Exploring the Feature Extraction and Relation Modeling For Light-Weight Transformer Tracking

Jikai Zheng[†1,4], Mingjiang Liang[†2], Shaoli Huang[3], and Jifeng Ning[*1,4]

[1] College of Information Engineering, Northwest A&F University, Shaanxi, China
[2] University of Technology Sydney, Sydney, Australia
[3] Tencent AI-Lab, Guangdong, China
[4] Shaanxi Engineering Research Center of Agricultural Information Intelligent Perception and Analysis, Shaanxi, China
jikai_zheng@163.com, mingjiang.liang@student.uts.edu.au, shaolihuang@tencent.com, jf_ning@sina.com

Appendix

A.1 More technical details

Pre-trained Backbone Network. We use ViT-Tiny [12], a light-weight visual Transformer pre-trained model, as the backbone network of our tracker. ViT-Tiny is based on ViT-Base [7], changing the dim of the feature after passing through the Patch Embedding layer from C_1=768 to C_2=192, reducing it to 1/4 of the original size. In this way, when $\mathrm{X} \in \mathbb{R}^{\mathrm{B}\times\mathrm{C}\times\mathrm{H}\times\mathrm{W}}$ is input to Patch Embedding, we will get $\mathrm{X}^{'} \in \mathbb{R}^{\mathrm{B}\times\mathrm{C}_2\times\mathrm{H}^{'}\times\mathrm{W}^{'}}$ instead of $\mathrm{X}^{'} \in \mathbb{R}^{\mathrm{B}\times\mathrm{C}_1\times\mathrm{H}^{'}\times\mathrm{W}^{'}}$. This approach undoubtedly speeds up the model, but along with it, the ability of feature extraction also decreases. In order to solve this problem, ViT-Tiny relies on the knowledge of MAE [8] and feature distillation [11, 18] to improve ViT's training method and obtain a light-weight visual model that is comparable to the performance of large models.

Design Basis for Attention Blocks. In the experimental phase, we implemented five Attention Blocks, namely A1 through A5. The core of our design idea is whether it has the ability to extract features or relation modeling. Hence, in this study, we have developed three distinct types of Attention Blocks: A1, which solely focuses on feature extraction; A2 and A3, which incorporate both feature interaction and relation modeling; and A4 and A5, which exclusively emphasize relation modeling. Among them, we have designed two forms for the relationship modeling part, namely one-way relationship modeling that only transfers template information to search and two-way relationship modeling that transfers information in both directions.

[†] Equal Contribution.
[*] Corresponding author.

A.2 More results on TNL2K and OTB100

TNL2K. We evaluate our tracker on TNL2K [13] benchmark, which includes 2000 video sequences, 1244340 frames, and 663 words. Among its 700 test sequences, each video is densely annotated with an English sentence and the corresponding bounding box of the target object. The results of AUC scores are shown in Tab. 1. FERMT not only achieves the best result among real-time trackers, but is also comparable to some high-performance trackers.

OTB100. OTB100 [15] is a widely utilized dataset for reviewing object trackers, comprising 100 test sequences. The video sequences within this dataset encompass diverse target types, including humans, animals, vehicles, and more. These videos encapsulate a range of challenges, such as variations in lighting conditions, complex backgrounds, occlusion of the target, and rapid motion. We report the AUC scores in Tab. 1. Among all real-time trackers, FERMT achieved the best result.

Method	Real-time				None-real-time			
	ATOM [6]	HCAT [2]	MixformerV2-S [5]	**FERMT**	TransT [3]	Stark-ST50 [17]	Ostrack [19]	ARTrack [14]
TNL2K	40.1	-	47.2	**53.3**	50.7	-	55.9	57.5
OTB100	66.9	68.1	-	**68.8**	69.4	67.3	-	-

Table 1: The results of our tracker are compared with state-of-the-art methods on TNL2K and OTB100 benchmarks. We represent our tracker in **bold** font.

A.3 More comparative results of computational complexity.

We have included a comparative analysis of FLOPs and parameters between our FERMT and other leading trackers in the field, as summarized in the below table. Notably, our approach has a slight increase in both FLOPs and parameters compared to lightweight trackers, but boasts substantial performance improvements, even on par with heavy trackers.

Method	LaSOT(AUC)	flops(G)	params(M)
LightTrack (CVPR21) [16]	53.8	0.79	3.13
FEAR-XS (ECCV22) [1]	53.5	1.59	1.37
MixFormerV2-S (NeurIPS23) [5]	60.6	4.40	16.04
HiT-B (ICCV23) [9]	64.6	4.34	42.14
FERMT (Ours)	**65.1**	2.40	7.97
TransT (CVPR21) [3]	64.9	29.3	21.3
Stark-ST50 (ICCV21) [17]	66.6	18.5	42.4
SwinTrack (NeurIPS22) [10]	67.2	69.1	91.0
MixFormer-22k (CVPR22) [4]	69.2	23.04	35.61
ARTrack(CVPR23) [14]	**70.4**	37.15	173.11

References

1. Borsuk, V., Vei, R., Kupyn, O., Martyniuk, T., Krashenyi, I., Matas, J.: Fear: Fast, efficient, accurate and robust visual tracker. In: ECCV (2022)
2. Chen, X., Kang, B., Wang, D., Li, D., Lu, H.: Efficient visual tracking via hierarchical cross-attention transformer. In: ECCV (2022)
3. Chen, X., Yan, B., Zhu, J., Wang, D., Yang, X., Lu, H.: Transformer tracking. In: CVPR (2021)
4. Cui, Y., Jiang, C., Wang, L., Wu, G.: Mixformer: End-to-end tracking with iterative mixed attention. In: CVPR (2022)
5. Cui, Y., Song, T., Wu, G., Wang, L.: Mixformerv2: Efficient fully transformer tracking. In: NeurIPS (2023)
6. Danelljan, M., Bhat, G., Khan, F.S., Felsberg, M.: Atom: Accurate tracking by overlap maximization. In: CVPR (2019)
7. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: ICLR (2021)
8. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: CVPR (2021)
9. Kang, B., Chen, X., Wang, D., Peng, H., Lu, H.: Exploring lightweight hierarchical vision transformers for efficient visual tracking. In: ICCV (2023)
10. Lin, L., Fan, H., Zhang, Z., Xu, Y., Ling, H.: Swintrack: A simple and strong baseline for transformer tracking. In: NeurIPS (2022)
11. Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., Jegou, H.: Training data-efficient image transformers & distillation through attention. In: ICML (2021)
12. Wang, S., Gao, J., Li, Z., Zhang, X., Hu, W.: A closer look at self-supervised lightweight vision transformers. In: ICML (2023)
13. Wang, X., Shu, X., Zhang, Z., Jiang, B., Wang, Y., Tian, Y., Wu, F.: Towards more flexible and accurate object tracking with natural language: Algorithms and benchmark. In: CVPR (2021)
14. Wei, X., Bai, Y., Zheng, Y., Shi, D., Gong, Y.: Autoregressive visual tracking. In: CVPR (2023)
15. Wu, Y., Lim, J., Yang, M.H.: Object tracking benchmark. IEEE TPAMI (2015)
16. Yan, B., Peng, H., Wu, K., Wang, D., Fu, J., Lu, H.: Lighttrack: Finding lightweight neural networks for object tracking via one-shot architecture search. In: CVPR (2021)
17. Yan, B., Penga, H., Fu, J., Wang, D., Lu, H.: Learning spatio-temporal transformer for visual tracking. In: CVPR (2021)
18. Yang, Z., Li, Z., Zeng, A., Li, Z., Yuan, C., Li, Y.: Vitkd: Practical guidelines for vit feature knowledge distillation. In: ICLR (2023)
19. Ye, B., Chang, H., Ma, B., Shan, S., Chen, X.: Joint feature learning and relation modeling for tracking: A one-stream framework. In: ECCV (2022)