Supplementary Material for HIPTrack: Visual Tracking with Historical Prompts

Wenrui Cai[1], Qingjie Liu[1,2,3,*], Yunhong Wang[1,3]
[1]State Key Laboratory of Virtual Reality Technology and Systems, Beihang University, Beijing, China
[2]Zhongguancun Laboratory, Beijing, China
[3]Hangzhou Innovation Institute, Beihang University, Hangzhou, China
{wenrui_cai, qingjie.liu, yhwang}@buaa.edu.cn

In this supplementary material, we present additional ablation studies on the historical prompt encoder and the historical prompt decoder in Section 1 to demonstrate the effectiveness of our design. In Section 2, we provide more comprehensive performance comparisons between our proposed HIPTrack and other trackers on LaSOT [3] *test* split, as well as their performance in different complex scenarios within LaSOT. In Section 3, we provide more qualitative visualization analyses.

1. Further Analyses

1.1. Ablation Studies on Historical Prompt Encoder

In the historical prompt encoder, we employ a lightweight network Φ to perform initial encoding on the input 4-channel image tensor and obtain the feature map $\boldsymbol{F}$. In the first row of Table 1, we investigate whether the encoder Φ can be made even lighter. We replace Φ from the first three stages of ResNet-18 [7] with a single convolutional layer. The results in the first and fourth rows indicate that replacing Φ with a single convolutional layer leads to performance degradation, which suggests that a stronger initial encoder Φ is required for encoding the historical target features.

After obtaining the feature map $\boldsymbol{F}_1'$ from the output of the residual block f_{RB1}, the historical prompt encoder employs spatial attention and channel attention to enhance the feature map $\boldsymbol{F}_1'$. In the experiments of the second and third rows in Table 1, we respectively remove these two branches from the historical prompt encoder to investigate their impact on tracking accuracy. The comparative results of the second, third, and fourth rows indicate that incorporating channel attention and spatial attention both yield positive benefits in tracking accuracy.

1.2. Ablation Studies on Historical Prompt Decoder

The historical prompt decoder is utilized to store the historical target features and adaptively aggregate them with the current search region feature to generate the historical prompt. In Table 2, we investigate the impact of different memory bank sizes on the tracking performance. In Table 3, we investigate the impact of different memory update intervals on the tracking performance.

Table 1. Ablation studies on lightweight encoder Φ and whether to use channel and spatial attention on LaSOT *test* set.

#	Φ	Channel	Spatial	AUC(%)	$\mathbf{P}_{\mathbf{Norm}}$(%)	P(%)
1	✘	✔	✔	72.1	82.2	78.7
2	✔	✘	✔	72.3	82.4	79.1
3	✔	✔	✘	72.4	82.5	79.1
4	✔	✔	✔	**72.7**	**82.9**	**79.5**

Memory Bank Size. The first four rows of Table 2 indicate that increasing the memory bank size leads to a performance improvement. In our approach, we set the memory bank size to 150 without carefully tuning, which also implies that there is still potential for performance improvement in our approach. The results from the fourth and fifth rows of Table 2 demonstrate that adding more historical information at the initial stage of tracking leads to a slight performance improvement. This may be because the result in the initial stage of tracking usually has higher accuracy, which facilitates the rediscovery of the target after it is lost.

Update Interval. As shown in Table 3, we conduct experiments using different update intervals on LaSOT. We find that setting the update intervals to 5 or 10 resulted in negligible performance improvement. Additionally, lower update intervals require more frequent calls to the historical prompt encoder, which can diminish efficiency. On the other hand, longer intervals such as 30 lead to a decline in performance. Therefore, we have chosen an update interval of 20, without carefully tuning as well.

2. More Detailed Results in Different Attribute Scenes on LaSOT

In Figure 1, we present more detailed quantitative comparisons of the success curves between our proposed HIP-

*Corresponding author.

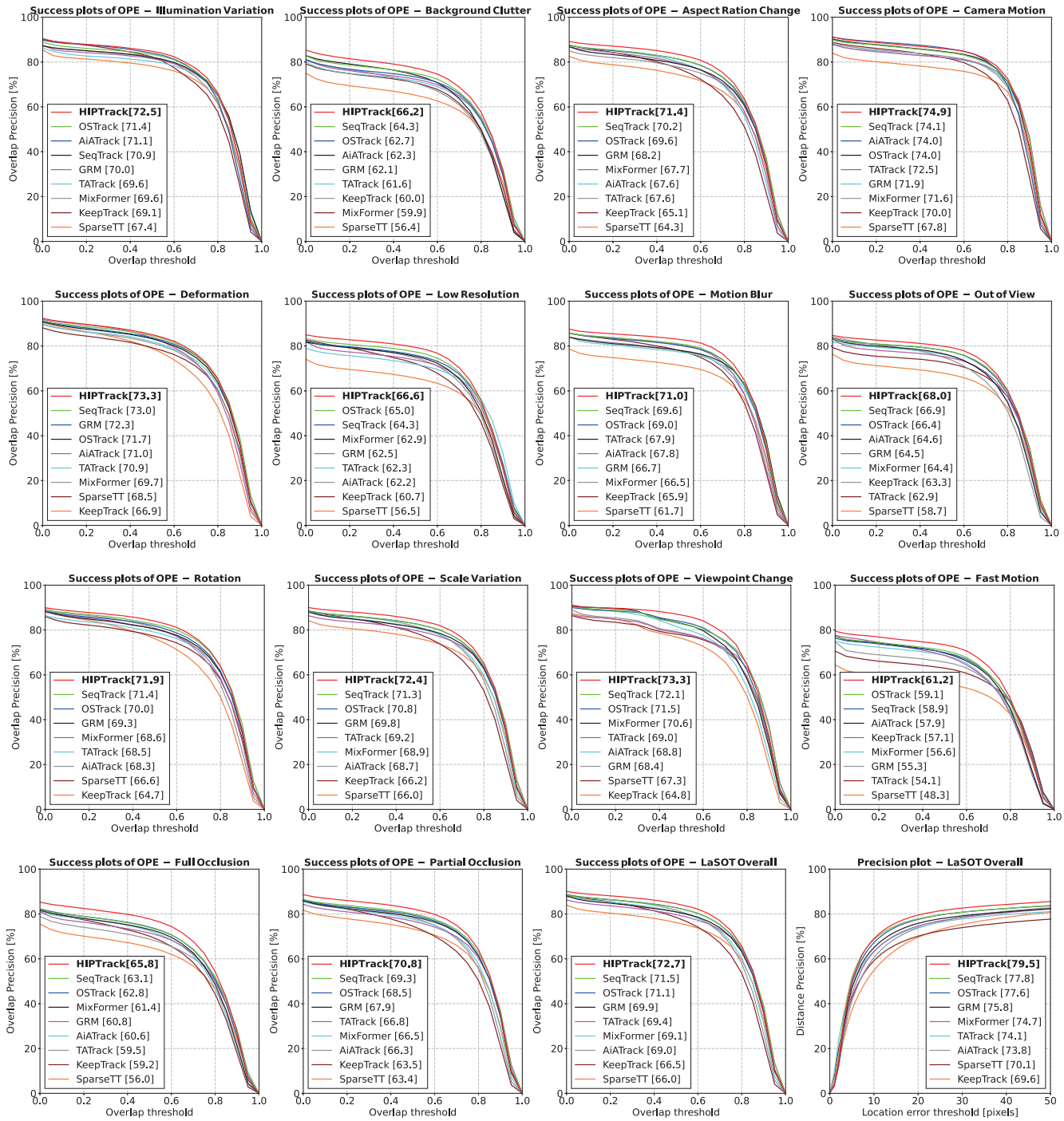
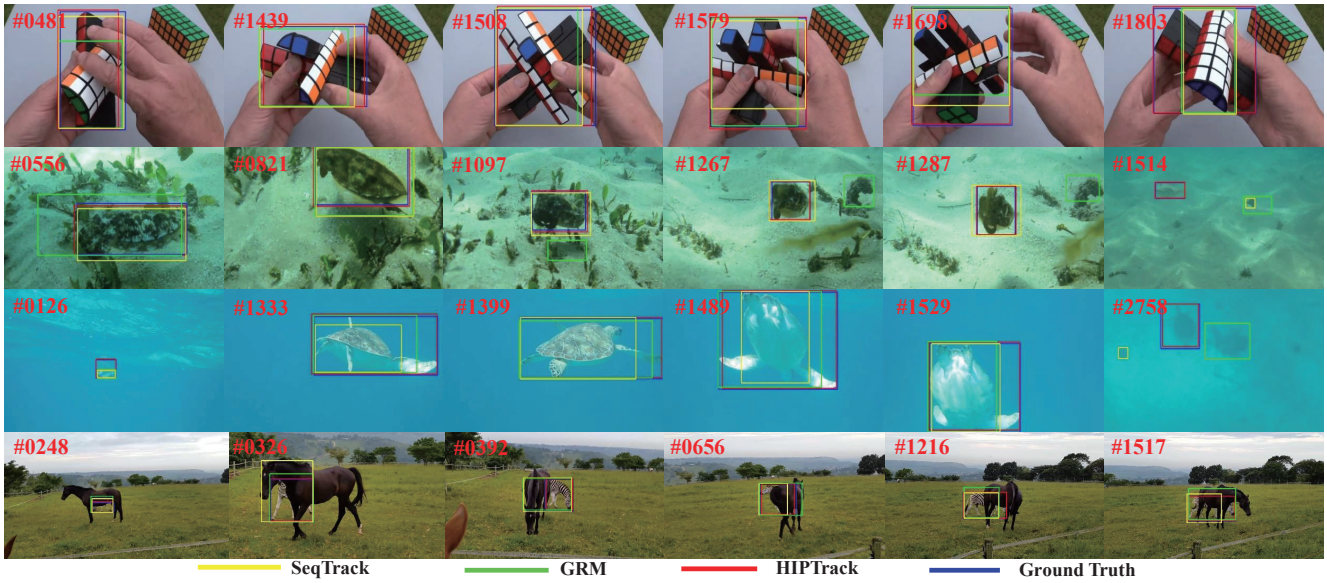
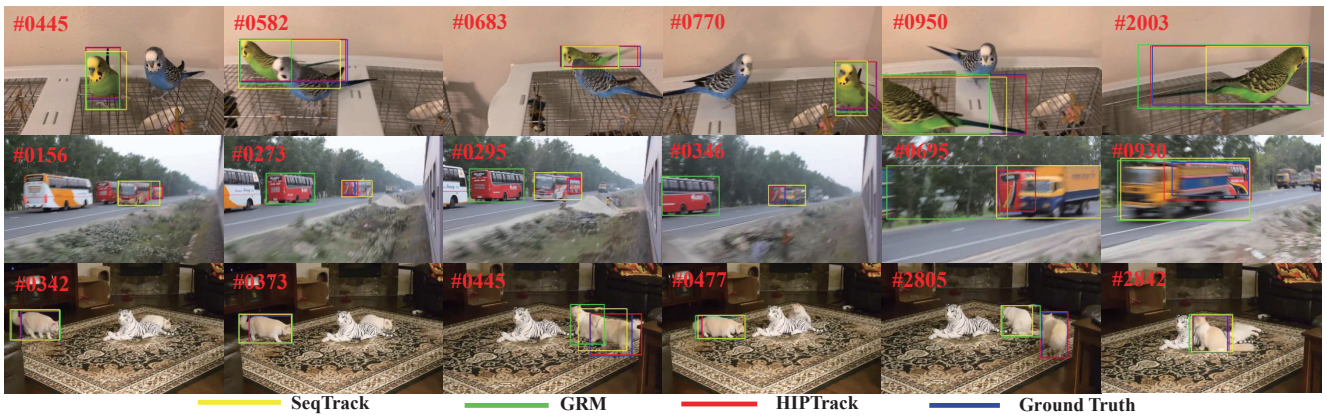


Figure 1. Comparisons of our proposed HIPTrack with other excellent trackers in the success curve on LaSOT *test* split, which includes eleven special scenarios such as Low Resolution, Motion Blur, Scale Variation, etc. We also provide the comparisons of the success and precision curves across the entire LaSOT *test* split.

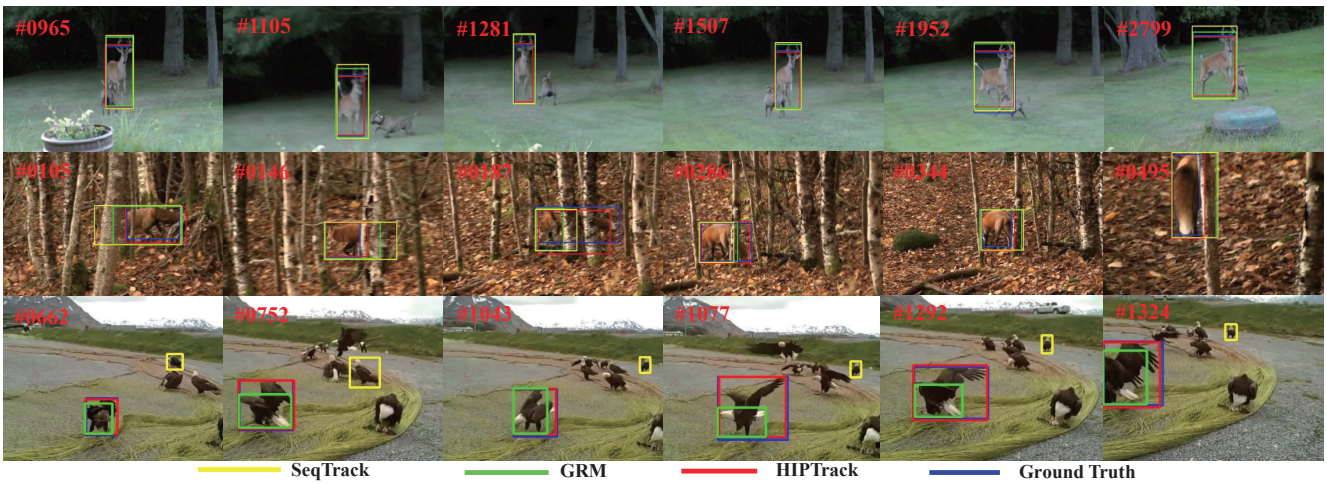
Track and other excellent trackers SeqTrack [1], GRM [6], TATrack [8], MixFormer [2], KeepTrack [9], OSTrack [10], AiATrack [5], and SparseTT [4] across various attribute scenes in LaSOT [3] *test* split. Figure 1 illustrates that our HIPTrack outperforms other trackers across all subsets of videos with special attributes in LaSOT. Particularly, when



(a) Qualitative results of three methods when the targets undergo large deformations.



(b) Qualitative results of three methods when the targets suffer from partial occlusions.



(c) Qualitative results of three methods when the targets have large scale variations.

Figure 2. This figure presents a visual comparison among our method, SeqTrack [1] and GRM [6] in the challenges of target deformation, partial occlusion and scale variation. It demonstrates that our method achieves more effective and accurate tracking in the aforementioned challenging scenarios. Zoom in for better view.

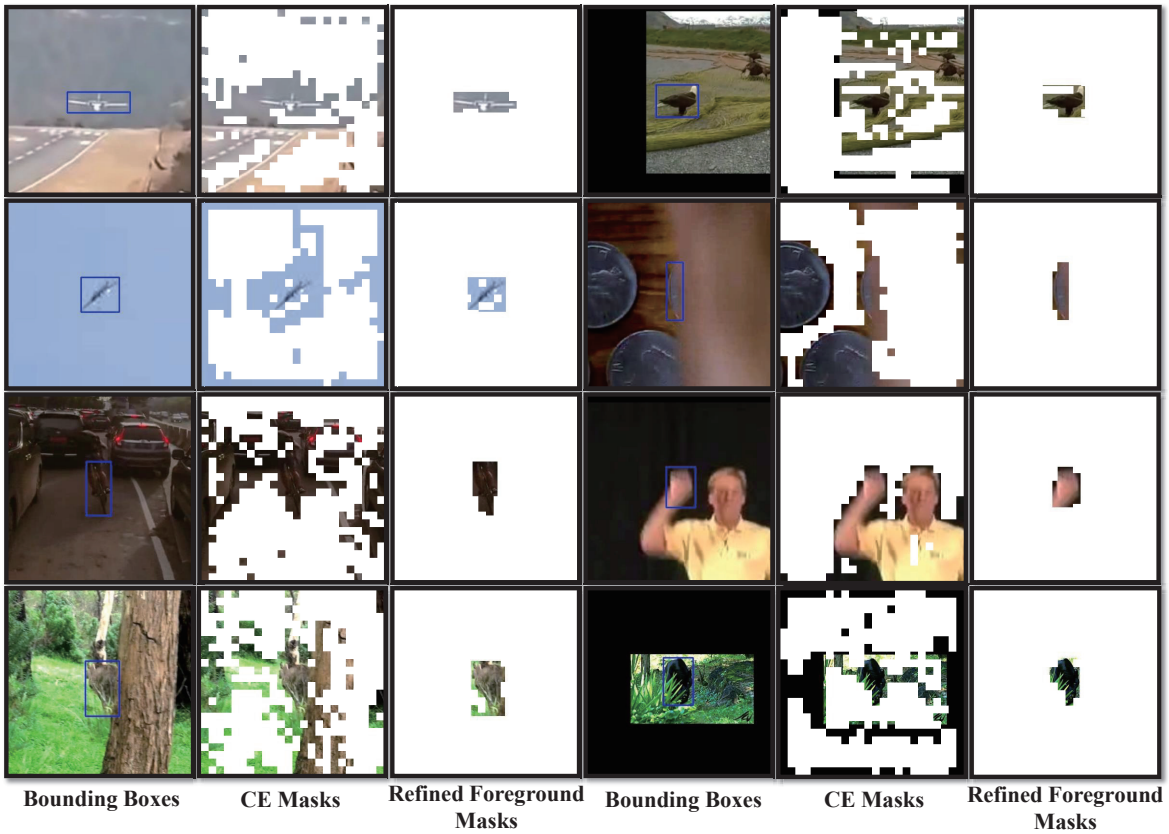


Figure 3. Visualization results of refined foreground masks. The construction process of refined foreground masks involves combining the bounding box masks generated from the predicted bounding boxes and the CE masks obtained from the candidate elimination module within the feature extraction network. The combining process is performed using the bitwise and operation.

Table 2. Ablation study on memory bank sizes and whether to preserve the first 10 memory frames on LaSOT *test* set.

#	Memory Bank Size	init 10	AUC(%)	$\mathbf{P_{Norm}}$(%)	P(%)
1	70	✔	72.5	82.7	79.4
2	100	✔	72.5	82.8	79.4
3	120	✔	72.6	82.8	79.5
4	150	✔	**72.7**	**82.9**	**79.5**
5	150	✘	72.7	82.8	79.5

Table 3. Ablation studies on different update intervals on LaSOT *test* set.

Interval	5	10	20	30
AUC(%)	72.7	72.7	**72.7**	72.6
P_{Norm}(%)	82.9	82.9	**82.9**	82.5
P(%)	79.5	79.5	**79.5**	79.0

dealing with scenarios involving partial occlusion, full occlusion, motion blur, and scale variation, our method surpasses the second best method by **+1.5%**, **+2.7%**, **+1.4%**, and **+1.1%** AUC, respectively. The results in Figure 1 demonstrate that our proposed HIPTrack maintains a high level of tracking accuracy and exhibits strong robustness in scenarios involving target appearance variations.

Furthermore, when the target goes out of view, our proposed HIPTrack also exhibits a performance improvement of **+1.1%** AUC compared to the second best method, which means that our proposed HIPTrack has a strong ability to rediscover the lost target. Figure 1 also includes the comparisons of the success and precision curves between our proposed HIPTrack and other approaches across the entire LaSOT *test* split. Our method achieves the highest performance in both two metrics.

3. More Qualitative Results

3.1. Tracking Results

In order to visually highlight the advantages of our method over existing approaches in challenging scenarios, we provide additional visualization results in Figure 2. All videos are from the *test* split of LaSOT. We compare our proposed HIPTrack with GRM [6] and SeqTrack [1] in terms of per-

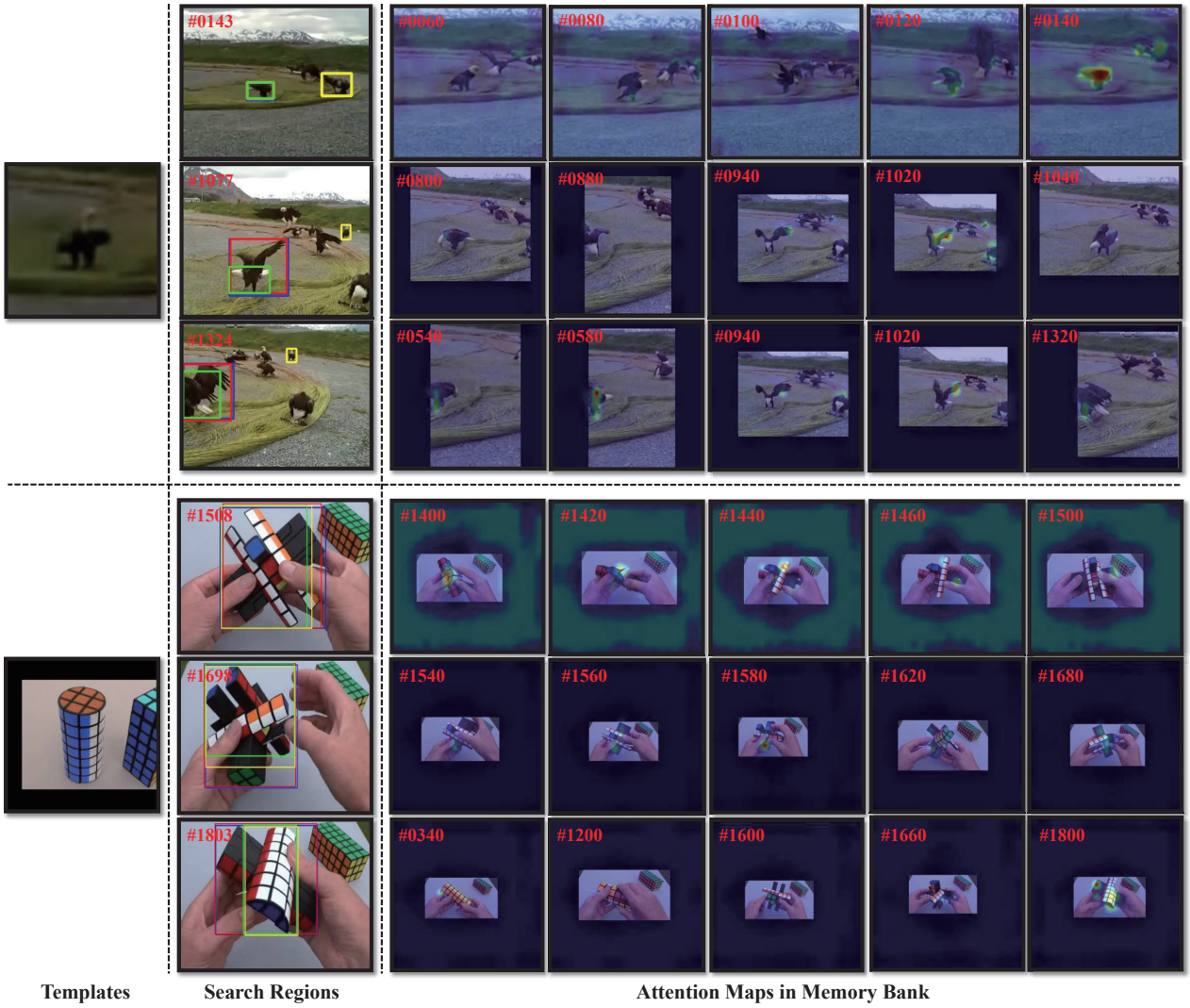


Figure 4. Visualization results of memory bank attention maps across different tracking frames in each video. We select the top 5 memory frames with the highest overall attention weights for visualization, arranging them in chronological order. Zoom in for a clearer view.

formance when the target undergoes deformation, occlusion, and scale variation. All the selected video segments are challenging , as described below:

- Figure 2(a) demonstrates the tracking results of three methods when the target suffers large deformations.
- Figure 2(b) demonstrates the tracking results of three methods when the target suffers partial occlusions.
- Figure 2(c) demonstrates the tracking results of three methods when the target suffers large scale variations.

3.2. Refined Foreground Masks

Our proposed historical prompt encoder utilizes the candidate elimination (CE) module within the feature extraction network to filter out background image patches to construct a CE Mask. The Bounding box mask is created based on the predicted bounding box of the current frame. These two masks are then combined using bitwise and operation, resulting in a refined target foreground mask. To investigate whether the refined foreground mask accurately captures the position information of the target, we visualize the predicted bounding boxes, CE Masks, and refined foreground masks in Figure 3. The visualization results in Figure 3 demonstrate that the refined foreground mask effectively filters out the majority of background regions, providing a more precise depiction of the position information of the target.

3.3. Attention Maps in Memory Bank

In the main body of this paper, we present visualization results of a subset of memory bank attention maps. In Figure

4, we further illustrate the visualization results of memory bank attention maps across different tracking frames within the same video.

As shown in Figure 4, the first row of results in the first video demonstrates that when the target has not undergone significant deformations or scale changes in recent frames, the most recent memory frame receives noticeably higher attention. However, the second and third rows in the first video indicate that when the target undergoes drastic changes in appearance, the historical prompt decoder directs attention towards earlier historical memory frames, thereby enhancing the prediction accuracy of the tracker.

In the second video, when the target undergoes severe deformations, the historical prompt decoder adaptively directs attention to the boundary regions of the target within the historical memory frames, leading to a significant improvement in the precision of boundary prediction. A similar phenomenon can also be observed in the second row of the first video.

References

[1] Xin Chen, Houwen Peng, Dong Wang, Huchuan Lu, and Han Hu. Seqtrack: Sequence to sequence learning for visual object tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14572–14581, 2023. 2, 3, 4

[2] Yutao Cui, Cheng Jiang, Limin Wang, and Gangshan Wu. Mixformer: End-to-end tracking with iterative mixed attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13608–13618, 2022. 2

[3] 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*, pages 5374–5383, 2019. 1, 2

[4] Zhihong Fu, Zehua Fu, Qingjie Liu, Wenrui Cai, and Yunhong Wang. Sparsett: Visual tracking with sparse transformers. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 905–912, 2022. 2

[5] 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

[6] Shenyuan Gao, Chunluan Zhou, and Jun Zhang. Generalized relation modeling for transformer tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18686–18695, 2023. 2, 3, 4

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 1

[8] Kaijie He, Canlong Zhang, Sheng Xie, Zhixin Li, and Zhiwen Wang. Target-aware tracking with long-term context attention. 2023. 2

[9] Christoph Mayer, Martin Danelljan, Danda Pani Paudel, and Luc Van Gool. Learning target candidate association to keep track of what not to track. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13444–13454, 2021. 2

[10] Botao Ye, Hong Chang, Bingpeng Ma, Shiguang Shan, and Xilin Chen. Joint feature learning and relation modeling for tracking: A one-stream framework. In *European Conference on Computer Vision*, pages 341–357. Springer, 2022. 2