Foreground-Background Distribution Modeling Transformer for Visual Object Tracking
—— Supplementary Material ——

Dawei Yang[1,*], Jianfeng He[1,*], Yinchao Ma[1], Qianjin Yu[1], Tianzhu Zhang[1,†]
[1] University of Science and Technology of China
{yangdawei,hejf,imyc,sa21010105}@mail.ustc.edu.cn, {tzzhang}@ustc.edu.cn

In the supplementary material, we first provide more details about pseudo-bbox generation in the method section. Then, we present more qualitative results to demonstrate the effectiveness of our tracker.

1. More Pseudo-bbox Generation Details

As described in the main text (Section 3.2), we conduct similarity matching to compute the matching point in search region feature $\widehat{\mathbf{E}}_x^l$ for each point in $\widehat{\mathbf{E}}_t^l$ according to the similarity probability, which can obtain a set of matching points $\{(\hat{x}_k^l, \hat{y}_k^l)\}_{k=1}^K$, where $K = h_t w_t$. To obtain pseudo-boxes of the search region, we refer to RepPoints [7] for deriving the target box $\hat{b}_x^l = (\hat{x}^l, \hat{y}^l, \hat{w}^l, \hat{h}^l)$ with the mean and the standard deviation of all keypoints. The mean of keypoints indicates the center of target object, and the standard deviation of keypoints represents the relationship between the height and width of the target bounding box. Formally,

$$
\begin{aligned}
(\hat{x}^l, \hat{y}^l) &= (\frac{1}{K}\sum_{k=1}^{K}\hat{x}_k^l, \frac{1}{K}\sum_{k=1}^{K}\hat{y}_k^l), \\
(\hat{w}^l, \hat{h}^l) &= (\sqrt{\sum_{k=1}^{K}\frac{(\hat{x}_k^l - \hat{x}^l)^2}{\lambda_w K}}, \sqrt{\sum_{k=1}^{K}\frac{(\hat{y}_k^l - \hat{y}^l)^2}{\lambda_h K}}),
\end{aligned} \tag{1}
$$

where λ_h, λ_w are learnable scale factors.

2. Qualitative Results

More Fore-background Agent Activation Maps. Here, we visualize more activation maps of fore-background agents (FB-agents) in the template and search region. As shown in Figure 1, the activation maps are computed by the similarity between FB-agents and feature maps of the template and search region, respectively. This reflects the ability of FB-agents to perceive the foreground and background of features. For any video sequence of the tracking scene, it can be seen that FB-agents generated by the FBAL module can provide sufficient foreground and background priors, whether for templates or search regions.

More Response Maps Visualization. We visualize more feature responses in complex scenes and distractor cases to qualitatively prove the effectiveness of our method. As shown in Figure 2, recent transformer trackers [8, 1, 5, 4, 6] utilize plain attention to indiscriminately interact among all pixels, which are easily disturbed by cluttered backgrounds due to insufficient consideration of fore-background relationship, as shown in the right half of Figure 2. Our tracker can accurately discriminate the specified target features, benefiting from target-aware capabilities provided by the FB-agents, driving distribution-aware attention can effectively suppress false responses to the background. In particular, our tracker has strong discriminative ability in distractor cases, which also proves that distribution-aware attention can avoid the influence of distractors more effectively than appearance-based plain attention.

More Visualization Comparisons between the plain attention mechanism and our DA^2 mechanism. In order to have a clearer understanding of the proposed distribution-aware attention (DA^2) module, as shown in Figure 3, we further qualitatively analyze the false response point on the plain attention response maps from the perspective of distribution similarity, and visually compare the distribution maps of the two attention mechanisms on the interference pixel. As can be seen from the first two rows of Figure 3, plain attention models the relationship between the template and search region based on the similarity of appearance, which is easily disturbed by cluttered backgrounds and distractors, due to the high correlation between similar appearance features of distractors, resulting in being incorrectly focused. The proposed DA^2 module reconstructs the plain attention from the similarity of the fore-background distributions, which can significantly distinguish the feature of distractors. Moreover, the last row of Figure 3 proves that our tracker can also maintain more features belonging to foreground targets, which leads to the object state estimation more accurate.

Tracking Results. To show the tracking performance of the proposed tracker more intuitively, we compare it with

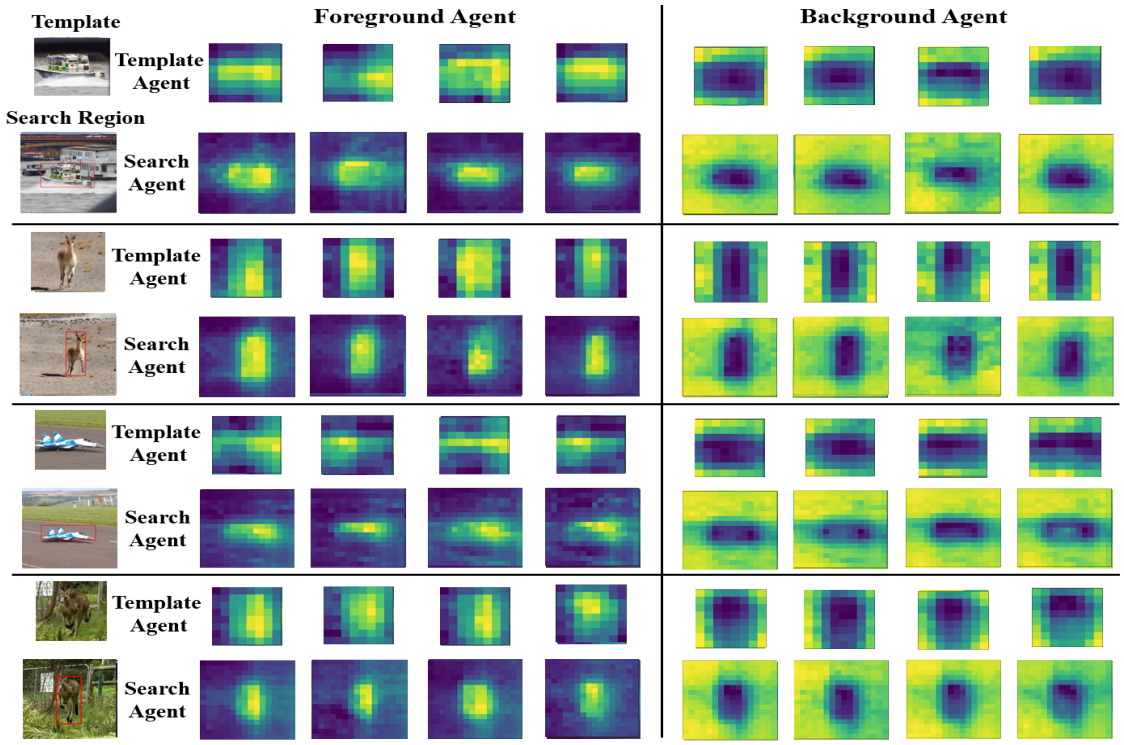

Figure 1. Visualizations of fore-background agent activation maps for the template and search region.

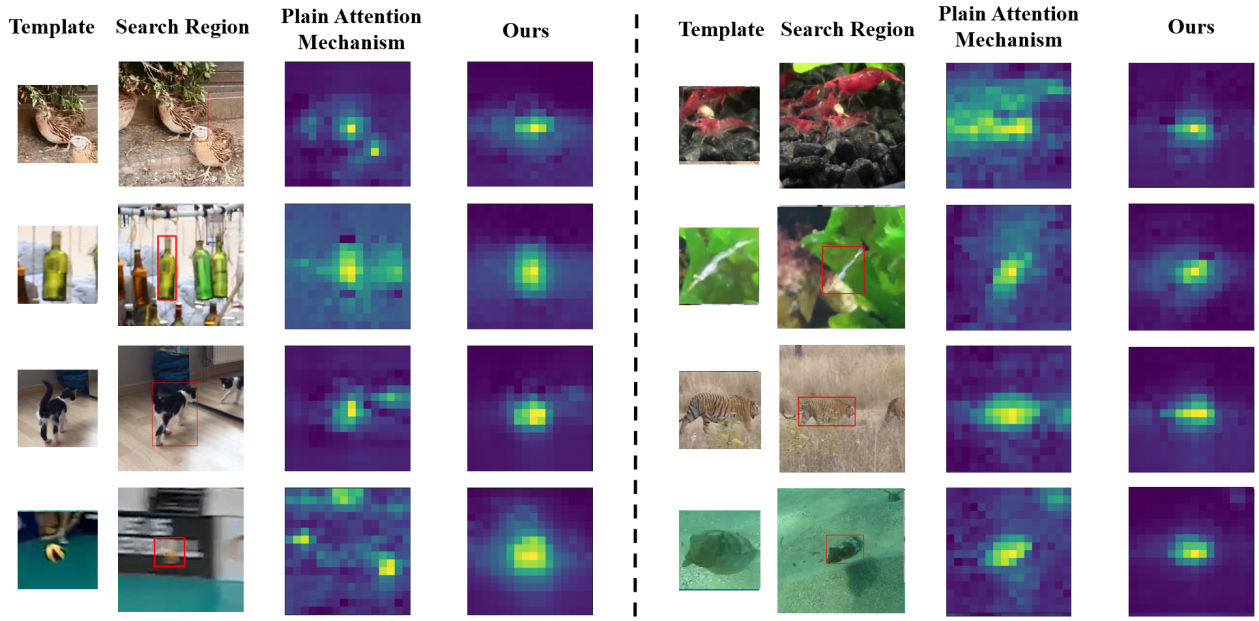

Figure 2. The response visualization of search region features to target objects. Maps are calculated by the similarity between the central of template and the feature map of search region. Our proposed tracker based on distribution-aware attention can greatly enhance the discriminability of features in complex scenarios and distractors.

some state-of-the-art trackers [8, 1, 2, 3] by directly visualizing the tracking results of video sequences in different challenging scenarios. As shown in Figure 4 (a) and (d), F-BDMTrack can still accurately track the target in the presence of severe distractors, and the results are consistent with groundtruth. However, some current trackers [8, 1, 2, 3] utilize appearance-based attention mechanisms, which are easily distracted by distractors near the target, resulting in

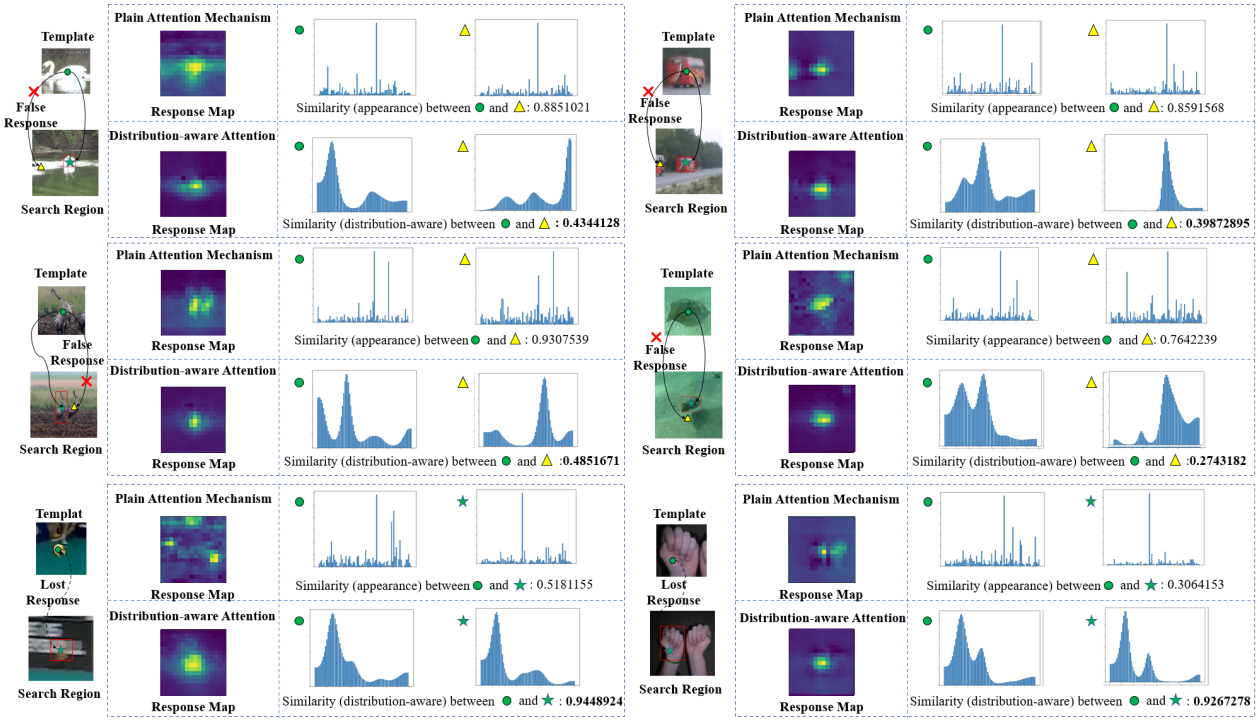

Figure 3. Visualization of distribution maps of proposed distribution-aware attention and plain attention mechanism on distracting pixels. Distribution-aware attention can effectively suppress the template response to cluttered backgrounds and distractors in the search region.

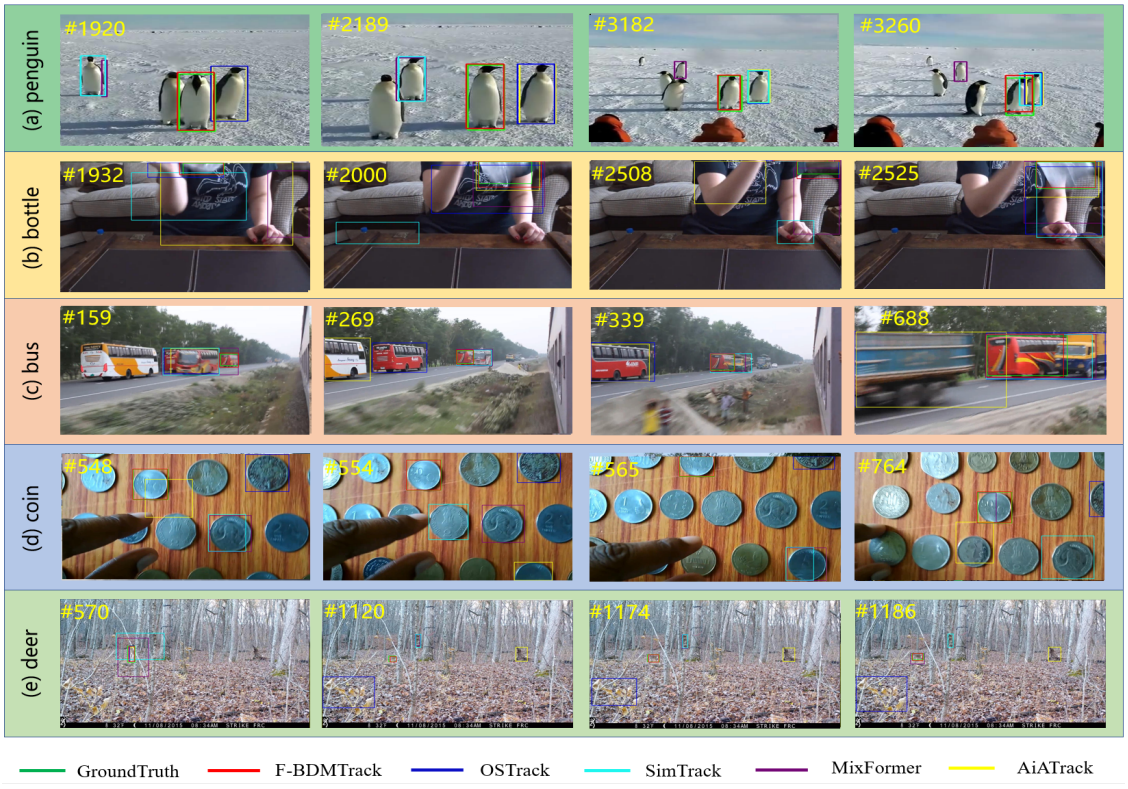

Figure 4. A comparison of our tracker with state-of-the-art trackers in different challenge scenes. Our tracker can achieve more accurate tracking in a variety of challenging scenes, such as distractor cases, motion blur and cluttered backgrounds.

mislocalization to distractors. In Figure 4 (b), our tracker can also accurately locate the target in the case of motion blur, while some current trackers [8, 1, 2, 3] often have shift in target state positioning. Because the plain attention based on appearance similarity is difficult to accurately perceive the target from the appearance similarity when the target is moving rapidly, while the proposed DA^2 module distinguishes the target from the perspective of fore-background distribution similarity, thus avoiding the dilemma that the appearance of the same object is not similar. Likewise, in the cluttered background scenes, as shown in Figure 4 (e), our tracker achieves superior tracking results, further proving the effectiveness of fore-ground distribution modeling via FB-agents on the visual object tracking task.

References

[1] Boyu Chen, Peixia Li, Lei Bai, Lei Qiao, Qiuhong Shen, Bo Li, Weihao Gan, Wei Wu, and Wanli Ouyang. Backbone is all your need: A simplified architecture for visual object tracking. In *Proceedings of the European Conference on Computer Vision*, 2022. 1, 2, 4

[2] Yutao Cui, Jiang Cheng, Limin Wang, and Gangshan Wu. Mixformer: End-to-end tracking with iterative mixed attention. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2022. 2, 4

[3] Shenyuan Gao, Chunluan Zhou, Chao Ma, Xinggang Wang, and Junsong Yuan. Aiatrack: Attention in attention for transformer visual tracking. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXII*, pages 146–164. Springer, 2022. 2, 4

[4] Liting Lin, Heng Fan, Yong Xu, and Haibin Ling. Swintrack: A simple and strong baseline for transformer tracking. *Advances of Neural Information Processing Systems*, 2021. 1

[5] Fei Xie, Chunyu Wang, Guangting Wang, Yue Cao, Wankou Yang, and Wenjun Zeng. Correlation-aware deep tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8751–8760, 2022. 1

[6] Bin Yan, Houwen Peng, Jianlong Fu, Dong Wang, and Huchuan Lu. Learning spatio-temporal transformer for visual tracking. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 10448–10457, 2021. 1

[7] Ze Yang, Shaohui Liu, Han Hu, Liwei Wang, and Stephen Lin. Reppoints: Point set representation for object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 9657–9666, 2019. 1

[8] Botao Ye, Hong Chang, Bingpeng Ma, and Shiguang Shan. Joint feature learning and relation modeling for tracking: A one-stream framework. *Proceedings of the European Conference on Computer Vision*, 2022. 1, 2, 4