box estimation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5289–5298 (2021) 3
22. Zhang, Z., Liu, Y., Li, B., Hu, W., Peng, H.: Toward accurate pixelwise object tracking via attention retrieval. IEEE Transactions on Image Processing **30**, 8553–8566 (2021) 3
23. Zhang, Z., Peng, H.: Deeper and wider siamese networks for real-time visual tracking. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4591–4600 (2019) 2