Supplementary Material for SPMTrack: Spatio-Temporal Parameter-Efficient Fine-Tuning with Mixture of Experts for Scalable Visual Tracking

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 the supplementary material, Section A presents additional ablation studies, further validating the effectiveness of other designs of SPMTrack and TMoE. In Section B, we provide more comprehensive performance comparisons between our SPMTrack and other excellent trackers. The comparisons consist of success curves on both overall LaSOT [6] *test* split and all challenging scenario subsets, along with overall precision curves. And in Section C, we further showcase the comparisons of tracking results of various trackers in complex scenarios and present more qualitative analyses of our method.

A. Further Analyses

A.1. Ablation Study on Target State Token

We employ a temporally propagated target state token to store historical target states, which is utilized in the prediction head for further adjustment and refinement of search region features. In Table A.1 and Table A.2, we remove the target state token and directly use the search region features from the output of feature extraction network for prediction in the prediction head. Table A.1 and Table A.2 present evaluation results on the GOT-10K [9] and TrackingNet [11] *test* splits, respectively. The results demonstrate that the integration of target state tokens, while introducing negligible additional parameters, improves performance on both two datasets, with a more significant improvement observed on TrackingNet.

Table A.1. Ablation study on whether adopt target state token. Results are evaluated on GOT-10K [9] *test* split.

model variants	#Params(M)	AO (%)	$SR_{0.5}$(%)	$SR_{0.75}$(%)
SPMTrack-B	115.331	**76.5**	**85.9**	**76.3**
w/o target state token	**115.329**	76.3	85.6	75.9

*Corresponding author.

Table A.2. Ablation study on whether adopt target state token. Results are evaluated on TrackingNet [11] *test* split.

model variants	#Params(M)	AUC (%)	P_{Norm}(%)	P(%)
SPMTrack-B	115.331	**86.1**	**90.2**	**85.6**
w/o target state token	**115.329**	85.7	89.9	85.1

A.2. Ablation Study on Token Type Embedding

In SPMTrack, we add not only positional embedding but also token type embeddings to the input of the feature extraction network to further enhance the token information. The additional parameter count and computational overhead introduced by the token type embeddings are negligible.

In Table A.3, we investigate the impact of introducing different token type embeddings on the performance of SPMTrack-B. Comparing the first row and the third row in Table A.3, adding three types of token type embeddings leads to a performance improvement. Comparing the first row and second row, we find that when the type embedding corresponding to the target foreground tokens is not introduced, which means all tokens in the reference frames use the same type embedding, there are no performance gains observed. The results indicate that the target foreground token type embedding is the most crucial among the three categories, as it enhances the ability of the tracker to discriminate target foreground regions, thereby improving performance.

Table A.3. Ablation study on different token type embeddings, where $\boldsymbol{TE}_o$, $\boldsymbol{TE}_b$ and $\boldsymbol{TE}_S$ represent the type embeddings of the foreground region tokens, background region tokens in reference frames and the search region tokens, respectively.

#	$\boldsymbol{TE}_o$	$\boldsymbol{TE}_b$	$\boldsymbol{TE}_S$	AO (%)	$SR_{0.5}$(%)	$SR_{0.75}$(%)
1	✘	✘	✘	76.2	85.6	75.8
2	✘	✔	✔	76.2	85.5	76.0
3	✔	✔	✔	**76.5**	**85.9**	**76.3**

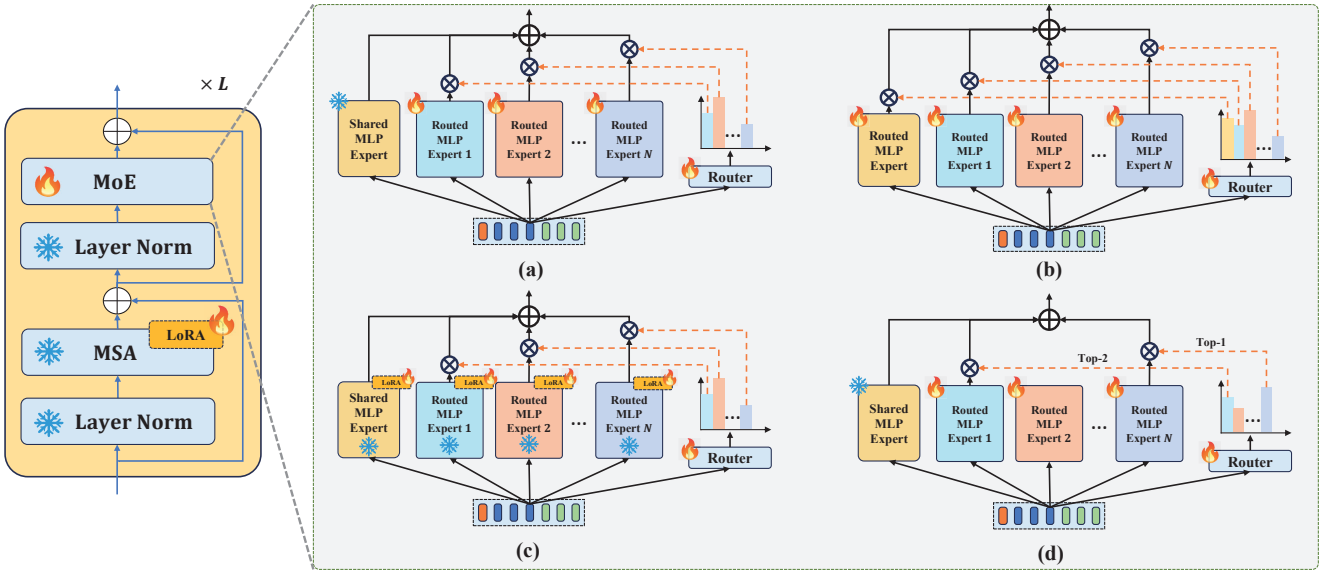


Figure A.1. Several conventional MoE architectural designs, where MoE modules exclusively replace feed-forward network (FFN) layers in the original Transformer Block, while linear layers in multi-head self-attention are fine-tuned using LoRA [8]. Zoom in for better view.

A.3. Ablation Study on Different Pre-trained Models

In this paper, we select DINOv2 [12] as the pre-trained model. In Table A.4, we investigate the impact of different pre-trained models on the final performance. All experiments are conducted based on SPMTrack-B, with pre-trained models based on the ViT-B [5] architecture. It is worth noting that for pre-trained models with an image patch size of 16, we use a reference frame size of 192×192 and a search region size of 384×384. As shown in Table A.4, experimental results demonstrate that SPMTrack maintains consistent performance on GOT-10K regardless of the choice of pre-trained model, indicating the robustness and the generalization ability of our approach. Eventually, in order to make a fair comparison with LoRAT [10] that also employs parameter-efficient fine-tuning, we still select DINOv2 [12] as the pre-trained model.

Table A.4. Ablation study on different pre-trained models. Results are evaluated on GOT-10K *test* split.

Pretrained Models	Patch Size	AO (%)	$SR_{0.5}$(%)	$SR_{0.75}$(%)
DINOv2 [12]	14	76.5	85.9	**76.3**
MAE [7]	16	76.3	87.0	75.1
DropMAE [14]	16	**76.6**	**87.7**	74.0

A.4. Ablation Study on MLP Prediction Head

In SPMTrack, in order to make a fair comparison with LoRAT [10], we also adopt a fully MLP-based prediction head. However, this does not represent our optimal configuration. As demonstrated in Table A.5, we replace the MLP prediction head with the convolution-based prediction head used in previous trackers like OSTrack [16]. Although the MLP prediction head has fewer parameters, the overall model parameter difference is minimal, and the convolution-based head achieves significantly superior performance. In contrast to LoRAT, which fails completely under parameter-efficient fine-tuning when using convolution-based heads, our method achieves even better performance with convolution-based heads, which further demonstrates the generality of our proposed TMoE for parameter-efficient fine-tuning and also further demonstrates the potential for performance improvement in SPMTrack.

Table A.5. Ablation study on MLP-based prediction head and convolutional-based prediction head. Results are evaluated on GOT-10K *test* split.

Prediction Head	#Params(M)	AO (%)	$SR_{0.5}$(%)	$SR_{0.75}$(%)
MLP Head	**2.37**	76.5	85.9	76.3
OSTrack [16] Conv Head	6.47	**76.8**	**86.4**	**77.0**

A.5. More Comparisons with Conventional MoE

In Table 7 of the main manuscript, we compare our method with conventional MoE in terms of parameter count and performance on LaSOT, the conventional MoE structure in Table 7 is illustrated in Figure A.1(a). Additionally, as shown in Figure A.1, we explore various alternative MoE structures. Figure A.1(b) shows a modification of the architecture in Figure A.1(a) where we replace the frozen shared

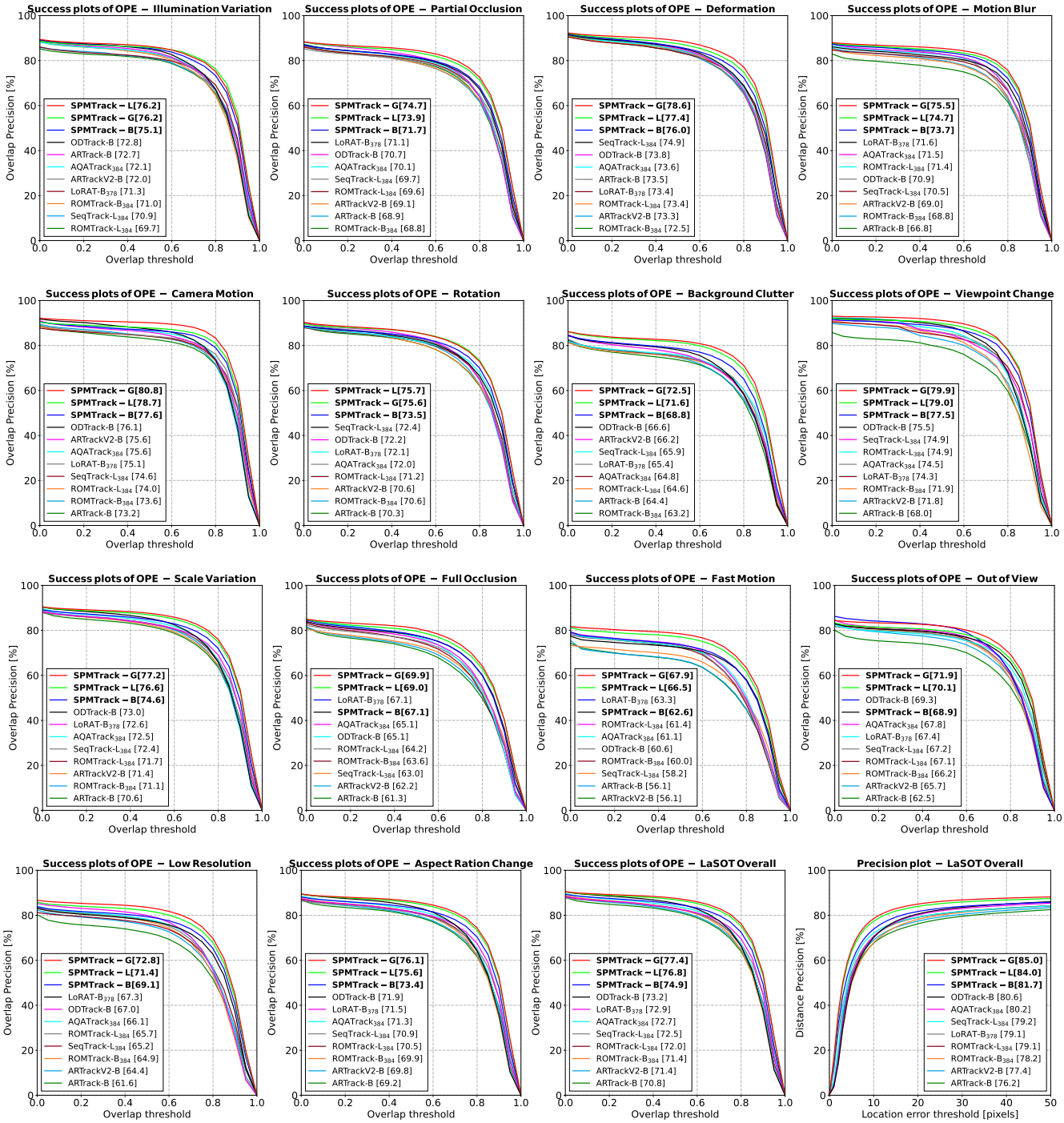
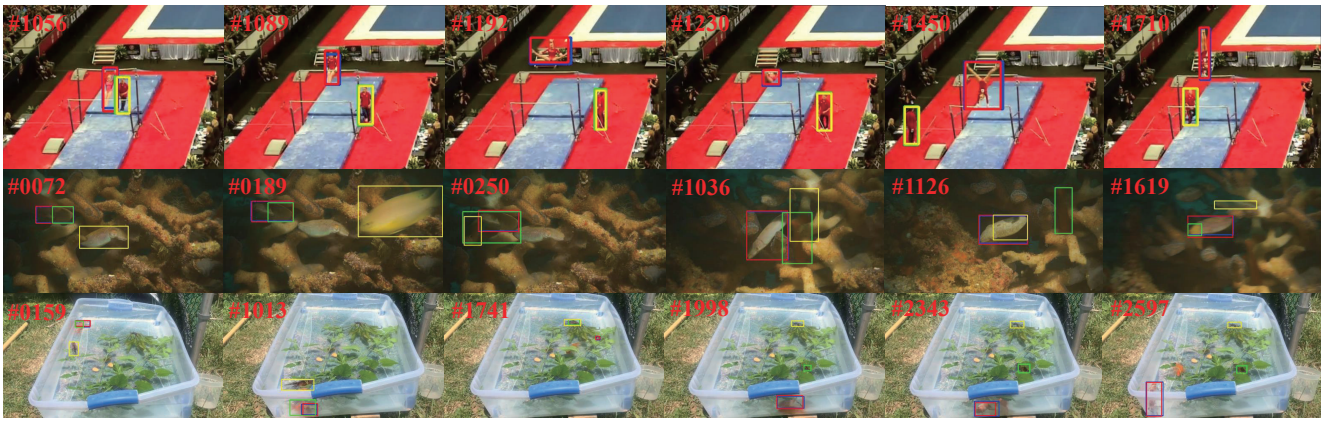
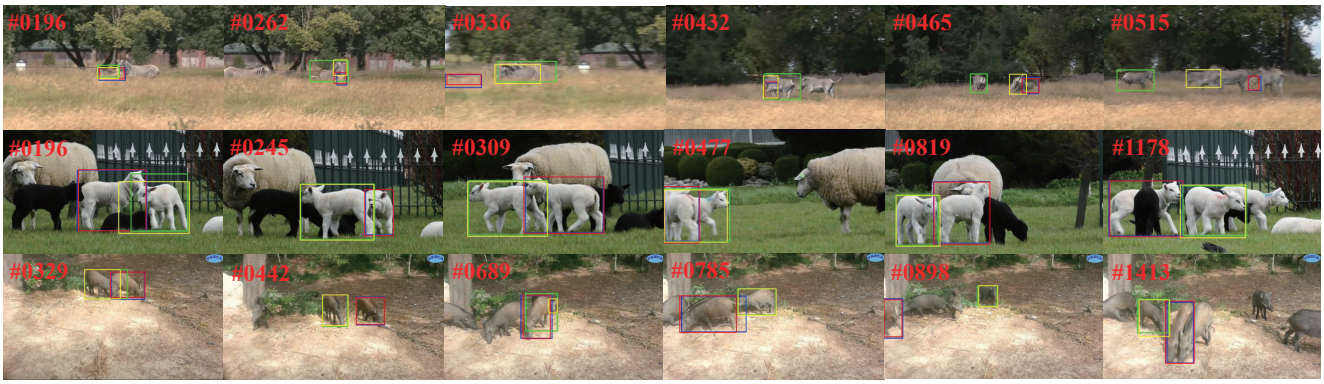


Figure A.2. Comparisons of our proposed SPMTrack with other excellent trackers in the success curve on LaSOT *test* split, which includes eleven challenging scenarios such as Low Resolution, Motion Blur, Scale Variation, etc. We also provide the comparisons of the success and precision curves across the entire LaSOT *test* split. Zoom in for better view.

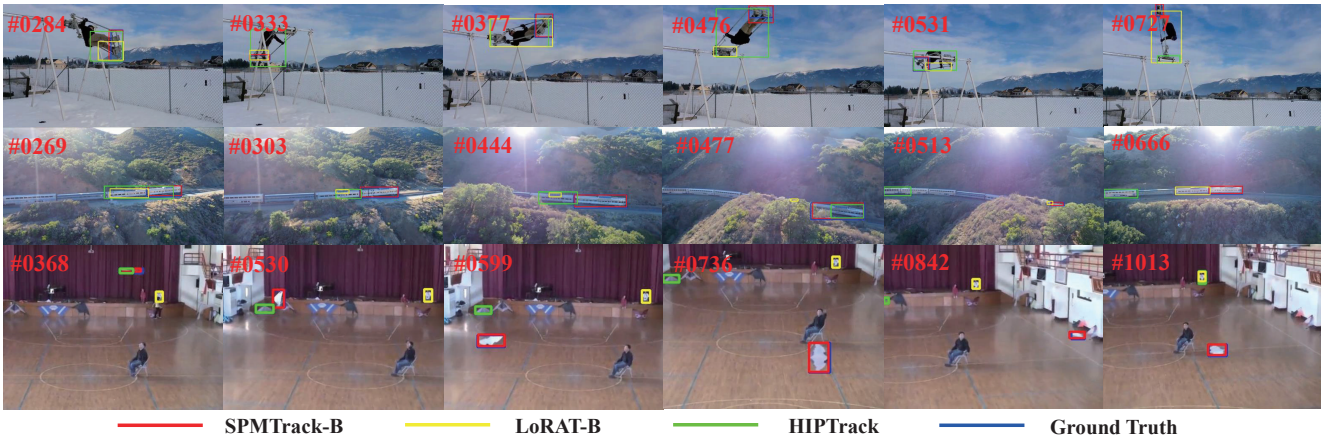
expert with a learnable routed expert and initialize all five routed experts with corresponding FFN weights from the pre-trained model. Figure A.1(c) extends the design in Figure A.1(a) by freezing all experts and fine-tuning each expert using LoRA [8]. In Figure A.1(d), we adopt the way of sparsely activating experts with top-k routing scores on the basis of Figure A.1(a). The experimental results of these different MoE architectures on LaSOT *test* split are pre-



(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 A.3. This figure presents a visual comparison among our proposed SPMTrack-B, LoRAT-B [10] and HIPTrack [2] 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.

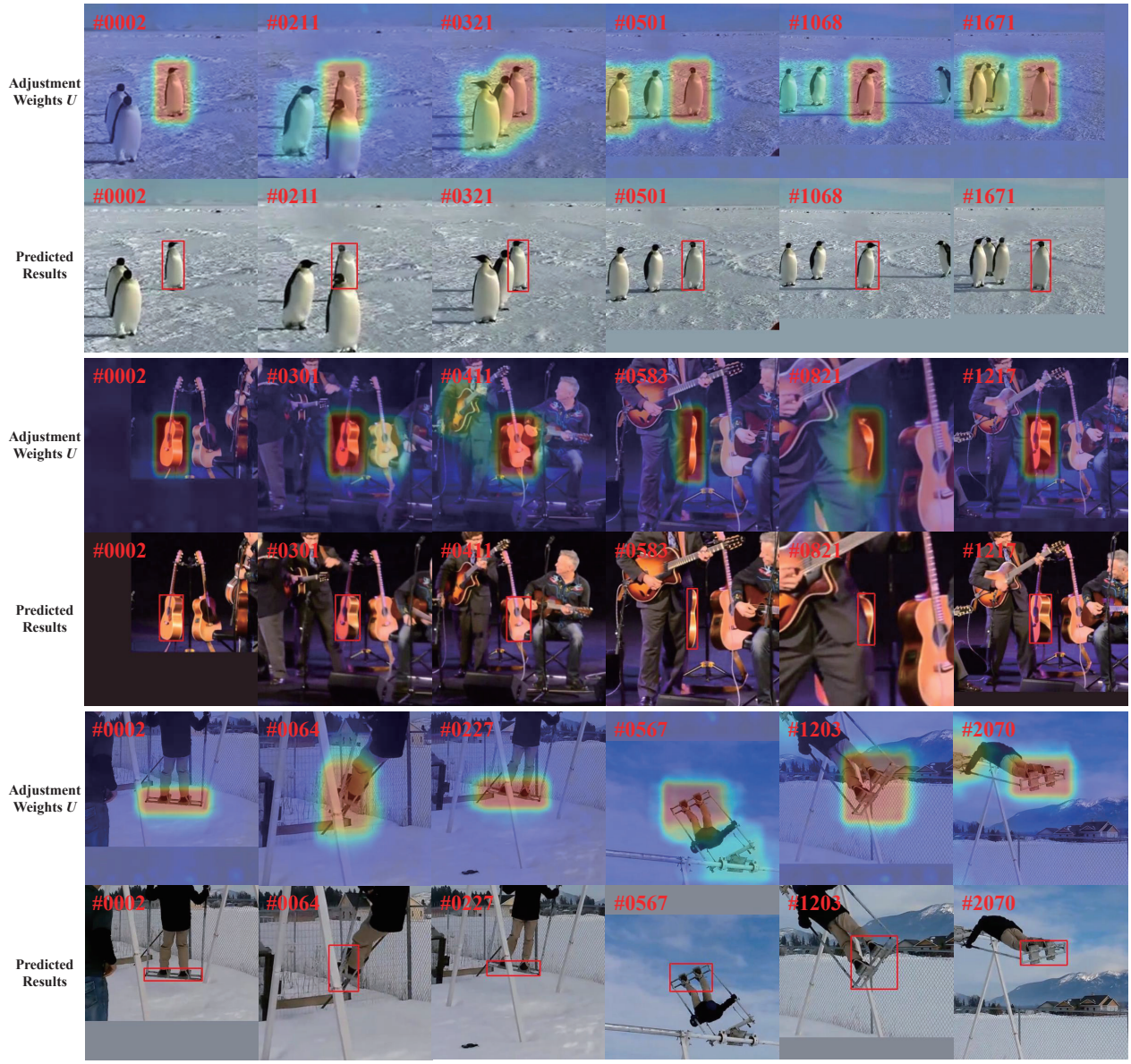


Figure A.4. Visualization of the search region feature adjustment weights $\boldsymbol{U}$ and the corresponding predicted bounding boxes.

sented in Table A.6.

As shown in Table A.6, in row 1, the MoE comprises one shared expert and four routed experts, with all experts participating in computation, hence the expert count is $1+4$ as well as the activated expert count. Row 3 follows the same rule. In row 2, there are only five routed experts, and all of them participate in the computation. In row 4, the MoE includes a total of one shared expert and four routed experts, but only the routed experts with the top-2 routing scores are activated, so the number of activated experts is $1 + 2$. In row 5, the MoE includes a total of one shared experts and eight routed experts, and the routed experts with the top-4 routing scores are activated.

Table A.6 demonstrates that TMoE achieves superior performance compared to all conventional MoE approaches that are only applied to replace FFN layers, while maintaining significantly lower total parameters (all conventional MoE variants exceed 300M parameters). At the same time, comparing row 2 with other rows, we observe that preserve pre-trained weights in MoE leads to substantially improved performance. This explains why row 3 achieves optimal performance among all conventional MoE approaches. Comparing rows 1, 4, and 5, we find that increasing the overall parameter count or the number of activated experts

results in only marginal performance gains, further validating the effectiveness of our TMoE design.

However, in the field of natural language processing, sparsely activated MoE has become mainstream. But sparsely activated MoE requires more engineering optimizations. Due to the limitations of time and resources, we have not further explored the performance of TMoE under sparse activation. We also hope that this work can inspire other researchers to conduct related explorations in the field of visual tracking.

Table A.6. Performance comparison between different conventional MoE approaches and TMoE. The number of experts is represented in the form of $\boldsymbol{a} + \boldsymbol{b}$, where $\boldsymbol{a}$ represents shared experts and $\boldsymbol{b}$ represents routed experts. All results are evaluated on LaSOT *test* split.

MoE Variants	#MLP Experts	#Activated Experts	AUC (%)	P_{Norm}(%)	P(%)
Figure A.1(a)	1+4	1+4	73.4	82.6	79.8
Figure A.1(b)	5	5	72.7	81.5	78.7
Figure A.1(c)	1+4	1+4	74.4	83.4	80.6
Figure A.1(d)	1+4	1+2	73.3	82.0	79.4
	1+8	1+4	73.8	82.8	79.9
TMoE	-	-	**74.9**	**84.0**	**81.7**

B. More Detailed Results in Different Attribute Scenes on LaSOT

In Figure A.2, we provide a more detailed comparison of our method with other current state-of-the-art trackers LoRAT [10], ODTrack [17], ARTrackV2 [1], AQATrack [15], ARTrack [13], ROMTrack [3] and SeqTrack [4] across various challenging scenario subsets in LaSOT [6]. Figure A.2 presents detailed success curves and AUC scores across individual subsets, along with the success and precision curves on the entire LaSOT *test* split. The results demonstrate that SPMTrack-B significantly outperforms current state-of-the-art approaches both overall and across the vast majority of subsets.

C. More Qualitative Results

C.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 A.3. All videos are from the *test* split of LaSOT. We compare our proposed SPMTrack-B with HIPTrack [2] and LoRAT-B [10] in terms of performance when the target undergoes deformation, occlusion, and scale variation. All the selected videos are challenging, as described below:

- Figure A.3(a) demonstrates the tracking results of three methods when the target suffers deformations.
- Figure A.3(b) demonstrates the tracking results of three methods when the target suffers partial occlusions.
- Figure A.3(c) demonstrates the tracking results of three methods when the target suffers scale variations.

C.2. Visualization of Search Region Feature Adjustment Weight

In the prediction head of SPMTrack, we compute matrix multiplication between the output $\boldsymbol{H}'$ corresponding to target state token and the output search region feature $\boldsymbol{X}'$ to obtain weight $\boldsymbol{U}$. The obtained weight $\boldsymbol{U}$ contains the historical state information of the target and is used to further adjust and refine the output search region feature $\boldsymbol{X}'$. In Figure A.4, we visualize the weight $\boldsymbol{U}$. The visualization results demonstrate that the adjustment weight $\boldsymbol{U}$ can significantly distinguish between the target foreground region and irrelevant background regions. Moreover, the heatmaps exhibit patterns resembling bounding boxes, with substantially higher weights inside the "bounding box". This further enhances the features at potential target locations, thereby improving the foreground-background discrimination capability of the search region features.

References

[1] Yifan Bai, Zeyang Zhao, Yihong Gong, and Xing Wei. Artrackv2: Prompting autoregressive tracker where to look and how to describe. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19048–19057, 2024. 6

[2] Wenrui Cai, Qingjie Liu, and Yunhong Wang. Hiptrack: Visual tracking with historical prompts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19258–19267, 2024. 4, 6

[3] Yidong Cai, Jie Liu, Jie Tang, and Gangshan Wu. Robust object modeling for visual tracking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9589–9600, 2023. 6

[4] 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. 6

[5] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 2

[6] 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, 6

[7] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference*

on Computer Vision and Pattern Recognition (CVPR), pages 16000–16009, 2022. 2

[8] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR*, 2022. 2, 3

[9] Lianghua Huang, Xin Zhao, and Kaiqi Huang. Got-10k: A large high-diversity benchmark for generic object tracking in the wild. *TPAMI*, 2019. 1

[10] Liting Lin, Heng Fan, Zhipeng Zhang, Yaowei Wang, Yong Xu, and Haibin Ling. Tracking meets lora: Faster training, larger model, stronger performance. In *Computer Vision – ECCV 2024*, pages 300–318, Cham, 2025. Springer Nature Switzerland. 2, 4, 6

[11] Matthias Muller, Adel Bibi, Silvio Giancola, Salman Alsubaihi, and Bernard Ghanem. Trackingnet: A large-scale dataset and benchmark for object tracking in the wild. In *ECCV*, pages 300–317, 2018. 1

[12] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *Transactions on Machine Learning Research Journal*, pages 1–31, 2024. 2

[13] Xing Wei, Yifan Bai, Yongchao Zheng, Dahu Shi, and Yihong Gong. Autoregressive visual tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9697–9706, 2023. 6

[14] Qiangqiang Wu, Tianyu Yang, Ziquan Liu, Baoyuan Wu, Ying Shan, and Antoni B. Chan. Dropmae: Masked autoencoders with spatial-attention dropout for tracking tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14561–14571, 2023. 2

[15] Jinxia Xie, Bineng Zhong, Zhiyi Mo, Shengping Zhang, Liangtao Shi, Shuxiang Song, and Rongrong Ji. Autoregressive queries for adaptive tracking with spatio-temporal transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19300–19309, 2024. 6

[16] 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

[17] Yaozong Zheng, Bineng Zhong, Qihua Liang, Zhiyi Mo, Shengping Zhang, and Xianxian Li. Odtrack: Online dense temporal token learning for visual tracking. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(7):7588–7596, 2024. 6