Adaptive and Background-Aware Vision Transformer for Real-Time UAV Tracking-Supplementary Material

Shuiwang Li[1], Yangxiang Yang[1], Dan Zeng[2,], and Xucheng Wang[1]
[1]College of Information Science and Engineering , Guilin University of Technology, China
[2]Research Institue of Trustworthy Autonomous Systems,
Southern University of Science and Technology, China
`lishuiwang0721@163.com, xyyang317@163.com, zengd@sustech.edu.cn, xcwang@glut.edu.cn`

A. Qualitative Results

More qualitative tracking results of Aba-ViTrack and eight top trackers are shown in Fig. 1 as a supplement to those in the main paper. As can be seen, only our tracker successfully tracks the targets in all challenging examples, where pose variations (i.e., in all sequences), background clusters (i.e., Animal1 and uav000088_0000_s), and scale variations (i.e., bike1 and S1701) are presented. Our method performs much better and is more visually pleasing in these cases, further supporting the effectiveness of the proposed method for UAV tracking.

More samples selected from UAVDT [5], VisDrone2018 [21], DTB70 [10], UAV123 [13], and UAVTrack112_L [6] are provided in Fig. 2 to visualize the token's depth that is adaptively controlled during inference with A-ViT* and our Aba-ViTrack, respectively. As can be seen, our background-aware token halting tends to stop background tokens earlier than A-ViT does, which is, therefore, effective in halting distractors and irrelevant tokens and their associated computations for UAV tracking. For example, in the case of animal and person classes, our method basically keeps only the target textures. The examples of cars, boats, and buildings also exhibit similar effects.

B. Comparison with Deep Trackers

Due to page length limit in the main paper, we only compared our method with 20 state-of-the-art deep trackers on DTB70 [10] dataset. In this section, our Aba-ViTrack is compared with five more state-of-the-art deep trackers, i.e., CSWinTT [14], SparseTT [7], SLT-TransT [9], SLT-TrDriMP [9], and ToMP [11], on three more datasets, i.e., UAVDT [5], VisDrone2018 [21], and UAV123@10fps [13]. The precision, AUC, and average FPS are shown in Table 1. The precision (PRC) and AUC are shown in form of (PRC, AUC). As can be seen, our Aba-ViT achieves the best precision on both DTB70 [10] and VisDrone2018 [21], and obtains the best and the second-best AUC on VisDrone2018 [21] and DTB70 [10], respectively. Although our method is significantly lower in precision and AUC compared with TrSiam and TrDimP on UAVDT [5], with differences up to about 5%, our method is close to five times faster than the two method. On UAV123@10fps [13], although our method is significantly inferior to the first-place KeepTrack [12] with margins of 4.7% and 2.7% on precision and AUC, respectively, our method outperforms KeepTrack [12] on both DTB70 [10] and VisDrone2018 [21] and is close to it on UAVDT [5] with difference less than 0.5%. Remarkably, our method is near 9 times faster than KeepTrack [12].

C. Detailed Analysis on Weighting the Proposed Ponder Loss

Detailed Analysis on the weight α_p of the proposed ponder loss $\mathcal{L}^*_{ponder}$ on the performance are evaluated on all six datasets are shown in Table 2. As can be seen, the best precision (PRC) over all datasets is achieved at $\alpha_p = 1.0 \times 10^{-4}$, while the best AUC is also obtained at $\alpha_p = 1.0 \times 10^{-4}$ except for UAV123@10fps [13] and UAV123 [13]. We also observe that the second and third-best performances over all these datasets are distributed both above and below $\alpha_p = 1.0 \times 10^{-4}$ without apparent patterns, which may be explained by the differences between these datasets. The maximal difference of precision is observed on UAVDT [5] by 5.6% and the maximal difference of AUC is observed on UAVDT [5] and UAVTrack112_L [6] by 3.4%, quite significant margins, suggesting that the weight α_p does greatly impact the tracking performance. More specifically, if appropriately weighted, the proposed ponder loss will lead to better tracking performance, otherwise, it may bring bad effects on the tracking task training.

D. Detailed Analysis on Weighting the Background Tokens

Detailed Analysis on the weight ω_b of the ponder loss of background tokens on the tracking performance are eval-

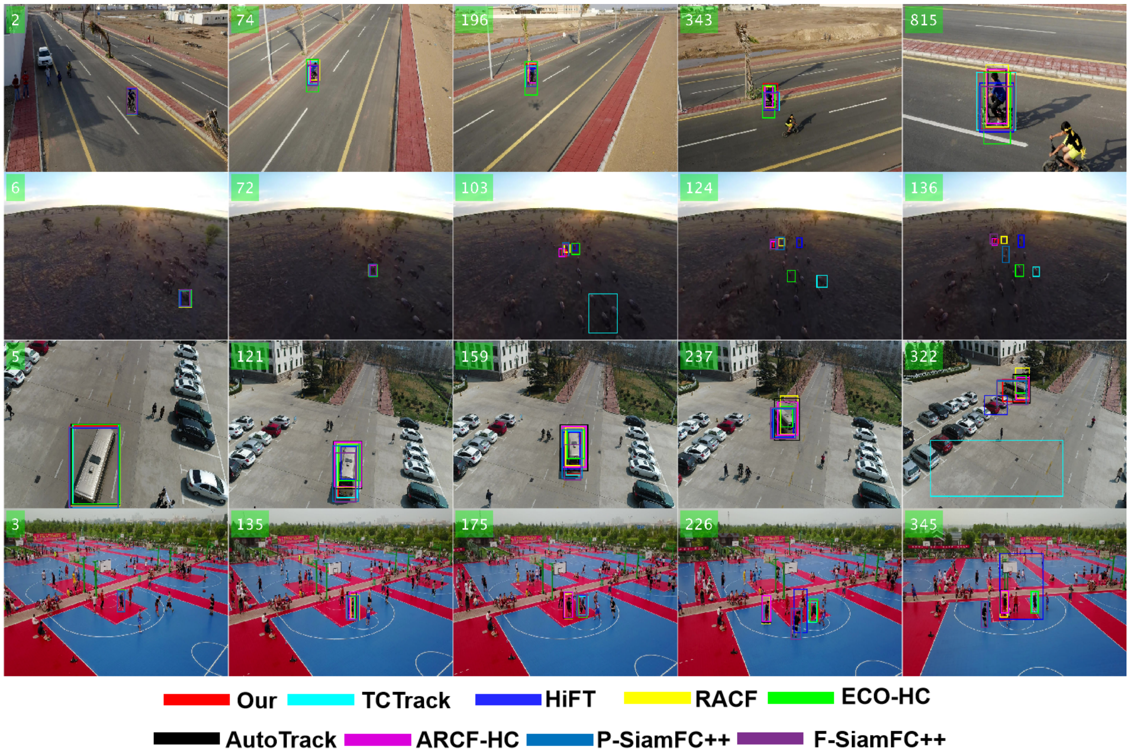

Figure 1. Qualitative evaluation on 4 video sequences from, respectively, UAV123@10fps [13], DTB70 [10], UAVDT [5], and VisDrone2018 [21] (i.e. bike1, Animal1, S1701, and uav000088_0000_s).

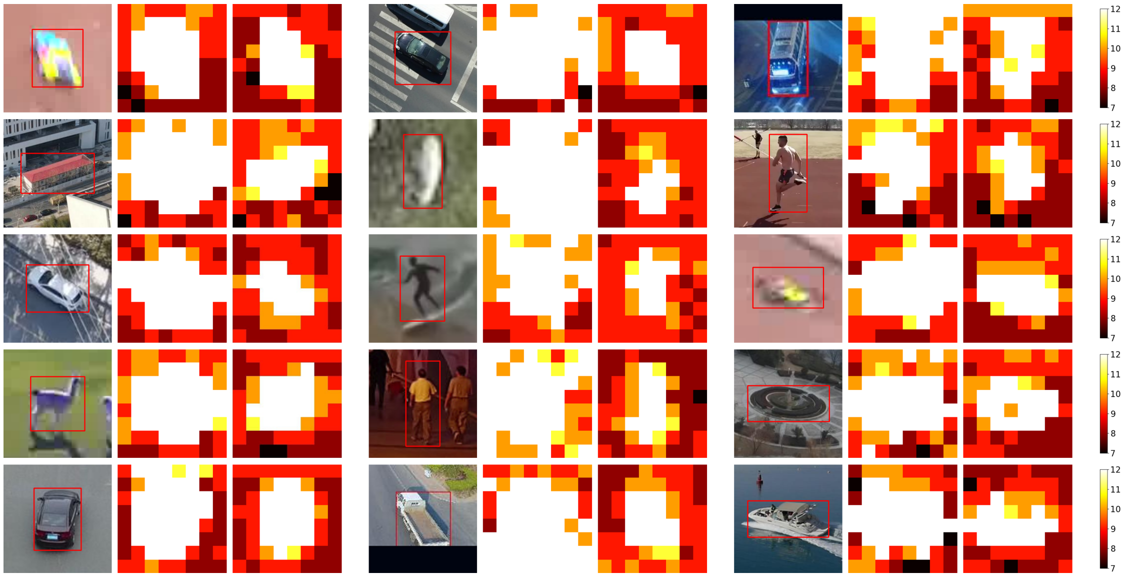

Figure 2. Original image (left), the dynamic token depth of A-ViT* (middle), and that of Aba-ViT (right) on samples from the DTB70 [10], UAVDT [5], VisDrone2018 [21], UAV123 [13]and UAVTrack112_L [6].

uated on all six datasets are shown in Table 3. ω_b goes from 1.0 to 3.0. Note that $\omega_b = 1.0$ reduces to the A-ViT* model. We can observe that the best precision is consistently achieved at $\omega_b = 1.5$. However, the best AUC is diversely distributed over the range of ω_b, despite that the second and the third-best AUCs are also obtained at $\omega_b = 1.5$. These results suggest that the impact of halting the background tokens strongly depends on datasets as well, which may be explainedy by the difficulties of discriminating the targets from backgrounds vary among different datasets.

Table 1. Precision and speed (FPS) comparison between Aba-ViTrack and deep-based trackers on DTB70[10], UAVDT [5], VisDrone2018 [21], and UAV123@10fps [13]. Note that the precision and AUC are shown in form of **(PRC, AUC)**, and the average GPU speed are shown in form of **GPU** fps. Red, blue and green indicate the first, second and third place.

Method	DTB70[10]	UAVDT[5]	VisDrone2018[21]	UAV123@10fps[13]	avg. FPS
SiamMask (CVPR 2018) [16]	(76.9, 57.1)	(80.5, 59.7)	(79.4, 58.1)	(78.8, 59.1)	**110.5**
DiMP18 (ICCV 2019) [1]	(79.8, 61.7)	(77.2, 55.8)	(76.4, 57.5)	(83.5, 63.8)	74.2
DiMP50 (ICCV 2019) [1]	(79.2, 61.3)	(78.3, 57.4)	(83.5, 63.0)	(85.1, 64.7)	51.3
SiamRPN++ (CVPR 2019) [9]	(79.9, 61.4)	(82.2, 61.0)	(79.1, 60.0)	(78.4, 59.4)	57.6
SiamDW (CVPR 2019) [19]	(73.5, 50.4)	(67.9, 43.6)	(79.7, 59.8)	(71.6, 50.6)	66.5
PrDiMP18 (CVPR 2020) [4]	(**84.0**, 64.3)	(75.8, 55.9)	(79.8, 60.2)	(83.9, 64.7)	53.9
PrDiMP50 (CVPR 2020) [4]	(76.4, 59.5)	(82.7, 60.1)	(79.4, 59.7)	(**87.9**, 67.5)	42.3
TransT (CVPR 2021) [3]	(83.6, 65.8)	(82.6, 64.2)	(**85.9**, **65.2**)	(84.8, 66.5)	55.0
SiamGAT (CVPR 2021) [8]	(75.1, 57.9)	(76.4, 58.9)	(78.3, 59.2)	(77.6, 59.7)	95.8
TrDiMP (CVPR 2021) [15]	(82.4, 63.9)	(**88.2**, **64.5**)	(84.1, 63.1)	(87.3, 66.5)	35.8
AutoMatch (ICCV 2021) [18]	(82.5, 63.4)	(82.1, 62.9)	(78.1, 59.6)	(78.1, 59.4)	63.6
SAOT (ICCV 2021) [20]	(83.1, 64.6)	(82.1, 60.7)	(76.9, 59.1)	(85.2, 65.7)	35.2
LightTrack (CVPR 2021) [17]	(76.4, 59.1)	(80.4, 61.1)	(74.8, 56.8)	(77.6, 59.9)	**103.6**
KeepTrack (ICCV 2021) [12]	(83.6, 64.3)	(83.8, 60.5)	(84.0, 63.5)	(**89.7**, **68.2**)	20.3
TrSiam (CVPR 2021) [15]	(82.7, 63.9)	(**88.9**, **65.0**)	(84.7, 64.0)	(85.3, 64.9)	38.1
CSWinTT (CVPR 2022) [14]	(80.3, 62.3)	(67.3, 54.0)	(75.2, 58.0)	(87.1, **68.1**)	9.6
SparseTT (IJCAI 2022) [7]	(82.3, **65.8**)	(82.8, **65.4**)	(81.4, 62.1)	(82.2, 64.9)	31.5
SLT-TransT (ECCV 2022) [9]	(83.4, 65.6)	(82.9, 62.5)	(**85.6**, **65.3**)	(86.2, 67.4)	32.6
SLT-TrDiMP (ECCV 2022) [9]	(83.6, 64.5)	(**87.9**, 63.8)	(85.1, 63.6)	(**88.0**, 67.1)	31.3
ToMP (CVPR 2022) [11]	(**85.6**, **67.1**)	(85.4, 64.1)	(84.1, **64.4**)	(87.5, **67.9**)	23.8
Aba-ViTrack (Ours)	(**85.9**, **66.4**)	(83.4, 59.9)	(**86.1**, **65.3**)	(85.0, 65.5)	**181.5**

Table 2. Ablation study of weighting the ponder loss $\mathcal{L}^*_{ponder}$ on DTB70 [10], UAVDT [5], VisDrone2018 [21], UAV123@10fps [13], UAV123 [13], and UAVTrack112_L [6]with α_p ranging from 0.5×10^{-4} to 1.5×10^{-4}. Note that $\times 10^{-4}$ is omitted for simplicity. And the precision and AUC are shown in form of (PRC, AUC).

α_p	DTB70[10]	UAVDT[5]	VisDrone2018[21]	UAV123@10fps[13]	UAV123[13]	UAVTrack112_L[6]
0.5	(82.9, 64.6)	(80.9, 58.4)	(83.6, 63.4)	(82.2, **65.4**)	(82.0, 65.0)	(**78.0**, **63.2**)
0.6	(**85.4**, **65.8**)	(80.8, 58.4)	(83.7, 63.5)	(81.4, 65.0)	(**84.0**, **66.4**)	(76.8, 61.8)
0.7	(84.2, 65.1)	(81.8, 59.1)	(84.1, 63.7)	(79.7, 63.7)	(83.2, 66.0)	(77.8, 63.0)
0.8	(83.6, 65.1)	(80.7, 58.4)	(83.1, 63.0)	(81.5, 64.6)	(80.7, 63.7)	(77.4, 62.5)
0.9	(83.4, 64.6)	(77.8, 56.5)	(82.7, 63.1)	(**83.2**, **65.9**)	(82.7, 65.5)	(**78.1**, 63.1)
1.0	(**85.9**, **66.4**)	(**83.4**, **59.9**)	(**86.1**, **65.3**)	(**85.0**, **65.5**)	(**86.4**, **66.4**)	(**81.1**, **64.2**)
1.1	(83.9, 65.2)	(**82.7**, **59.3**)	(82.1, 62.8)	(**82.3**, **65.4**)	(83.2, 65.7)	(77.2, 62.3)
1.2	(**85.1**, **65.7**)	(80.4, 58.2)	(**84.3**, **63.9**)	(80.9, 64.5)	(83.1, 65.7)	(**78.1** **63.3**)
1.3	(82.9, 64.4)	(**81.9**, **59.6**)	(83.9, 63.3)	(82.3, **65.5**)	(83.6, **66.2**)	(77.8, 63.0)
1.4	(**85.1**, **65.7**)	(79.7, 57.4)	(84.0, 63.3)	(82.1, **65.4**)	(83.2, 65.9)	(77.7, 62.8)
1.5	(83.8, 65.5)	(79.5, 57.4)	(**85.4**, **65.1**)	(81.5, 64.9)	(**84.7**, **67.0**)	(76.9, 62.4)

Therefore, the weighting of the ponder loss of background tokens should be set appropriately, since too large weight may stop too many background tokens so that the discriminative learning lacks sufficient negative samples, thus resulting in degraded performance, whereas, small weight reduces the model to the baseline A-ViT* without prior information about background is exploited. As can be seen, when ω_b is appropriately set with fixed α_p, our proposed background-aware ponder loss can improve PRC and AUC of the baseline A-ViT* on all datasets.

Table 3. Ablation study of weighting the background tokens on DTB70 [10], UAVDT [5], VisDrone2018 [21], UAV123@10fps [13], UAV123 [13], and UAVTrack112_L [6] with ω_b ranging from 1.0 to 3.0. Note that the precision and AUC are shown in form of (PRC, AUC).

ω_b	DTB70[10]	UAVDT[5]	VisDrone2018[21]	UAV123@10fps[13]	UAV123[13]	UAVTrack112_L[6]
1.0	(84.1, 64.7)	(78.2, 56.7)	(84.4, 63.9)	(82.1, 65.3)	(82.9, 65.6)	(76.8, 62.1)
1.1	(**85.6**, **65.9**)	(**83.3**, **60.3**)	(82.1, 62.9)	(**82.7**, **65.8**)	(83.7, **66.5**)	(76.1, 61.9)
1.2	(83.1, 64.6)	(80.3, 57.9)	(**86.0**, **66.1**)	(80.7, 64.2)	(81.9, 64.9)	(78.9, 63.8)
1.3	(83.9, 64.7)	(82.1, 59.1)	(82.3, 62.4)	(81.5, 64.9)	(82.9, 65.6)	(77.9, 63.2)
1.4	(83.2, 64.4)	(78.6, 57.3)	(**85.8**, **64.9**)	(82.3, **65.4**)	(82.6, 65.4)	(77.0, 61.9)
1.5	(**85.9**, **66.4**)	(**83.4**, **59.9**)	(**86.1**, **65.3**)	(**85.0**, **65.5**)	(**86.4**, **66.4**)	(**81.1**, **64.2**)
1.6	(84.1, 64.9)	(81.5, 58.9)	(85.1, 64.5)	(81.3, 64.9)	(82.6, 65.7)	(78.7, 63.7)
1.7	(84.4, 65.3)	(79.7, 57.7)	(82.9, 62.8)	(80.6, 64.3)	(82.6, 65.6)	(**79.8**, **64.4**)
1.8	(**85.5**, **65.5**)	(80.0, 58.3)	(84.3, 64.2)	(82.0, 65.2)	(**84.8**, **66.8**)	(76.9, 62.6)
1.9	(83.8, 64.5)	(82.3, 59.4)	(84.9, 64.7)	(80.8, 64.3)	(82.4, 65.4)	(78.0, 62.8)
2.0	(82.5, 64.1)	(**84.6**, **61.5**)	(80.6, 61.5)	(**83.0**, 65.9)	(83.4, 66.0)	(78.9, **64.1**)
2.5	(84.6, 65.3)	(78.2, 56.6)	(83.9, 63.7)	(81.6, 64.9)	(**83.9**, **66.4**)	(**79.1**, 63.9)
3.0	(84.4, 64.9)	(83.0, 59.4)	(84.9, 64.7)	(80.7, 64.3)	(83.3, 66.0)	(78.2, 63.4)

E. Attribute-based Evaluation

To further examine and comprehend the performances of different trackers, we conduct performance evaluations based on 11 attributes. Our Aba-ViTrack achieves the best PRC and AUC on most attributes. To limit the page length, we demonstrate the success plots and precision plots on the DTB70 [10] dataset in this section and leave the results on UAV123@10fps [13], UAVDT [5], and VisDrone2018 [21] to the zip file.

As illustrated in Fig. 3, we can observe that Aba-ViTrack achieves the best AUC on all attributes except on 'Out-of-view'. Specifically, Aba-ViTrack significantly outperforms the second tracker on 'Aspect ratio variation', 'Scale variation', 'Background clutter', 'Deformation', and 'Out-of-plane rotation', with gains of 5.0%, 5.5%, 7.0%, 6.8%, 19.1%, respectively. It is worthy of note that only a small margin of 0.7% between Aba-ViTrack and the first tracker TCTrack [2] is observed, although Aba-ViTrack is inferior to TCTrack [2].

In terms of precision, our Aba-ViTrack achieves the best performance on all attributes except on 'Occlusion', as shown in Fig. 4. More specifically, Aba-ViTrack significantly surpasses the second tracker on 'Aspect ratio variation', 'Motion blur', 'Scale variation', 'Background clutter', 'Deformation', 'Out-of-plane rotation', and 'In-plane rotation', by a significant margin of 5.8%, 4.8%, 6.5%, 4.9%, 7.5%, 17.9%, and 5.1%, respectively. Although Aba-ViTrack is surpassed by TCTrack [2] on 'Occlusion', the difference is only 0.3%, quite small a margin. These results justify the effectiveness of our method in raising tracking performance.

References

[1] Goutam Bhat, Martin Danelljan, Luc Van Gool, and Radu Timofte. Learning discriminative model prediction for tracking. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6182–6191, 2019.

[2] Ziang Cao, Ziyuan Huang, Liang Pan, Shiwei Zhang, Ziwei Liu, and Changhong Fu. Tctrack: Temporal contexts for aerial tracking. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14778–14788, 2022.

[3] Xin Chen, Bin Yan, Jiawen Zhu, Dong Wang, Xiaoyun Yang, and Huchuan Lu. Transformer tracking. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8122–8131, 2021.

[4] Martin Danelljan, Luc Van Gool, and Radu Timofte. Probabilistic regression for visual tracking. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7181–7190, 2020.

[5] Dawei Du, Yuankai Qi, Hongyang Yu, Yi-Fan Yang, Kaiwen Duan, Guorong Li, W. Zhang, Qingming Huang, and Qi Tian. The unmanned aerial vehicle benchmark: Object detection and tracking. In *ECCV,*, pages 375–391, 2018.

[6] Changhong Fu, Ziang Cao, Yiming Li, Junjie Ye, and Chen Feng. Onboard real-time aerial tracking with efficient siamese anchor proposal network. *IEEE Transactions on Geoscience and Remote Sensing*, PP(99):1–13, 2021.

[7] Zhihong Fu, Zehua Fu, Qingjie Liu, Wenrui Cai, and Yunhong Wang. Sparsett: Visual tracking with sparse transformers. *arXiv preprint arXiv:2205.03776*, 2022.

[8] Dongyan Guo, Yan Shao, Ying Cui, Zhenhua Wang, Liyan Zhang, and Chunhua Shen. Graph attention tracking. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9538–9547, 2020.

[9] Minji Kim, Seungkwang Lee, Jungseul Ok, Bohyung Han, and Minsu Cho. Towards sequence-level training for visual tracking. *ArXiv*, abs/2208.05810, 2022.

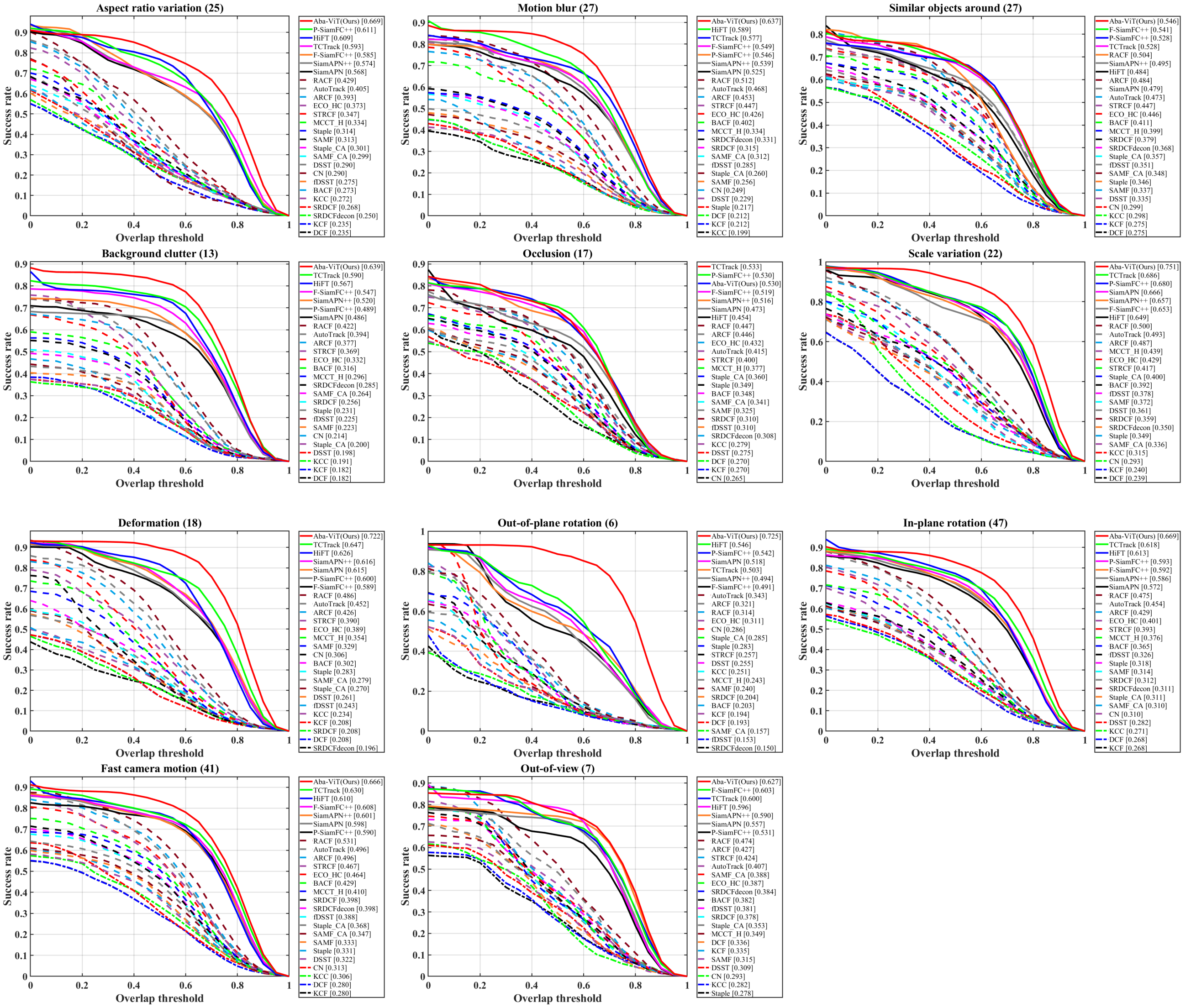

Figure 3. Success plots of attribute-based evaluation of all trackers on DTB70 [10]. The area under curve (AUC) are used for ranking and marked in the success plots. Our Aba-ViTrack achieves the best AUC at most attributes.

[10] Siyi Li and D. Y. Yeung. Visual object tracking for unmanned aerial vehicles: A benchmark and new motion models. In *AAAI Conference on Artificial Intelligence*, pages 4140–4146, 2017.

[11] Christoph Mayer, Martin Danelljan, Goutam Bhat, Matthieu Paul, Danda Pani Paudel, Fisher Yu, and Luc Van Gool. Transforming model prediction for tracking. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8721–8730, 2022.

[12] Christoph Mayer, Martin Danelljan, Danda Pani Paudel, and Luc Van Gool. Learning target candidate association to keep track of what not to track. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 13424–13434, 2021.

[13] Matthias Mueller, Neil G. Smith, and Bernard Ghanem. A benchmark and simulator for uav tracking. In *European Conference on Computer Vision*, 2016.

[14] Zikai Song, Junqing Yu, Yi-Ping Phoebe Chen, and Wei Yang. Transformer tracking with cyclic shifting window attention. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8781–8790, 2022.

[15] Ning Wang, Wen gang Zhou, Jie Wang, and Houqiang Li. Transformer meets tracker: Exploiting temporal context for robust visual tracking. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1571–1580, 2021.

[16] Qiang Wang, Li Zhang, Luca Bertinetto, Weiming Hu, and Philip H. S. Torr. Fast online object tracking and segmentation: A unifying approach. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1328–1338, 2018.

[17] B. Yan, Houwen Peng, Kan Wu, Dong Wang, Jianlong Fu,

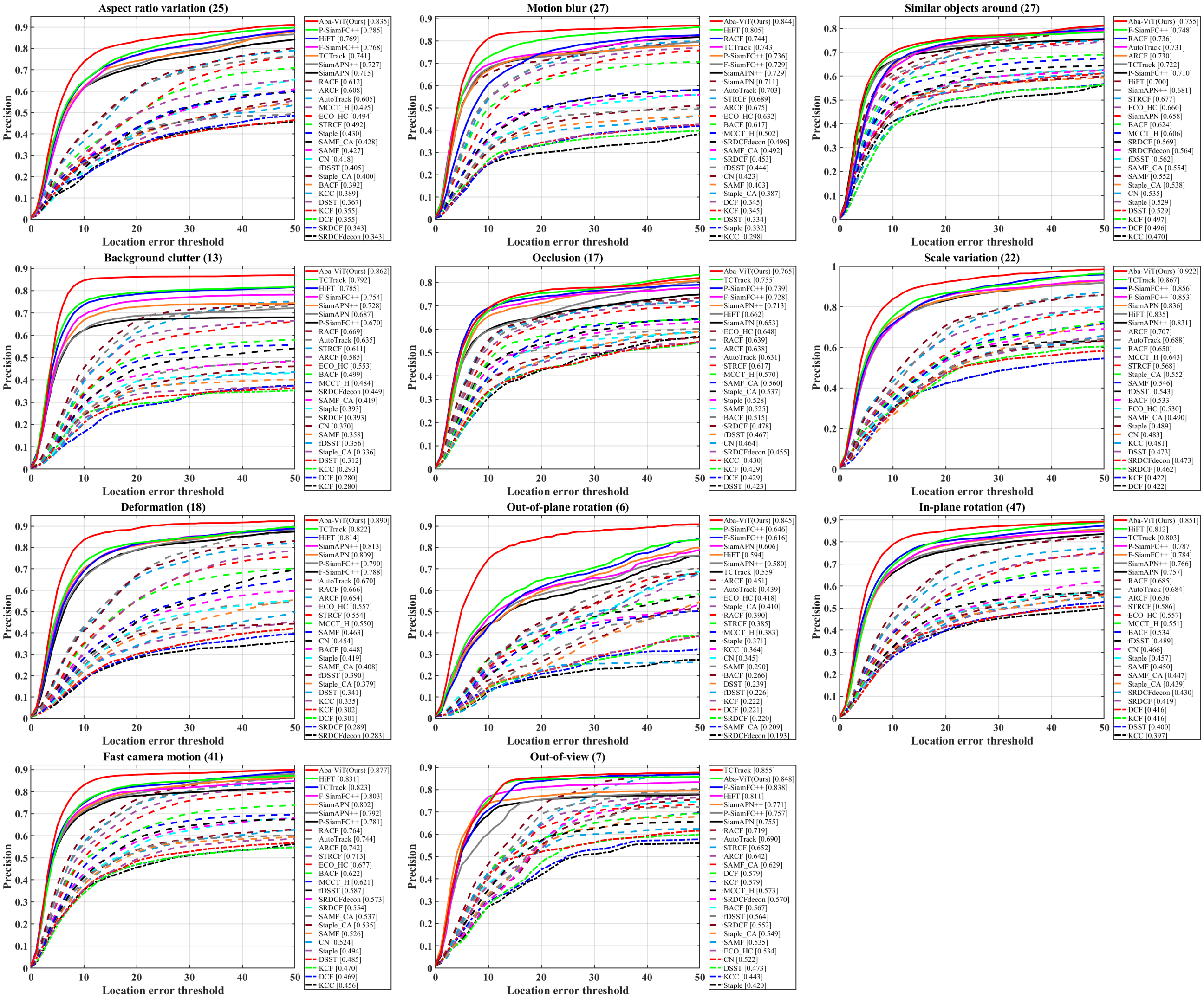

Figure 4. Precision plots of attribute-based evaluation of all trackers on DTB70 [10]. The precision at 20 pixels are used for ranking and marked in the precision plots. Our Aba-ViTrack achieves the best precision at most attributes.

and Huchuan Lu. Lighttrack: Finding lightweight neural networks for object tracking via one-shot architecture search. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15175–15184, 2021.

[18] Zhipeng Zhang, Yihao Liu, Xiao Wang, Bing Li, and Weiming Hu. Learn to match: Automatic matching network design for visual tracking. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 13319–13328, 2021.

[19] Zhipeng Zhang, Houwen Peng, and Qiang Wang. Deeper and wider siamese networks for real-time visual tracking. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4586–4595, 2019.

[20] Zikun Zhou, Wenjie Pei, Xin Li, Hongpeng Wang, Feng Zheng, and Zhenyu He. Saliency-associated object tracking. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9846–9855, 2021.

[21] Pengfei Zhu, Longyin Wen, and et al. Visdrone-sot2018: The vision meets drone single-object tracking challenge results. In *ECCV Workshops*, pages 469–495, 2018.