A Unified Transformer based Tracker for Anti-UAV Tracking

Qianjin Yu, Yinchao Ma, Jianfeng He,Dawei Yang
University of Science and Technology of China
{sa21010105, imyc}@mail.ustc.edu.cn, {yangdawei,hejf}@mail.ustc.edu.cn

In the supplementary material, we first provide more details for model training and inference. Then, we present more qualitative results to demonstrate the effectiveness of our method.

1. Training and Inference

1.1. More Training Details

In order to evaluate the performance of our tracker on the 1st and 2nd Anti-UAV test-dev [5, 15] benchmarks, we fine-tune our Multi-Region Local Tracking (MRLT) module 50 epochs with 6×10^5 samples per epoch on the train split. The learning rates decrease by a factor of 10 after 40 epochs. The other training settings are consistent with the models trained on entire datasets [3, 4, 7, 9] without UAV classes. Besides, in the Global Detection (GD) module, we fine-tune our global detector 100 epochs on the train split for global detection at each frame. We filter out the drone objects smaller than 5×5 in the train dataset to avoid noise during training.

1.2. Inference Details

Table 1. Effect of different region search strategies on performance.

strategy	Score(%)
single-region search	72.61
multi-region search	75.13

During the tracking process, we try the single-region search and multi-region search strategies in our MRLT module. Specifically, the single-region search means that we only compute the target score in the local tracker search region. Then the target score compares with the global detect scores, and the optimal target location corresponding to the highest score is selected as the final result. In contrast, the multi-region search means that we rescore the detected proposals using the local tracker, which can identify the tracking drone. In particular, we crop search regions based on the detected drone proposals and concatenate these regions in a batch. The local tracker detects targets in these search regions parallelly and outputs the corresponding target score for selecting the optimal result. As shown in Table 1, if we use the multi-region search strategy, the performance raise by 2.97%. This demonstrates that the multi-region search strategy can obtain a more accurate target score for selecting the final result. Thus, our tracker can achieve a robust tracking post-process.

Table 2. Performance of different types of matching algorithms.

matching algorithm	Score(%)
SIFT [8]	73.33
SuperGlue [10]	74.21
LoFTR [11]	75.13

In the process of tracking, we compare different matching algorithms in the Background Correction (BC) module. Specifically, SIFT [8] and Superglue [10] are detector-based local feature matching algorithms. SIFT [8] are arguably the most successful hand-crafted local features and are widely adopted in image feature matching. Recently, SuperGlue [10] proposes a learning-based approach for local feature matching. In contrast, LoFTR [11] proposes to tackle the repeatability issue of feature detectors with a detector-free design. These matching algorithms are popular in many computer vision tasks. As shown in Table 2, the LoFTR model serves as the matching algorithm that can make our UTTracker get the optimal tracking results.

2. Qualitative Results

2.1. Tracking Results

We make a qualitative analysis of our UTTracker under different tracking challenges. As shown in Figure 1, the first row displays an appearance variation challenge,thanks to the MRLT module, our UTTracker can adapt target scale change timely; the second row describes a target out-of-view challenge, by assembling the GD module to our UTTracker, it can be solved effectively; the third row shows a camera motion challenge, with the use of the BC module, our UTTracker can still track the target in the face of camera movement; the final row exposes a small target tracking challenge, and it can be well solved by the DSOD module that contained in our UTTracker. In a word, our UTTracker shows clear advantages in dealing with challenging UAV tracking in TIR mode.

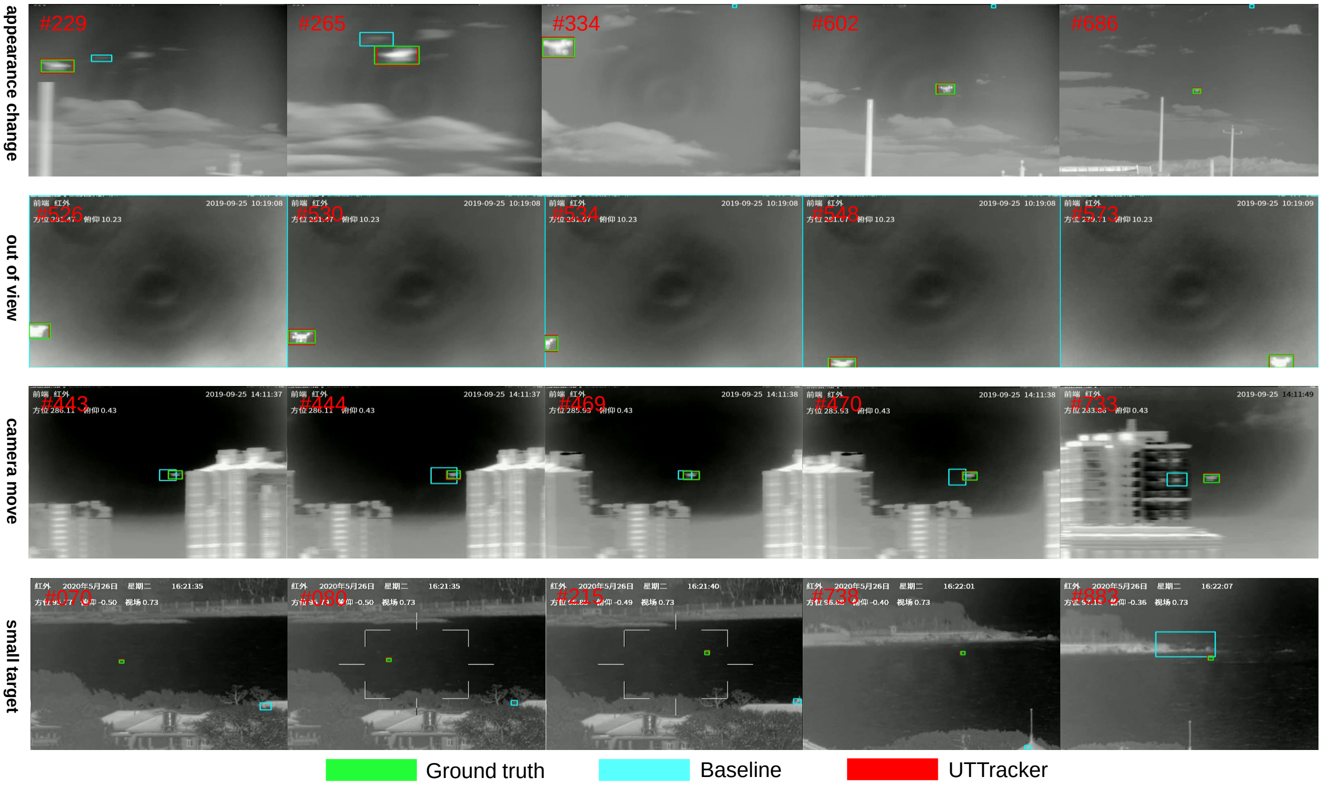

Figure 1. Qualitative comparison of UTTracker with OSTrack Baseline in face of different scenarios. Our tracker can achieve more accurate tracking results in such challenging scenarios.

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.

[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.

[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 IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 1

[4] 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*, 2019. 1

[5] Nan Jiang, Kuiran Wang, Xiaoke Peng, Xuehui Yu, Qiang Wang, Junliang Xing, Guorong Li, Qixiang Ye, Jianbin Jiao, Zhenjun Han, et al. Anti-uav: a large-scale benchmark for vision-based uav tracking. *IEEE Transactions on Multimedia*, 2021. 1

[6] Bo Li, Wei Wu, Qiang Wang, Fangyi Zhang, Junliang Xing, and Junjie Yan. SiamRPN++: Evolution of siamese visual tracking with very deep networks. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[7] 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 *Proceedings of the European Conference on Computer Vision*, 2014. 1

[8] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60:91–110, 2004. 1

[9] 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*, 2018. 1

[10] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superglue: Learning feature matching with graph neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4938–4947, 2020. 1

[11] Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. LoFTR: Detector-free local feature matching with transformers. *CVPR*, 2021. 1

[12] Yinda Xu, Zeyu Wang, Zuoxin Li, Ye Yuan, and Gang Yu. SiamFC++: Towards robust and accurate visual tracking

with target estimation guidelines. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020.

[13] 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.

[14] 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.

[15] Jian Zhao, Gang Wang, Jianan Li, Lei Jin, Nana Fan, Min Wang, Xiaojuan Wang, Ting Yong, Yafeng Deng, Yandong Guo, et al. The 2nd anti-uav workshop & challenge: methods and results. *arXiv preprint arXiv:2108.09909*, 2021. 1