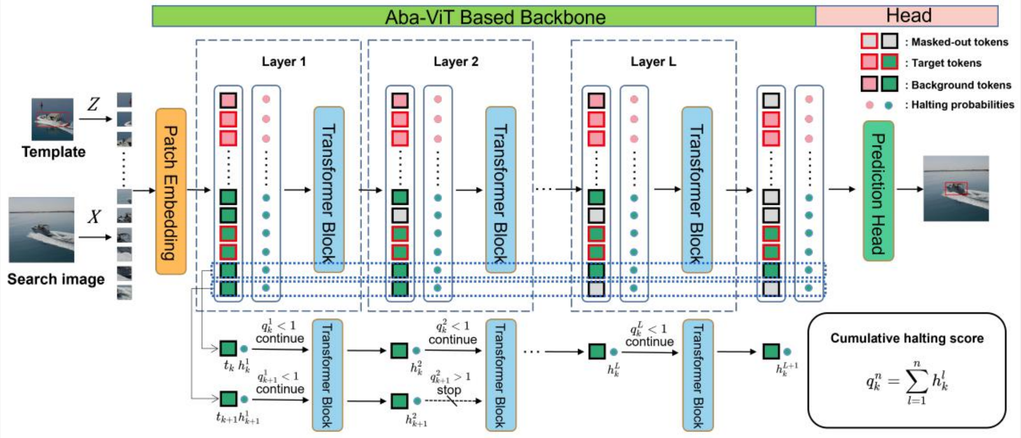
Adaptive and Background-Aware Vision Transformer for Real-Time UAV Tracking

Shuiwang Li, Yangxiang Yang, Dan Zeng, Xucheng Wang

Introduction

In this paper, we propose an efficient ViT-based tracking framework, Aba-ViTrack, for UAV tracking. In our framework, feature learning and template-search coupling are integrated into an efficient one-stream ViT to avoid an extra heavy relation modeling module. The proposed Aba-ViT exploits an adaptive and background-aware token computation method to reduce inference time. This approach adaptively discards tokens based on learned halting probabilities, which a priori are higher for background tokens than target ones.

Method



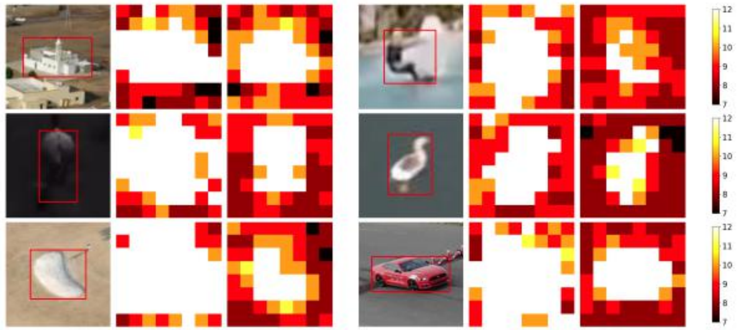


Figure 4. Original image (left), the dynamic token depth of A-ViT (middle), and that of Aba-ViT (right) on samples from the DTB70 [35], UAV123 [48], and UAVTrack112_L [20].

Experiments

Table 1. Precision and speed (FPS) comparison between Aba-ViTrack and deep-based trackers on DTB70 [35] . Red, blue and green indicate the first, second and third place.

Tracker	PRC	FPS	Tracker	PRC	FPS
Aba-ViTrack	85.9	185.4	DiMP18 [2]	79.8	73.0
PrDiMP18 [12]	84.0	55.7	DiMP50 [2]	79.2	52.4
PrDiMP50 [12]	76.4	42.1	SiamMask [63]	76.9	109.6
SiamRPN++ [29]	79.9	58.2	AutoMatch [76]	82.5	65.2
SiamDW [78]	73.5	65.0	SAOT [79]	83.1	34.0
TransT [7]	83.6	53.7	TrSiam [61]	82.7	36.3
SiamGAT [24]	75.1	92.3	KeepTrack [46]	83.6	19.5
CSWinTT [51]	82.4	9.6	SparseTT [21]	82.3	31.5

Table 2. Ablation study of weighting the ponder loss $\mathcal{L}^*_{ponder}$ on DTB70 [35] with α_p ranging from 0.5×10^{-4} to 1.5×10^{-4}. Note that $\times 10^{-4}$ is omitted for simplicity. PRC stands for precision.

α_p	0.5	0.6	0.7	0.8	0.9	1.0	1.1	1.2	1.3	1.4	1.5
PRC	82.9	85.4	84.2	83.6	83.4	85.9	83.9	85.1	82.9	85.1	83.8
AUC	64.6	65.8	65.1	65.1	64.6	66.4	65.2	65.7	64.4	65.7	65.5

Table 3. Ablation study of weighting the background tokens on DTB70 [35] with ω_b ranging from 1.0 to 3.0.

ω_b	1.0	1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8	1.9	2.0	2.5	3.0
PRC	84.1	85.6	83.1	83.9	83.2	85.9	84.1	84.4	85.5	82.6	82.5	84.6	84.4
AUC	64.7	65.9	64.6	64.7	64.4	66.4	64.9	65.3	65.5	64.0	64.1	65.3	64.9

Table 4. Evaluation of efficient ViT-based Trackers. Four lightweight ViTs, i.e. ViT-tiny [16], DeiT-tiny [54], A-ViT [73], and Aba-ViT, are integrated into the proposed tracking framework, denoted by ViT-tiny*, DeiT-tiny*, A-ViT*, and Aba-ViT*, respectively. Note that the precision and AUC are shown in form of **(PRC, AUC)**, and the average GPU and CPU speed are shown in form of **[GPU fps, CPU fps]**.

Method		UAV123@10fps [48]	DTB70 [35]	UAVDT [17]	VisDrone2018 [80]	UAV123 [48]	UAVTrack112_L [20]	Avg. FPS [GPU, CPU]
DCF-based	ECO-HC [11]	(64.0, 46.8)	(63.5, 44.8)	(69.4, 41.6)	(80.8, 58.1)	(71.0, 49.6)	(64.8, 41.7)	[— , 83.5]
	ARCF [28]	(66.6, 47.3)	(69.4, 47.2)	(72.0, 45.8)	(79.7, 58.4)	(67.1, 46.8)	(64.0, 39.9)	[— , 34.2]
	AutoTrack [37]	(67.1, 47.7)	(71.6, 47.8)	(71.8, 45.0)	(78.8, 57.3)	(68.9, 47.2)	(67.5, 40.2)	[— , 57.8]
	RACF [33]	(69.4, 48.6)	(72.5, 50.5)	(77.3, 49.4)	(83.4, 60.0)	(70.2, 47.7)	(62.6, 40.0)	[— , 35.6]
CNN-based	HiFT [4]	(74.9, 57.0)	(80.2, 59.4)	(65.2, 47.5)	(71.9, 52.6)	(78.7, 59.0)	(73.4, 55.1)	[160.3, —]
	P-SiamFC++ [66]	(73.1, 54.9)	(80.3, 60.4)	(80.7 ,55.6)	(80.1, 58.5)	(74.5, 48.9)	(70.4, 53.1)	[240.5, 46.1]
	F-SiamFC++ [67]	(72.1, 54.5)	(81.4, 60.5)	(79.4, 55.5)	(80.7, 59.6)	(78.9, 59.2)	(74.2, 54.5)	[255.4, 51.6]
	TCTrack [5]	(78.0, 59.9)	(81.2, 62.2)	(72.5, 53.0)	(79.9, 59.4)	(80.0, 60.5)	(78.6, 58.3)	[139.6, —]
Efficient ViT-based	ViT-tiny*	(82.1 , 64.8)	(79.3, 62.4)	(77.0, 55.6)	(83.0, 62.7)	(83.2, 65.5)	(78.9, 63.6)	[166.2, 47.1]
	DeiT-tiny*	(83.5, 65.8)	(83.6, 64.9)	(81.2, 58.2)	(83.6, 63.8)	(82.8, 65.2)	(80.3, 64.6)	[164.6, 46.3]
	A-ViT*	(82.1, 65.3)	(84.1, 64.7)	(78.2, 56.7)	(84.4, 63.9)	(82.9, 66.4)	(76.8, 62.1)	[176.4, 49.6]
	Aba-ViTrack	(85.0, 65.5)	(85.9, 66.4)	(83.4, 59.9)	(86.1, 65.3)	(86.4, 66.4)	(81.1, 64.2)	[181.5, 50.3]

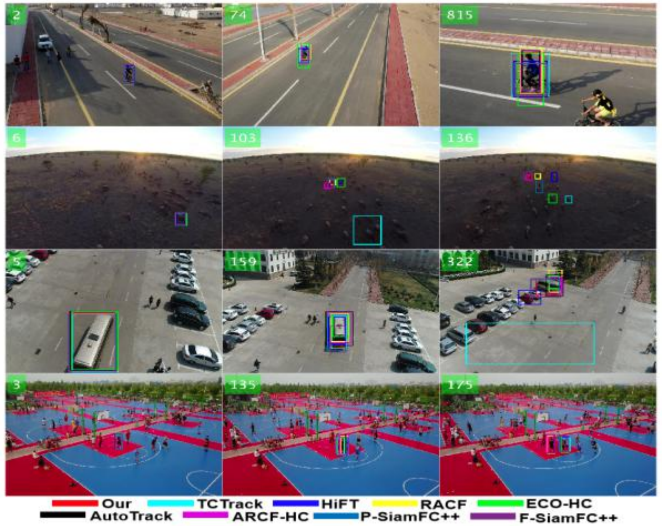


Figure 5. Qualitative evaluation on 4 video sequences from, respectively, UAV123@10fps [48], DTB70 [35], UAVDT [17], and VisDrone2018 [80] (i.e. bike1, Animal1, S1701, and uav000088_0000_s).