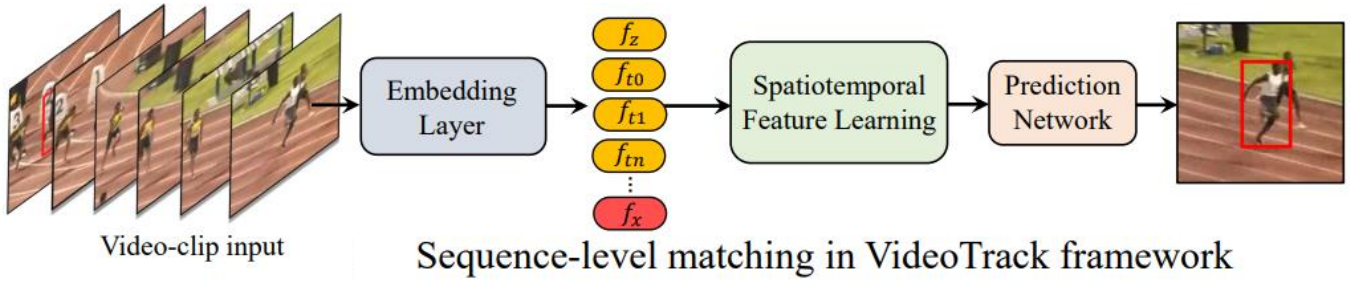
VideoTrack: Learning to Track Objects via Video Transformer

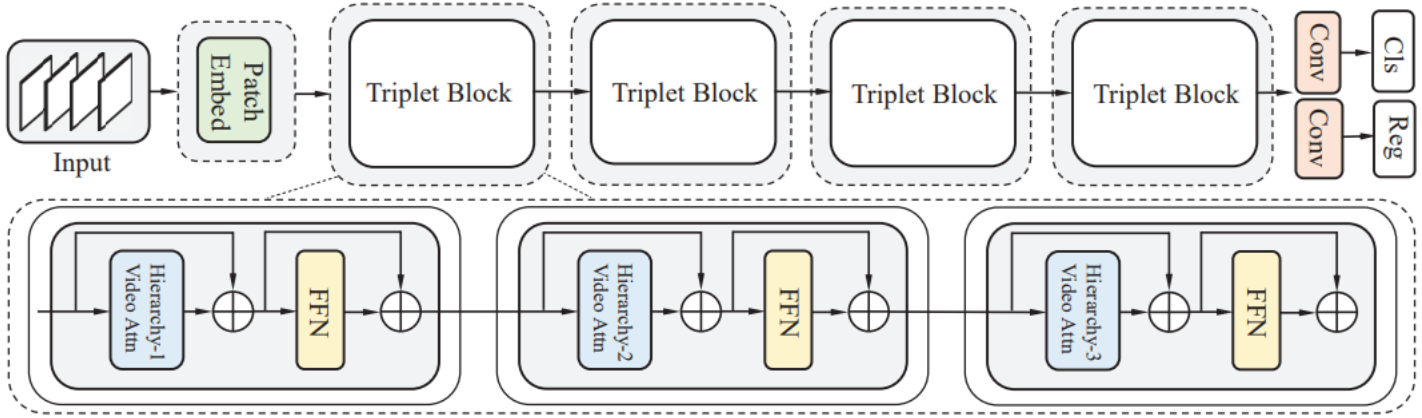
Fei Xie[1*], Lei Chu[2], Jiahao Li[2], Yan Lu[2] and Chao Ma[1]

[1] MoE Key Lab of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University
[2]Microsoft Research Asia

* This work was done when Fei Xie was an intern at Microsoft Research Asia

Brief introduction



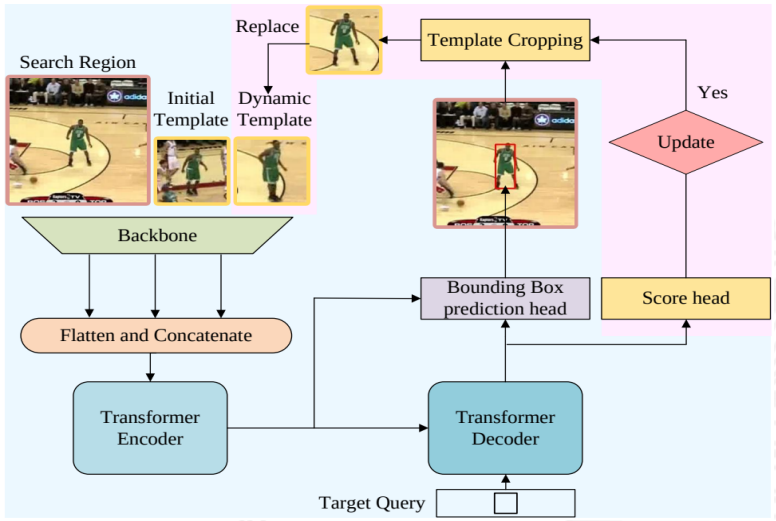


Sequence-level matching

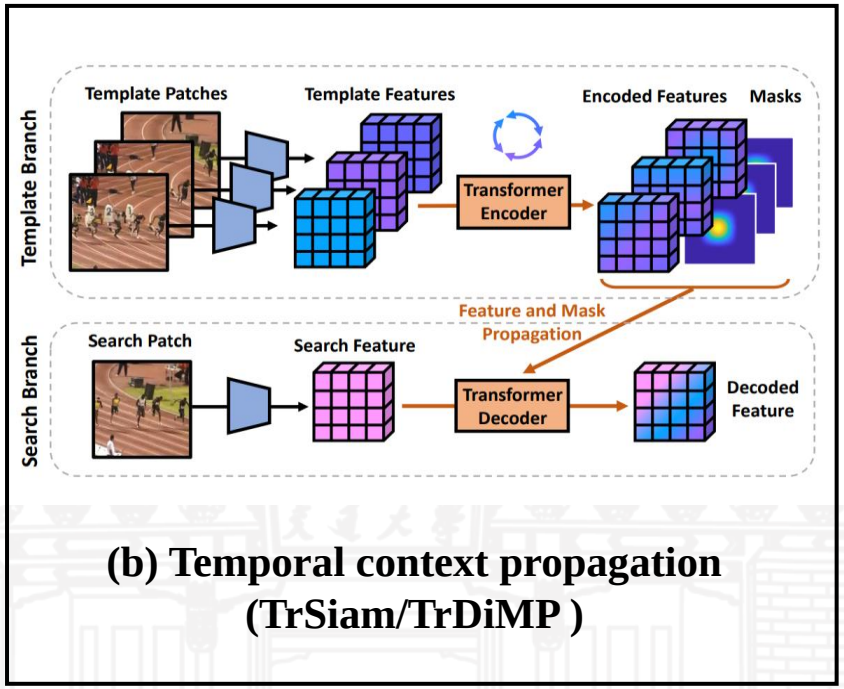
On top of video transformer network

Get rid of complex temporal modelling

Background

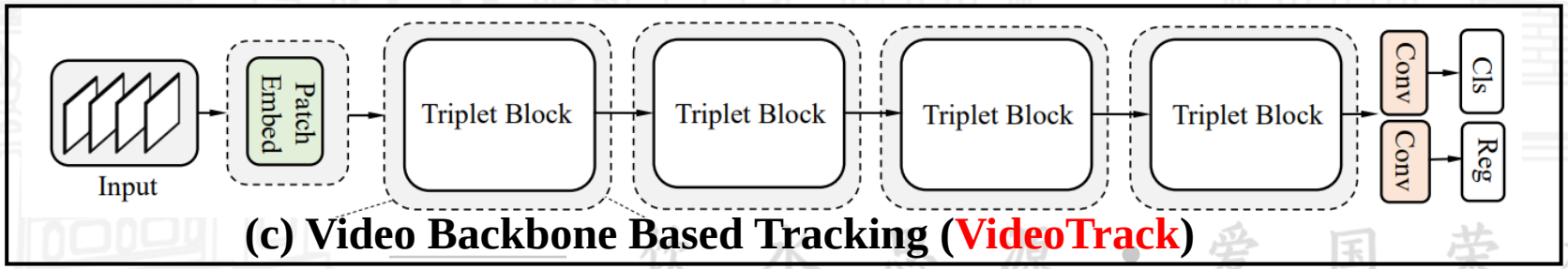


(a) Online template update (STARK)



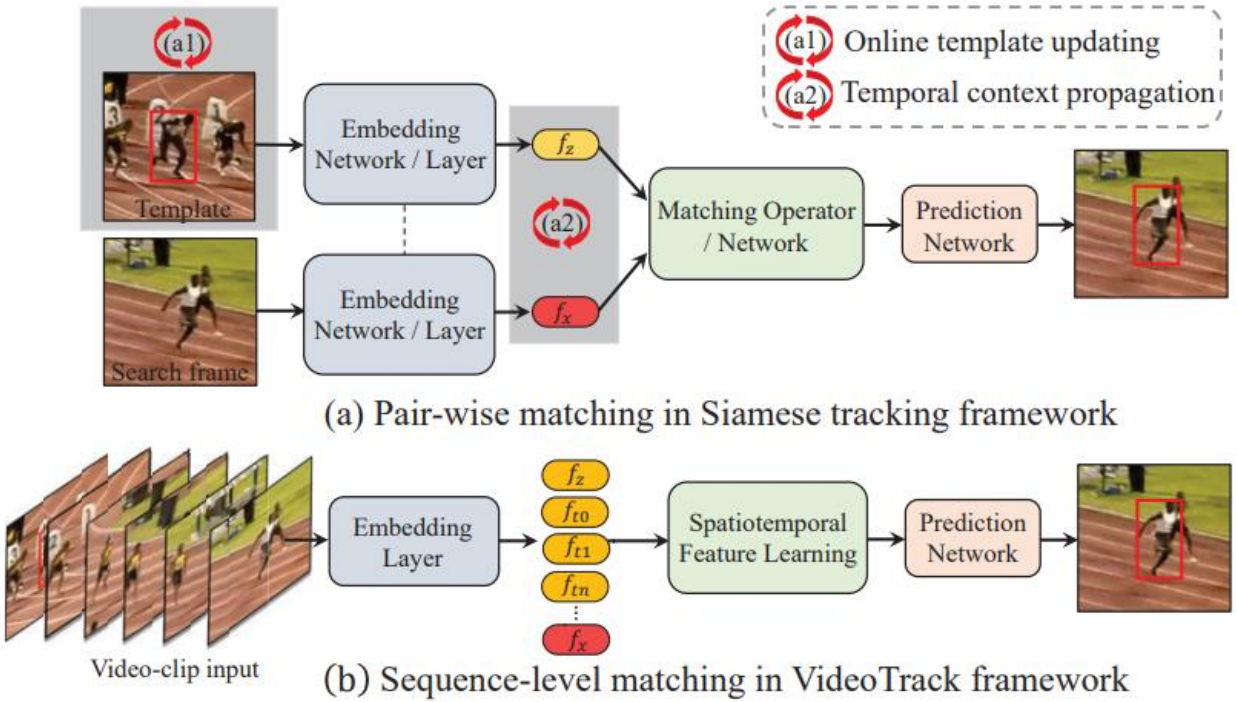
(b) Temporal context propagation (TrSiam/TrDiMP)

Traditional Siamese matching based trackers requires sophisticated mechanisms to exploit temporal contexts



(c) Video Backbone Based Tracking (VideoTrack)

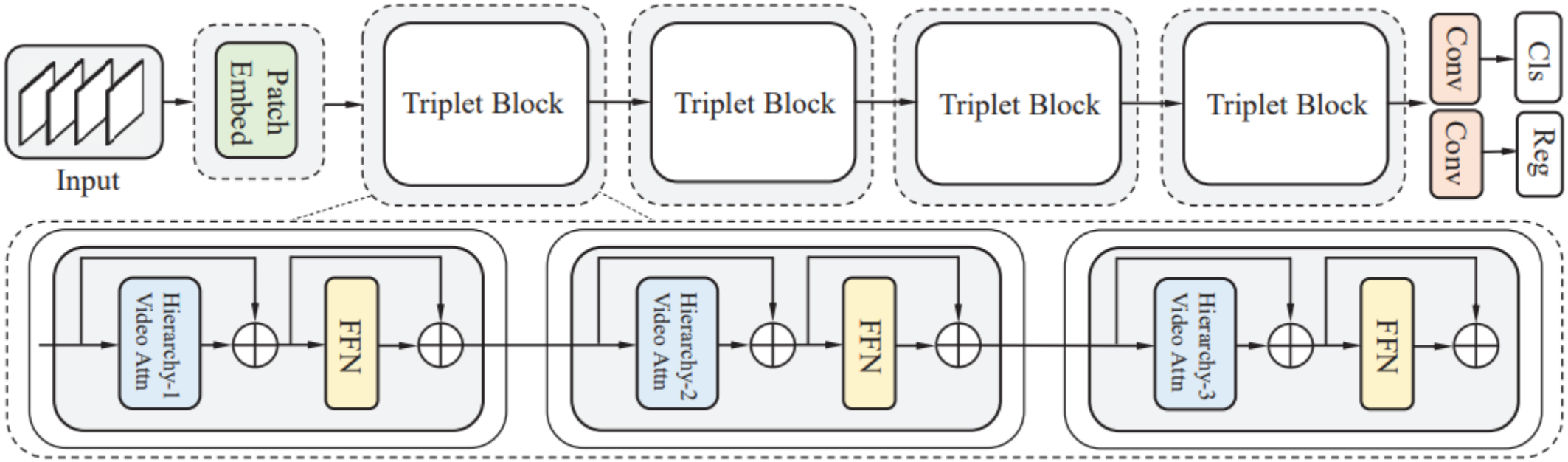
Introduction



Comparing to traditional Siamese matching network trackers:

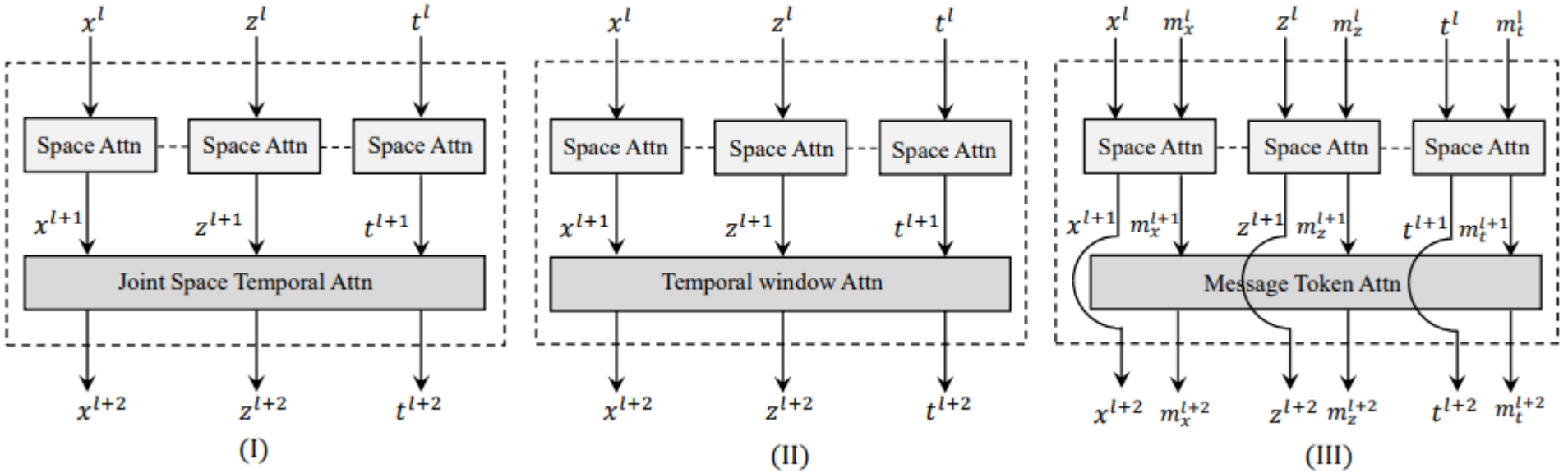
- 2D pair-wise matching is lifted to spatiotemporal domain, encoding temporal context at the feature-level via a neat feedforward video model.
- The first to adopt Video Transformer based tracking framework.
- VideoTrack exhibits encouraging results on multiple VOT benchmarks.

Overall framework of VideoTrack



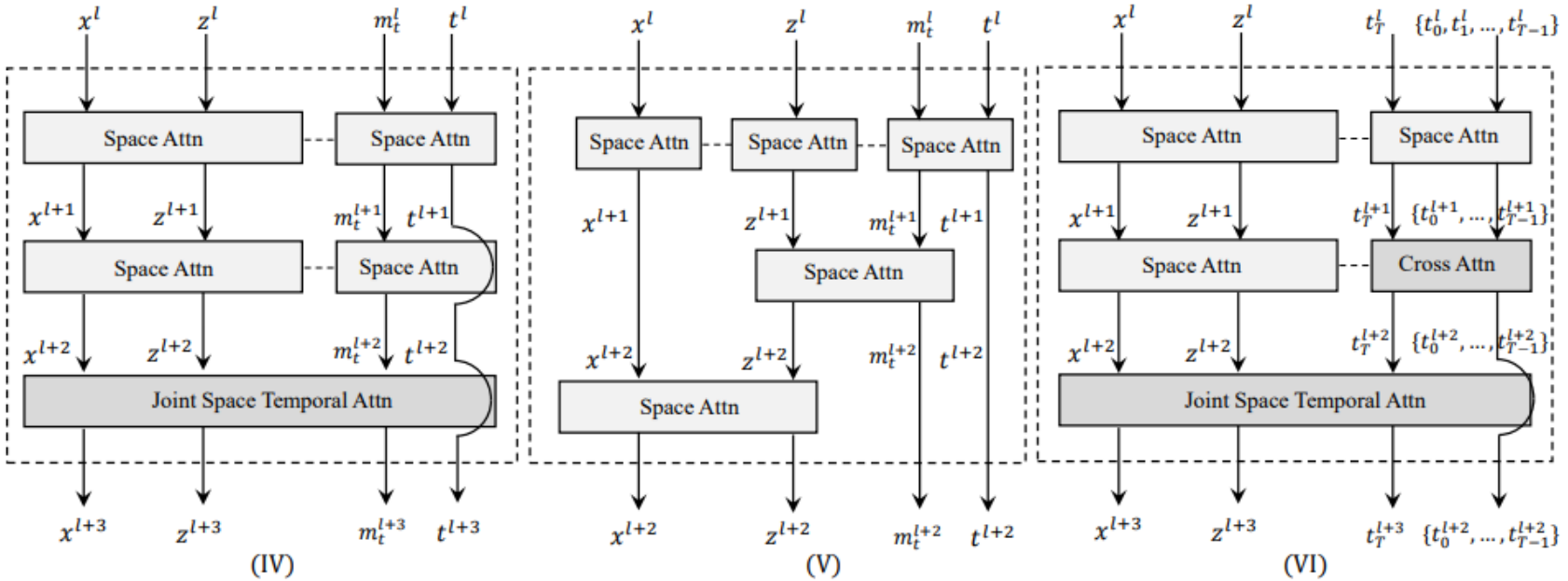
Overall architecture of video transformer model for tracking (VideoTrack). It is constructed by stacking multiple basic building units, named as triplet-block which consists of three hierarchical attention layers. Inside each layer of triplet-block, video attention module mixes the multi-branch information flows among inputs asymmetrically.

Temporal modelling inside the transformer block



Three basic temporal modelling methods in the building unit with two-layer structure : (I) denotes the divided space time pattern; (II) denotes the temporal window pattern; (III) denotes the pattern with message token for temporal modelling.

Temporal modelling inside the transformer block



Disentangled Dual-Template Mechanism,

- Leverage the strong static appearance information from the first template and the dynamic factors of the intermediate templates through the efficient temporal modelling.
- It reduces the computation & temporal redundancy
- Propagating the appearance information more thoroughly than the message token communication.

Empirical study on model choices

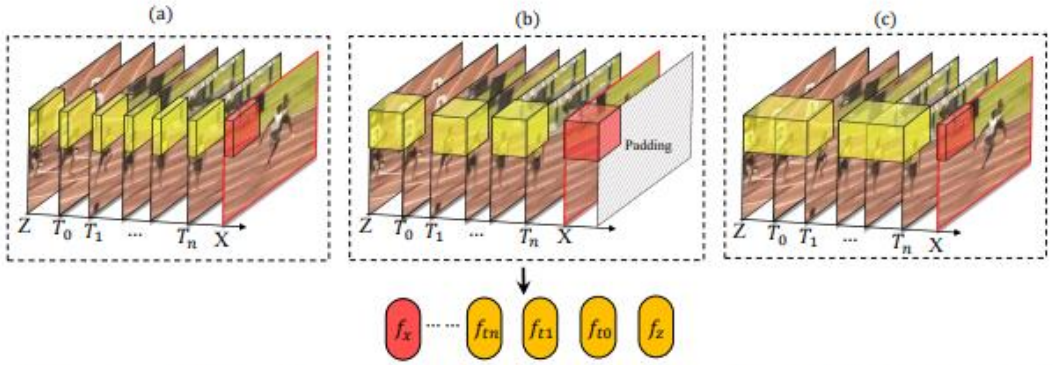


Figure 3. Three different embedding methods. (a) separated embedding. (b) tubelet embedding. (c) combination of separated (search frame) and tubelet embedding (templates).

Extensive model designs are investigated :

token embedding,

video attention scheme, temporal modelling,

varying the number of input frames,

pretraining and Foreground label map.

Empirical study on model choices

case	Pattern I	Pattern VI
Separate	**71.3**	**70.6**
Tubelet for t	66.4	65.8
Tubelet	65.3	63.1

Table 1. Ablations on input encoding. The performance is AO in GOT-10k [24].

case	Pattern I	Pattern VI
None	70.1	69.5
Space-only	70.5	69.8
Space-Time	**71.3**	**70.7**

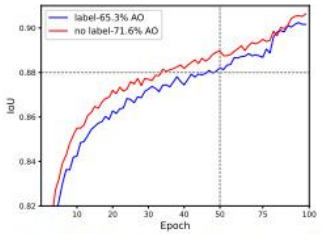
Table 2. Ablations on positional embedding. The performance is AO in GOT-10k [24].

case	Pattern I	Flops	Param.
Vanilla ViT [14] Attn	71.3	67.7 G	**85.4** M
VideoSwin [32] Attn	69.2	**49.6** G	90.9 M
Trajectory [38] Attn	**71.5**	70.5 G	**85.4** M

Table 3. Ablations on video attention design. The performance is AO in GOT-10k [24].

case	Pattern I	Pattern VI
None	62.1	60.5
ImageNet-1k [14]	70.4	70.2
ImageNet-21k [14]	71.2	70.8
VideoMAE [41]	64.1	63.5
MAE [21]	**73.3**	**72.6**

Table 4. Ablations on pretrain. MAE pre-training is more effective. The performance is AO in GOT-10k [24].



case	Pattern I
Label	65.3
No label	**71.6**

Figure 6. Ablations on foreground label. It shows the IoU curve during training and their best performance in GOT-10k [24].

VOT benchmark results

	SiamRPN++ [26]	ATOM [12]	DiMP [5]	Tr Siam [47]	TransT [7]	Stark st101 [58]	Mixformer 1k large [10]	SBT large [16]	OStrack 256 [61]	VideoTrack
$AO\uparrow$	51.8	55.6	61.1	66.0	67.1	68.8	67.9	**70.4**	**71.0**	**72.9**
$SR_{50}\uparrow$	61.6	63.4	71.7	76.6	76.8	78.1	77.3	**80.8**	**80.4**	**81.9**
$SR_{75}\uparrow$	32.5	40.2	49.2	57.1	60.9	64.1	63.9	**64.7**	**68.2**	**69.8**

Table 7. Comparison on the GOT-10k [24] test set.

	SiamRPN++ [26]	ATOM [12]	DiMP [5]	Tr Siam [55]	TransT [47]	DTT [7]	Stark s50 [62]	Mixformer 1k [58]	SBT large [10]	Ostrack 256 [16]	VideoTrack [61]
AUC↑	49.6	51.5	56.9	62.4	64.9	60.1	65.8	**67.9**	66.7	**69.1**	**70.2**
Prec↑	49.1	50.5	56.7	60.0	69.0	-	69.7	**73.9**	71.1	**75.2**	**76.4**

Table 8. Comparison on the LaSOT [15] test set.

	ECO [11]	SiamFC [4]	SiamFC++ [57]	PrDiMP [13]	D3S [33]	AutoMatch [65]	TransT [7]	Stark st50 [58]	Mixformer 1k [10]	Ostrack 256 [61]	VideoTrack
AUC↑	55.4	57.1	75.4	75.8	72.8	76.0	81.4	81.3	**82.6**	**83.1**	**83.8**
Norm.Prec↑	61.8	66.3	80.0	81.6	76.8	82.4	86.7	86.1	**87.7**	**87.8**	**88.7**
Prec↑	49.2	53.3	70.5	70.4	66.4	72.5	80.3	-	**81.2**	**82.0**	**83.1**

Table 9. Comparison on the TrackingNet [36] test set.

	ATOM [12]	Ocean [67]	AutoMatch [65]	SiamGAT [19]	Tr DiMP [47]	TransT [7]	Stark st50 [58]	Mixformer 1k [10]	OStrack 256 [61]	VideoTrack
AUC↑	61.7	62.1	64.4	64.6	67.0	68.1	**69.2**	**68.7**	68.3	**69.7**
Prec↑	82.7	82.3	83.8	84.3	87.6	87.6	**88.2**	**89.5**	-	**89.9**

Table 10. Comparison on the UAV123 [35] test set.

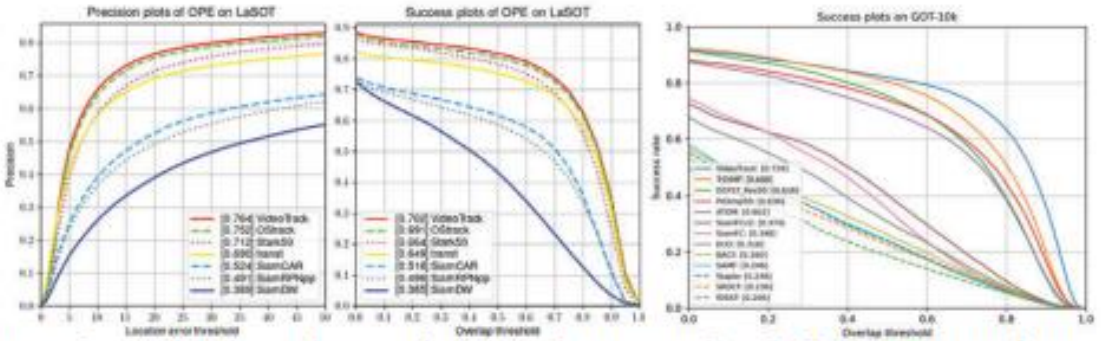


Figure 8. AUC and precision plots on LaSOT [15] and success plot on GOT-10k [24]. Better viewed with zooming in.

VideoTrack exhibits encouraging results on multiple VOT benchmarks, including LaSOT, GOT-10k, LaSOT and UAV123.

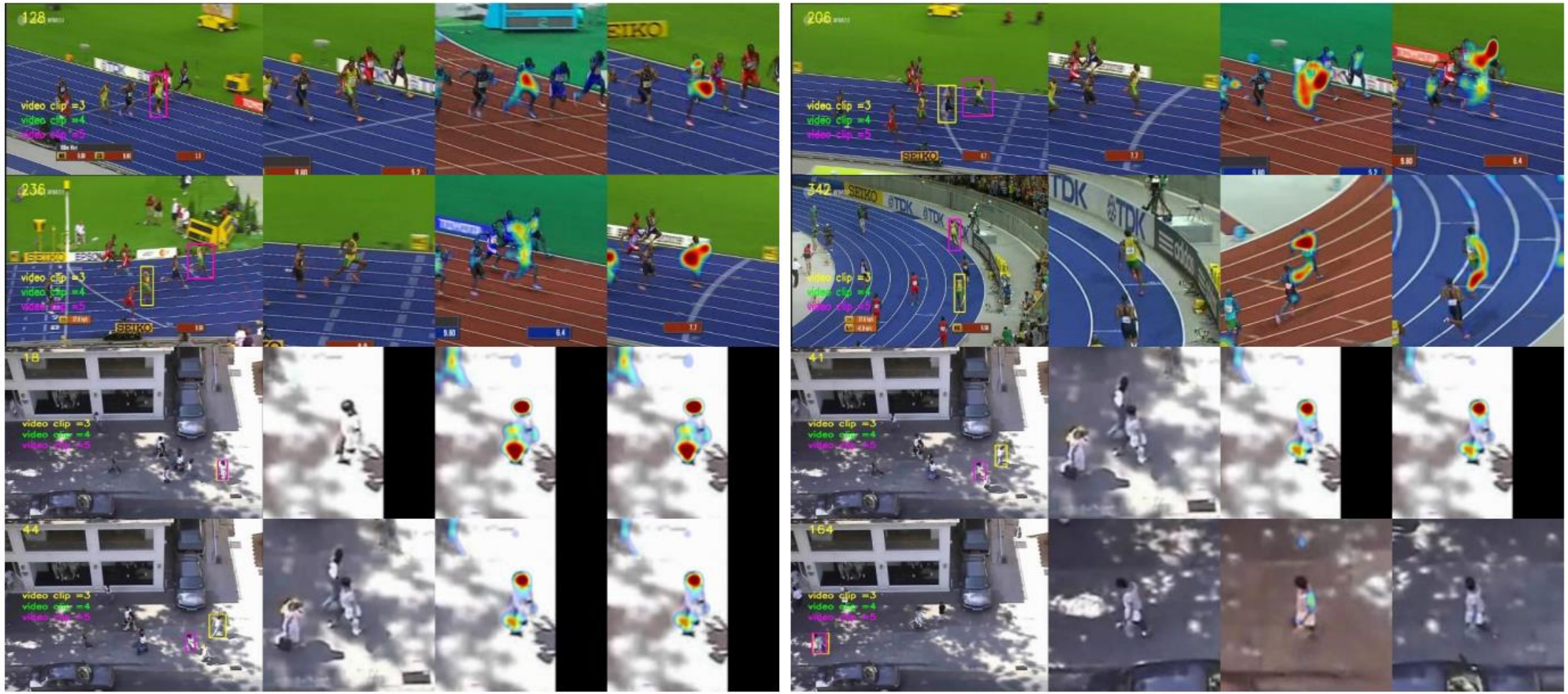
Qualitative results

- Over-long temporal extends do not provide more useful clues but redundancy. It is consistent with our observation that the appearance clue plays a more important role than motion clue in the matching.

- Longer video sequences can help tracker to handle the challenging scenarios, such as appearance variations, occlusion and similar distractor objects.

Qualitative results

Qualitative results



Ending