AiATrack: Attention in Attention for Transformer Visual Tracking

Shenyuan Gao

sygao@connect.ust.hk

06/08/2022

ECCV TEL AVIV 2022

Paper

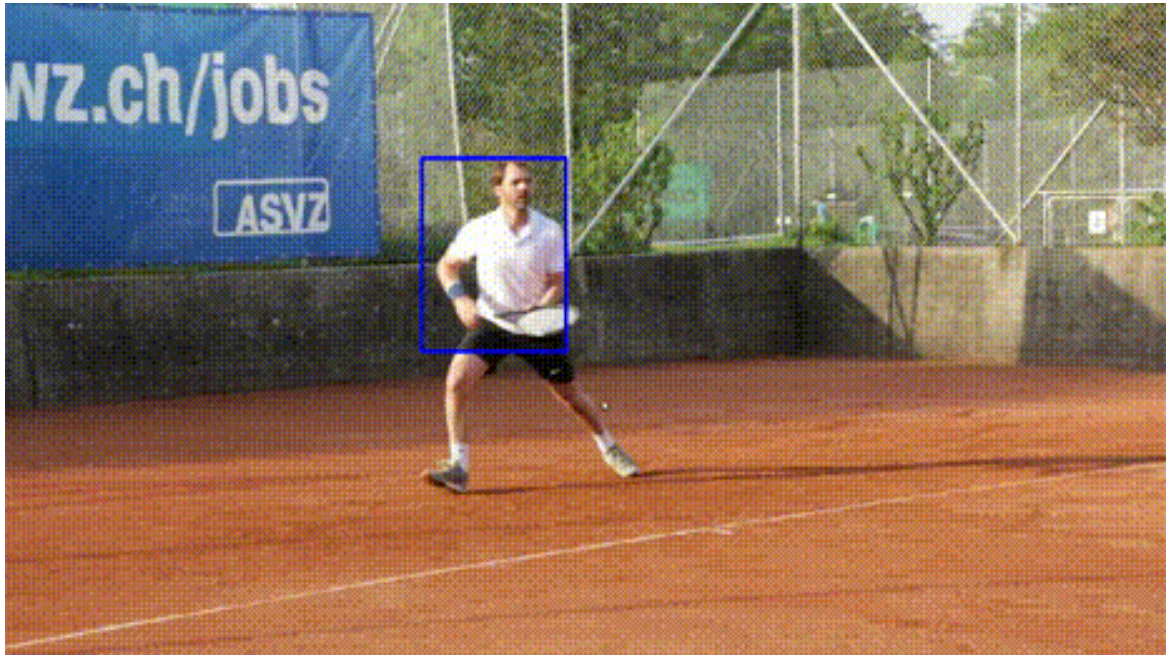
Code

Background: Single Object Tracking

- Objective:

  Given a target with bounding box annotation in the initial frame, localize the target in successive frames.

- Keywords:
  - Single-object
  - Model-free
  - One-shot
  - Real-time



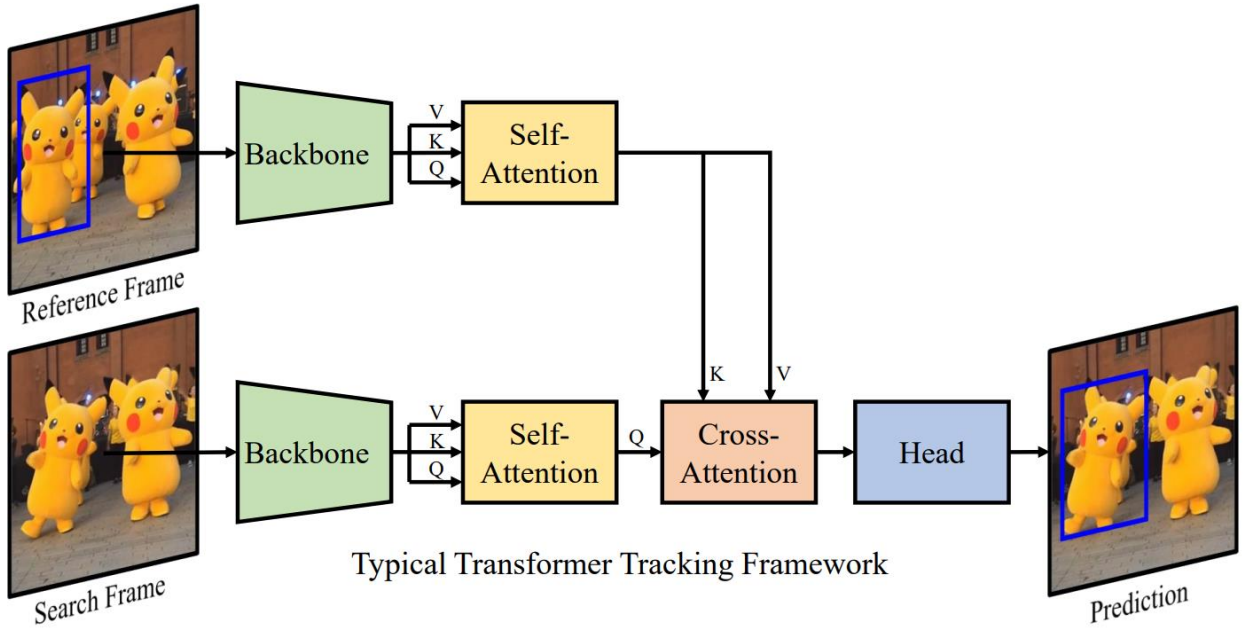
Background: Typical Transformer Tracker

- Self-attention blocks:

  Aggregate and enhance the extracted features.

- Cross-attention blocks:

  Propagate the information for target prediction.



Typical Transformer Tracking Framework

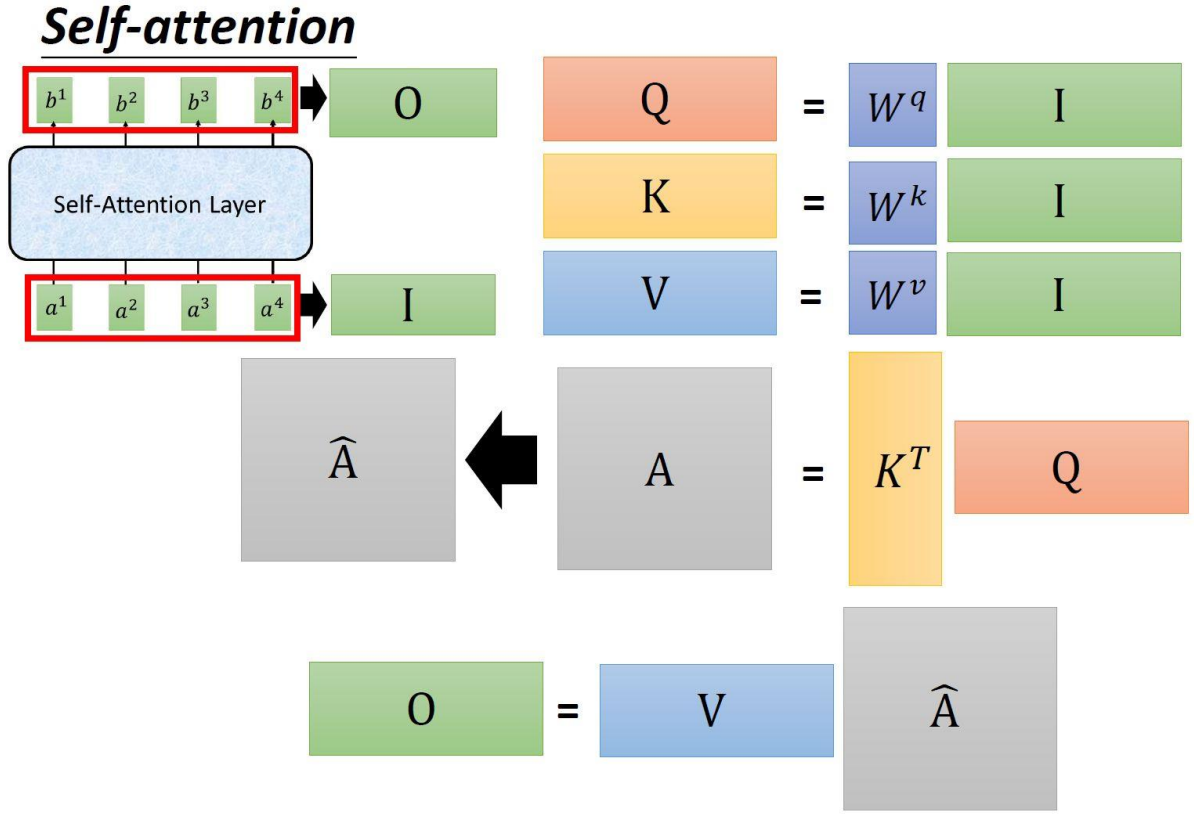
Chen et al., Transformer Tracking, CVPR 2021
Wang et al., Transformer Meets Tracker: Exploiting Temporal Context for Robust Visual Tracking, CVPR 2021
Yu et al., High-Performance Discriminative Tracking with Transformers, ICCV 2021

Background: Attention Mechanism

- Conventional attention mechanism:

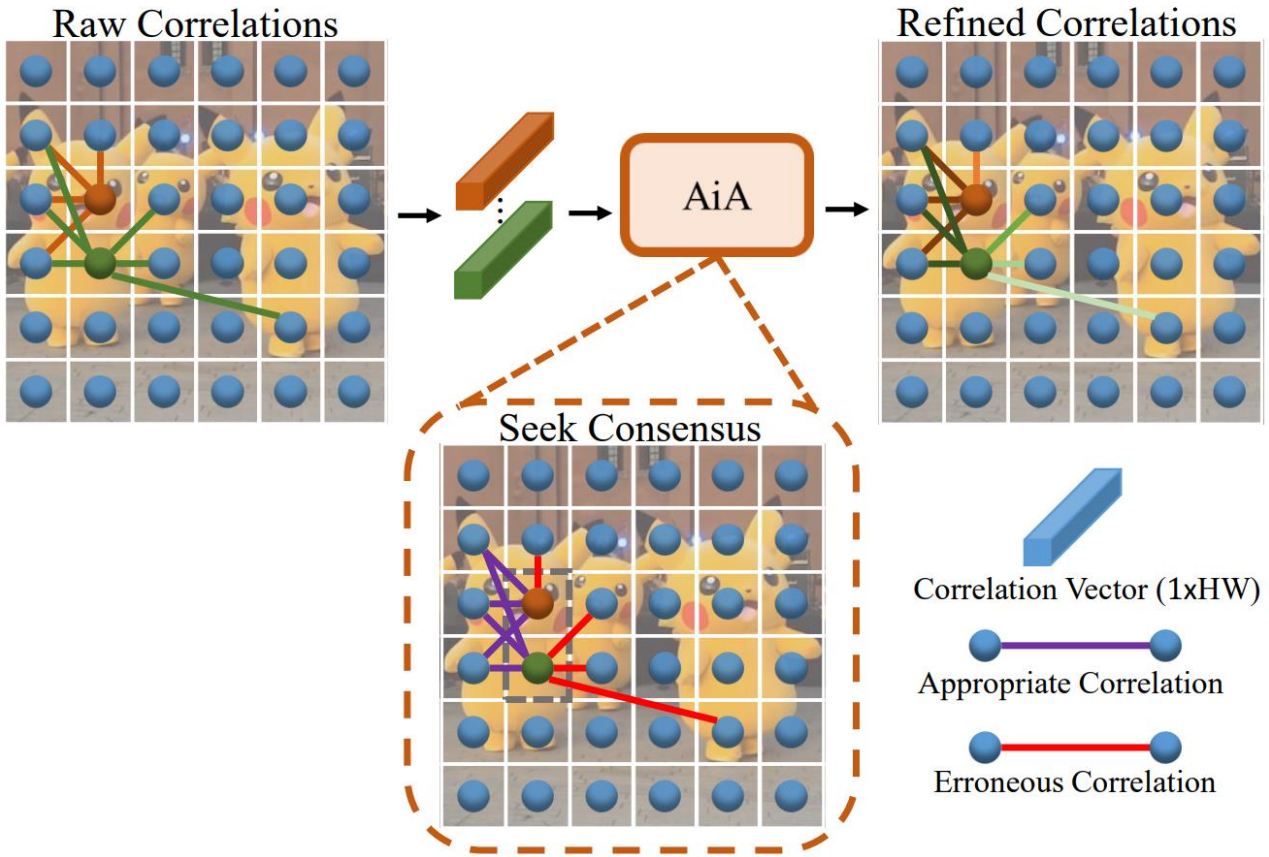
$$\mathrm{ConvenAttn}(\mathbf{Q},\mathbf{K},\mathbf{V}) = (\mathrm{Softmax}\left(\frac{\bar{\mathbf{Q}}\bar{\mathbf{K}}^{\mathbf{T}}}{\sqrt{C}}\right)\bar{\mathbf{V}})\mathbf{W_o}$$

- Limitation:
  - The correlation of each key-query pair is computed independently, which ignores the correlations of other query-key pairs.
  - Result in noisy and ambiguous attention weights, which may inhibit the potential of Transformer trackers.



Our Insight

- If a key has a high correlation with a query, its neighboring keys should also have relatively high correlations with that query.
- Seek consensus among raw correlations with a global receptive field.

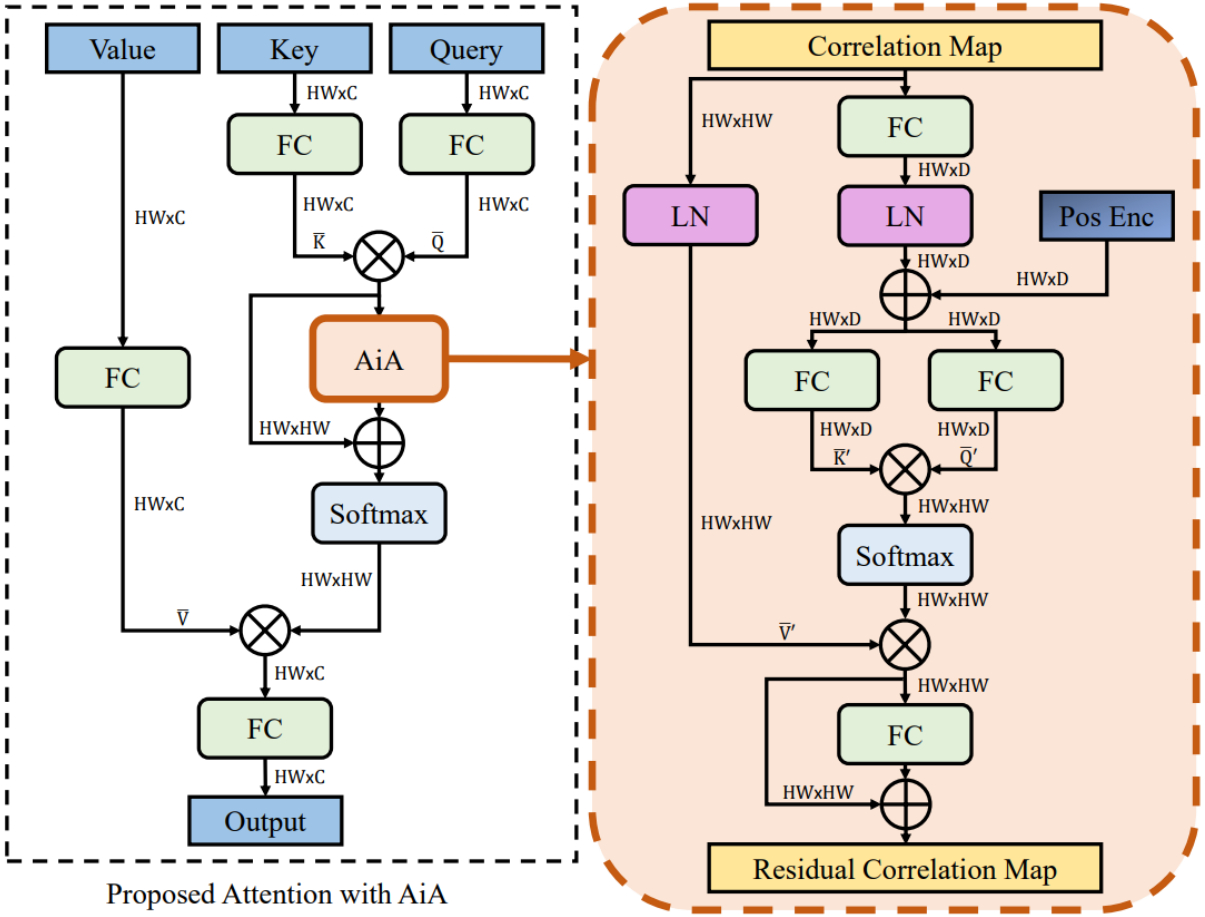


Methods: Attention in Attention (AiA)

- Generate the residual correlation map using another attention module:

$$\mathrm{AttninAttn}(\mathbf{Q},\mathbf{K},\mathbf{V}) = (\mathrm{Softmax}(\mathbf{M} + \mathrm{InnerAttn}(\mathbf{M}))\bar{\mathbf{V}})\mathbf{W_o}$$

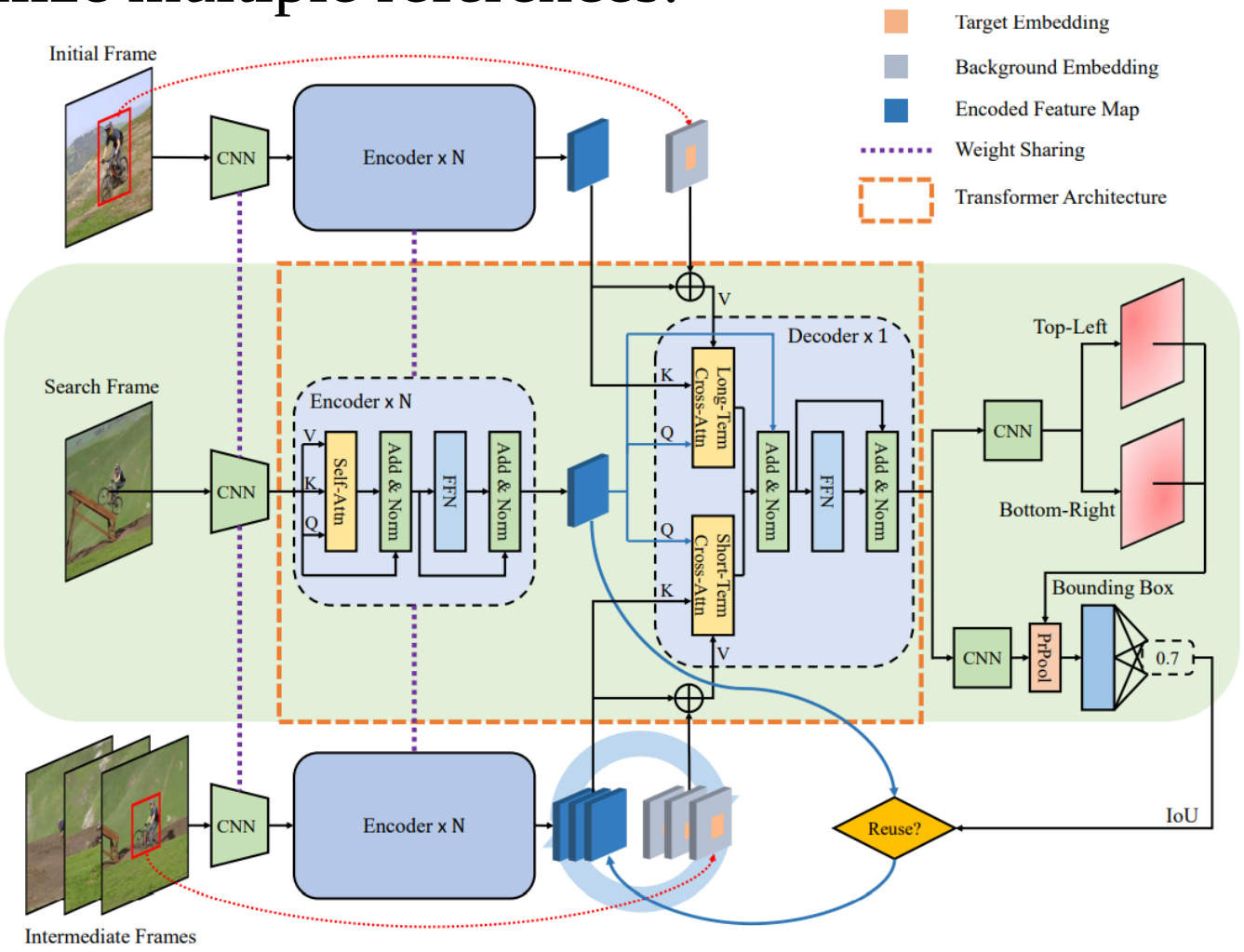
- Unified design:
  - Can be readily applied to both self-attention and cross-attention blocks.
- Performance gain:
  - A margin of 2.5% on LaSOT.
- Complexity increase:
  - A margin of 0.8% in Params.



Proposed Attention with AiA

Methods: Framework (AiATrack)

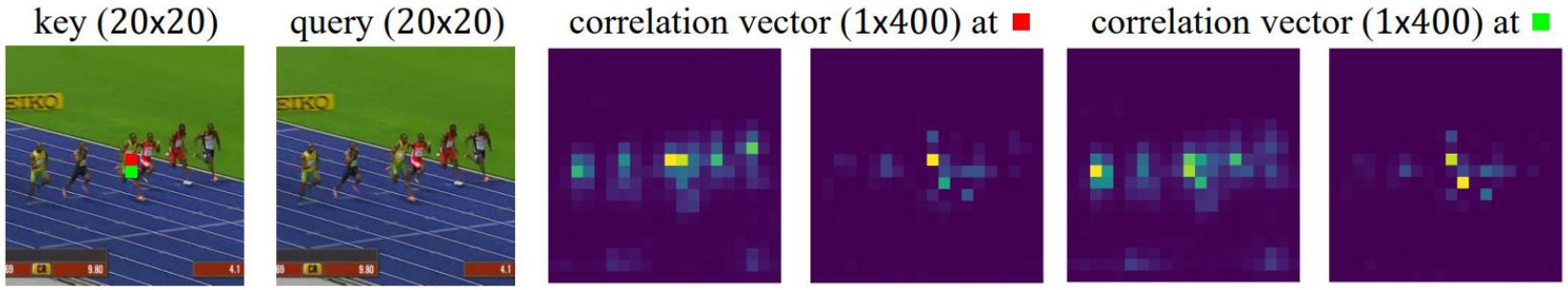
- A streamlined framework to utilize multiple references:
- More efficient:
  - Encoded feature reuse.
- More discriminative:
  - Target-background embeddings.
- More accurate:
  - IoU-guided update.
- More robust:
  - Two-branch design.



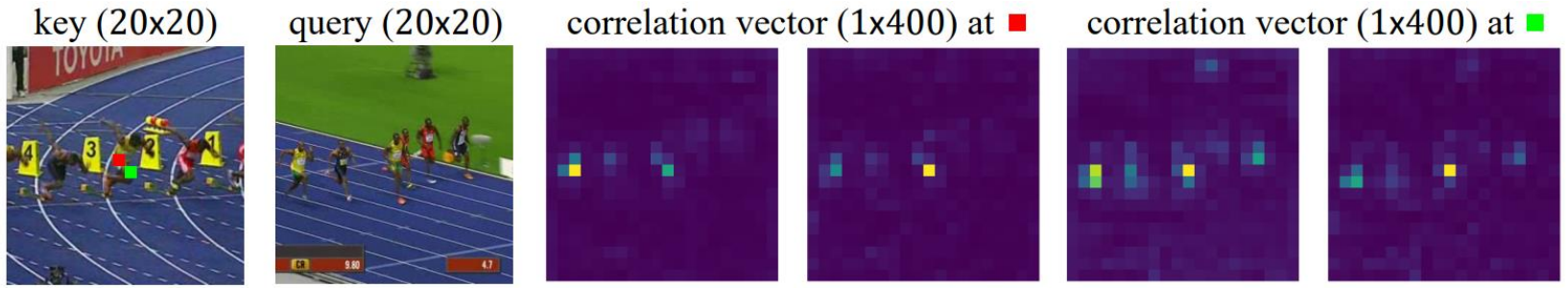
Results: Effectiveness of AiA

Modification	LaSOT			$\text{LaSOT}_{\text{Ext}}$		
	AUC	P_{Norm}	P	AUC	P_{Norm}	P
w/o AiA†	67.0	77.0	71.3	44.7	52.7	51.5
AiA in self-attn	68.6	78.7	72.9	46.2	**54.4**	53.4
AiA in cross-attn	67.5	77.9	71.8	46.2	54.2	53.3
w/o pos in both	68.0	78.2	72.7	46.2	54.0	53.0
AiA in both‡	**68.7**	**79.3**	**73.7**	**46.8**	**54.4**	**54.2**

Self-attention



Cross-attention



Results: Superiority of AiA

Modification	Correlation Refinement	LaSOT			$\text{LaSOT}_{\text{Ext}}$		
		AUC	P_{Norm}	P	AUC	P_{Norm}	P
w/o AiA†	✗	67.0	77.0	71.3	44.7	52.7	51.5
w/o AiA cascade	✗	67.1	77.0	71.7	44.6	52.9	51.6
conv in both	✓	67.9	78.2	72.8	46.0	53.4	52.8
AiA in both‡	✓	**68.7**	**79.3**	**73.7**	**46.8**	**54.4**	**54.2**

- Simply increasing the number of attention blocks in our framework does not help much.
- Inserting a convolutional bottleneck can also bring positive effects, which suggests the necessity of correlation refinement.
- The proposed AiA can boost much more performance than the convolutional substitute.

Results: Overall Performance

Tracker	Source	LaSOT [17]			TrackingNet [45]			GOT-10k [25]		
		AUC	P_{Norm}	P	AUC	P_{Norm}	P	AO	$SR_{0.75}$	$SR_{0.5}$
AiATrack	Ours	69.0	79.4	73.8	82.7	87.8	80.4	69.6	63.2	80.0
STARK-ST50 [58]	ICCV2021	66.4	76.3	71.2	81.3	86.1	78.1	68.0	62.3	77.7
KeepTrack [41]	ICCV2021	67.1	77.2	70.2	-	-	-	-	-	-
DTT [61]	ICCV2021	60.1	-	-	79.6	85.0	78.9	63.4	51.4	74.9
TransT [8]	CVPR2021	64.9	73.8	69.0	81.4	86.7	80.3	67.1	60.9	76.8
TrDiMP [53]	CVPR2021	63.9	-	61.4	78.4	83.3	73.1	67.1	58.3	77.7
TrSiam [53]	CVPR2021	62.4	-	60.0	78.1	82.9	72.7	66.0	57.1	76.6
KYS [4]	ECCV2020	55.4	63.3	-	74.0	80.0	68.8	63.6	51.5	75.1
Ocean-online [67]	ECCV2020	56.0	65.1	56.6	-	-	-	61.1	47.3	72.1
Ocean-offline [67]	ECCV2020	52.6	-	52.6	-	-	-	59.2	-	69.5
PrDiMP50 [12]	CVPR2020	59.8	68.8	60.8	75.8	81.6	70.4	63.4	54.3	73.8
SiamAttn [62]	CVPR2020	56.0	64.8	-	75.2	81.7	-	-	-	-
DiMP50 [3]	ICCV2019	56.9	65.0	56.7	74.0	80.1	68.7	61.1	49.2	71.7
SiamRPN++ [34]	CVPR2019	49.6	56.9	49.1	73.3	80.0	69.4	51.7	32.5	61.6

Concluding Remarks

- We propose a novel attention in attention module to enhance appropriate correlations and suppress unreliable ones by seeking consensus among all correlation vectors, which further unveils the potential of the Transformer tracker.

- We propose a concise and efficient framework to utilize multiple references, which we believe would shed light on future works.

- We hope that the proposed methods could serve as possible solutions for future contributions in related tasks, such as video object segmentation, video object detection, and multi-object tracking.

Thanks!

Paper

Code

Transformer Tracking