OneTracker: Unifying Visual Object Tracking with Foundation Models and Efficient Tuning
Supplementary Materials

1. Discussion

To the best knowledge of ours, we are the first to unify visual object tracking in a general framework. Although there exists some works [1, 10, 13, 16, 18, 19] which tackle multiple tracking tasks in a single model, these works only consider RGB modality and ignore multimodal informtaion. Moreover, some methods [21, 23] take multimodal information into consideration, but they only focus on specific modalities and still treat RGB and RGB+X tracking as separate entities. We consider these two tasks as a unified whole. Our work unifies several tracking tasks, RGB tracking and RGB+N/D/T/E/M tracking, and achieves competitive performance on 11 benchmarks across the 6 tasks.

Diverging from conventional approaches that perform full finetuning on downstream datasets, we break the widely-used full finetuning manner, and introduce the parameter-efficient transfer learning (PETL), which is popular in NLP, into tracking. In NLP, a large-scale foundation model is trained on broad data and owns a strong logical reasoning and generative ability. Then, PETL techniques are adopted to transfer foundation model to downstream tasks by freezing the pretrained weights and training inserted parameters. Due to the similar temporal matching mechanisms in RGB and RGB+X tracking tasks, we follow the large-scale training and PETL manner in NLP. Our framework begins with the pretraining of Foundation Tracker on large-scale RGB tracking datasets, enabling it to acquire a strong temporal matching ability. After that, we incorporate multimodal information as prompt and introduce CMT Prompters to enhance Prompt Tracker with multimodal features, boosting overall performance. Despite similar structure is discussed in ProTrack and ViPT, they do not take language and mask into account. Besides, TTP Transformer layers are utilized to adapt Prompt Tracker to downstream tasks better. Through adjusting a set of additional parameters (about 2.8M), Prompt Tracker inhibits the strong ability from Foundation Tracker, and have better adaptability than full finetuning. Importantly, the parameter efficiency makes it particularly suitable for resource-constrained devices where only a small number of parameters need to be distributed to end-side deployments for the generalization of downstream scenarios.

Limitations. Despite the high effectiveness and efficiency, our framework still has some limitations. Firstly, for different tracking tasks within the RGB+X domain, our Prompt Tracker still needs to be trained on specific datasets separately. This implies that if we want to handle multiple RGB+X tracking tasks, we need to adjust the parameters of the CMT Prompters and TTP Transformer layers accordingly. Although the parameters of these two modules are iightweight and can be almost negligible, it still results in inconvenience. Exploring methods to handle multiple tasks within a general model through joint training is an important direction for future research. Secondly, although our model is capable of handling 6 tracking tasks across various modalities, there are still other modalities that have not been considered. We will continue to extending our model to more downstream tasks. Thirdly, as the landscape of downstream RGB+X tasks evolves, it is crucial to make our Prompt Tracker adaptive to new tasks while maintaining its original capabilities. Ensuring the flexibility of our framework to accommodate emerging tasks without sacrificing its performance on existing tasks is an important challenge that requires further investigation. Addressing these limitations will contribute to the continuous development and improvement of our framework, making it more versatile, adaptable, and effective for a broader range of tracking tasks.

2. Experiment Details

2.1. Foundation Tracker Training

Foundation Tracker are trained on a combination of several RGB tracking and object detection datasets, including LaSOT [3], TrackingNet [11], GOT-10K [5], and COCO [9], following [2, 4, 22]. We only used the training sets of these dataset for training.Data augmentations, such as horizontal flip and brightness jittering, are adopted during training.

Compared to previous trackers, the training datasets and setting, such as the number of training epochs, remain consistent. Despite the **same** training setting, our Foundation Tracker achieves superior performance, outperforming

Table 1. Training setting for Foundation Tracker on RGB traking datasets.

Config	Value
optimizer	AdamW [6]
learning rate in head	4×10^{-4}
learning rate in backbone	4×10^{-5}
weight decay	10^{-4}
batch size	128
epoch	300
learning rate decay epoch	240
learning rate decay factor	10
learning rate schedule	steplr
maximum sampling frame gap	200

Table 2. Finetuning setting for Prompt Tracker on RGB+N/D/T/E tracking.

Config	Value
optimizer	AdamW [6]
learning rate	4×10^{-5}
weight decay	10^{-4}
batch size	128
epoch	60
learning rate decay epoch	48
learning rate decay factor	10
learning rate schedule	steplr
maximum sampling frame gap	200

Table 3. Finetuning setting for Prompt Tracker on RGB+M tracking datasets.

Config	Value
optimizer	AdamW [6]
base learning rate	1×10^{-5}
weight decay	1×10^{-7}
batch size	8
Iterations	150,000
learning rate decay iteration	125,000
learning rate schedule	steplr
maximum sampling frame gap	25

other trackers by at least 0.6 AUC on LaSOT. Models like UNINEXT and OmniTracker, which aim to address multiple vision tasks, utilze a larger set of datasets in addition to RGN tracking dataset6s. The training of UNINEXT and OmniTracker require significantly more time and GPU resources, typically taking several days and utilizing more GPUs. In contrast, our Foundation Tracker can be trained in about one day on 4 NVIDIA RTX 3090 GPUs. Compared to these models which required much more training data and training cost than our Foundation Tracker, our Foundation Tracker achieves better performance on RGB tracking (at least 1.3 AUC on LaSOT). Considering that our Foundation Tracker achieves better performance on RGB tracking while utilizing the **same or smaller** amount of training data and computational resources compared to other models, the comparison on LaSOT and TrackingNet benchmarks is both fair and favourable to our approach.

2.2. Prompt Tracker Finetuning

RGB+N/D/T/E tracking. For the parameter-efficient finetuning of Prompt Tracker on downstream RGB+X tracking tasks, we finetune Prompt Tracker on each task separately. The size for template and search frame is 192×192 and 384×384. For RGB+N tracking, we adopt OTB99 [8], LaSOT [3], and TNL2K [15] as training sets. For RGB+D tracking, DepthTrack [20] is chosen for training. For RGB+T tracking, LasHeR [7] is utilized for training. For RGB+E tracking, VisEvent [14] is leveraged for training. The hyper-parameters are in Table 2.

RGB+M tracking. We choose the popular RGB+M tracking datasets, DAVIS17 [12] and YouTube-VOS [17] for finetuning. We select the first frame and previous frame as template frame, and do not implement any cropping operation on the template and search frames. The finetuning details are in Table 3

References

[1] Ali Athar, Alexander Hermans, Jonathon Luiten, Deva Ramanan, and Bastian Leibe. Tarvis: A unified approach for target-based video segmentation. *arXiv preprint arXiv:2301.02657*, 2023. 1

[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*, pages 13608–13618, 2022. 1

[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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5374–5383, 2019. 1, 2

[4] Shenyuan Gao, Chunluan Zhou, and Jun Zhang. Generalized relation modeling for transformer tracking. *arXiv preprint arXiv:2303.16580*, 2023. 1

[5] Lianghua Huang, Xin Zhao, and Kaiqi Huang. 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

[6] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 2

[7] Chenglong Li, Wanlin Xue, Yaqing Jia, Zhichen Qu, Bin Luo, Jin Tang, and Dengdi Sun. Lasher: A large-scale high-diversity benchmark for RGBT tracking. *IEEE Transactions on Image Processing*, 31:392–404, 2021. 2

[8] Zhenyang Li, Ran Tao, Efstratios Gavves, Cees GM Snoek, and Arnold WM Smeulders. Tracking by natural language specification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6495–6503, 2017. 2

[9] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *ECCV*, 2014. 1

[10] Fan Ma, Mike Zheng Shou, Linchao Zhu, Haoqi Fan, Yilei Xu, Yi Yang, and Zhicheng Yan. Unified transformer tracker for object tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8781–8790, 2022. 1

[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 *Proceedings of the European conference on computer vision (ECCV)*, pages 300–317, 2018. 1

[12] Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alex Sorkine-Hornung, and Luc Van Gool. The 2017 davis challenge on video object segmentation. *arXiv preprint arXiv:1704.00675*, 2017. 2

[13] Junke Wang, Dongdong Chen, Zuxuan Wu, Chong Luo, Xiyang Dai, Lu Yuan, and Yu-Gang Jiang. Omnitracker: Unifying object tracking by tracking-with-detection. *arXiv preprint arXiv:2303.12079*, 2023. 1

[14] Xiao Wang, Jianing Li, Lin Zhu, Zhipeng Zhang, Zhe Chen, Xin Li, Yaowei Wang, Yonghong Tian, and Feng Wu. Visevent: Reliable object tracking via collaboration of frame and event flows. *arXiv preprint arXiv:2108.05015*, 2021. 2

[15] Xiao Wang, Xiujun Shu, Zhipeng Zhang, Bo Jiang, Yaowei Wang, Yonghong Tian, and Feng Wu. Towards more flexible and accurate object tracking with natural language: Algorithms and benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13763–13773, 2021. 2

[16] Zhongdao Wang, Hengshuang Zhao, Ya-Li Li, Shengjin Wang, Philip Torr, and Luca Bertinetto. Do different tracking tasks require different appearance models? *Advances in Neural Information Processing Systems*, 34:726–738, 2021. 1

[17] Ning Xu, Linjie Yang, Yuchen Fan, Jianchao Yang, Dingcheng Yue, Yuchen Liang, Brian Price, Scott Cohen, and Thomas Huang. Youtube-vos: Sequence-to-sequence video object segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 585–601, 2018. 2

[18] Bin Yan, Yi Jiang, Peize Sun, Dong Wang, Zehuan Yuan, Ping Luo, and Huchuan Lu. Towards grand unification of object tracking. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXI*, pages 733–751. Springer, 2022. 1

[19] Bin Yan, Yi Jiang, Jiannan Wu, Dong Wang, Ping Luo, Zehuan Yuan, and Huchuan Lu. Universal instance perception as object discovery and retrieval. *arXiv preprint arXiv:2303.06674*, 2023. 1

[20] Song Yan, Jinyu Yang, Jani Käpylä, Feng Zheng, Aleš Leonardis, and Joni-Kristian Kämäräinen. Depthtrack: Unveiling the power of rgbd tracking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10725–10733, 2021. 2

[21] Jinyu Yang, Zhe Li, Feng Zheng, Ales Leonardis, and Jingkuan Song. Prompting for multi-modal tracking. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 3492–3500, 2022. 1

[22] Botao Ye, Hong Chang, Bingpeng Ma, Shiguang Shan, and Xilin Chen. Joint feature learning and relation modeling for tracking: A one-stream framework. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXII*, pages 341–357. Springer, 2022. 1

[23] Jiawen Zhu, Simiao Lai, Xin Chen, Dong Wang, and Huchuan Lu. Visual prompt multi-modal tracking. *arXiv preprint arXiv:2303.10826*, 2023. 1