ICCV23 PARIS

Robust Object Modeling for Visual Tracking

Yidong Cai, Jie Liu*, Jie Tang, Gangshan Wu

State Key Laboratory for Novel Software Technology, Nanjing University, China

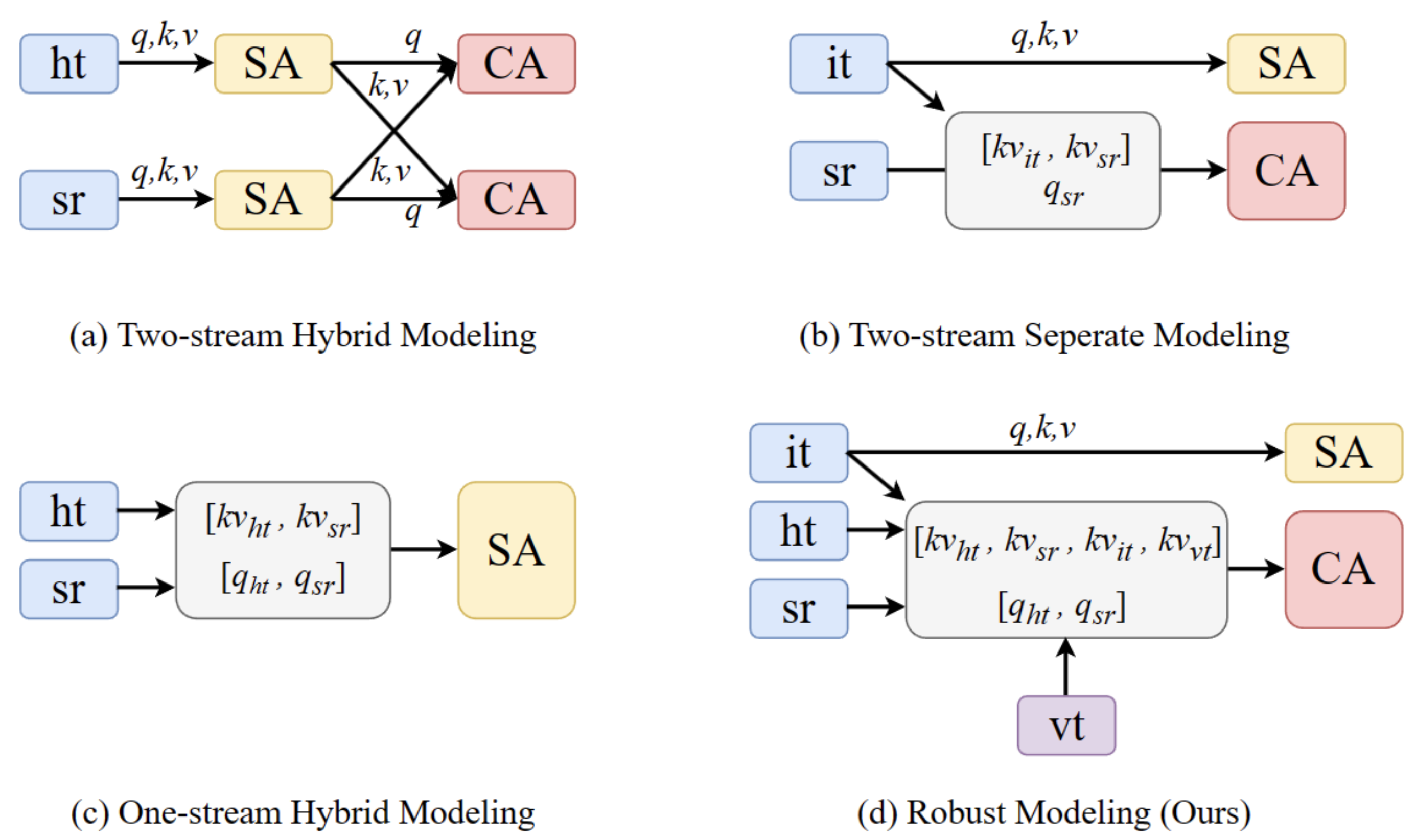
NANJING UNIVERSITY

https://arxiv.org/abs/2308.05140

https://github.com/dawnyc/ROMTrack

Introduction

- Object Modeling and Robustness of visual tracking are two core parts of recent tracking framworks.
- We compare 3 typical object modeling methods for template-search feature learning in Figure 1, together with our Robust Modeling design.



(a) Two-stream Hybrid Modeling

(b) Two-stream Seperate Modeling

(c) One-stream Hybrid Modeling

(d) Robust Modeling (Ours)

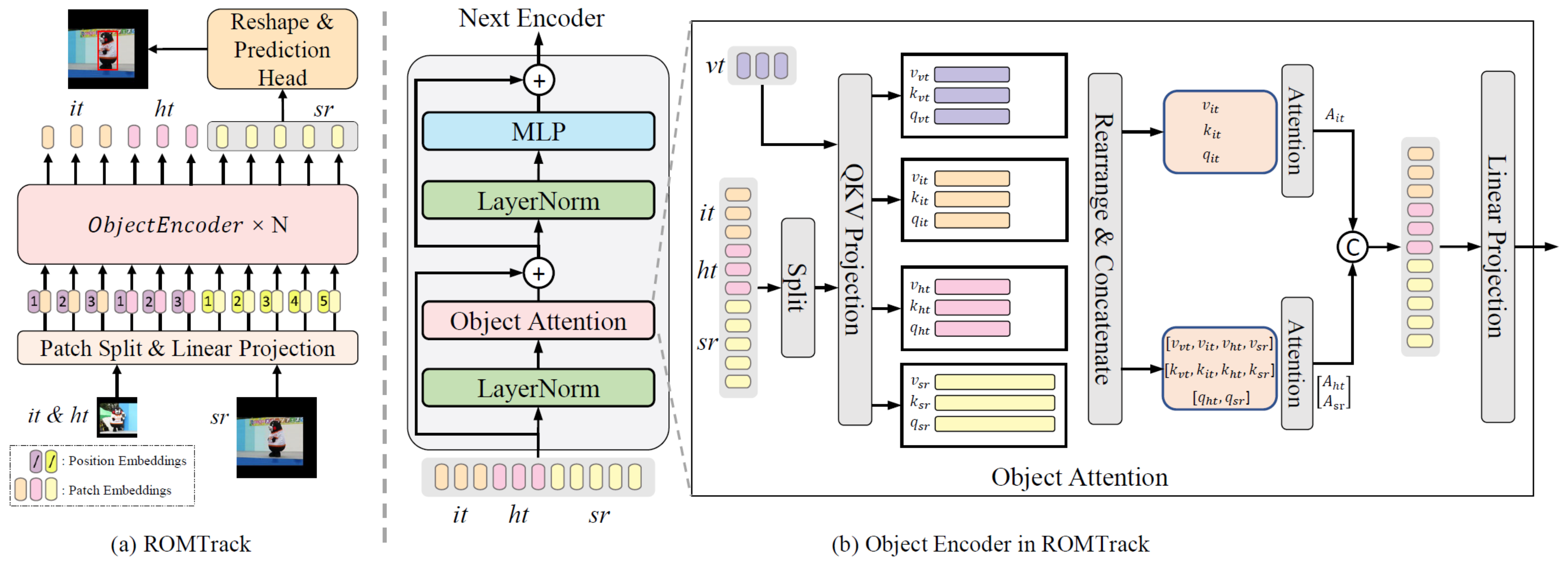
Figure 1: ht, it, sr, and vt represent hybrid template, inherent template, search region, and variation tokens, respectively. SA and CA denote self-attention and cross-attention.

Contribution

- We propose ROMTrack, which can keep the inherent information of the target template and enables mutual feature matching between the target and the search region simultaneously.
- We present a neat and effective variation-token design that embeds the appearance context during tracking into the attention calculation of hybrid features.
- The proposed ROMTrack sets a new state-of-the-art performance on 6 challenging benchmarks.

Technical Details

- ROMTrack is a Transformer-based tracker with N Object Encoder layers. Each encoder layer performs object attention for robust tracking, as shown in Figure 2.


Reshape & Prediction Head
it
ht
sr
ObjectEncoder × N
1 2 3 1 2 3 1 2 3 4 5
Patch Split & Linear Projection
it & ht
sr
: Position Embeddings
: Patch Embeddings
Next Encoder
MLP
LayerNorm
Object Attention
LayerNorm
it
ht
sr
vt
it
ht
sr
Split
QKV Projection
v_{vt} k_{vt} q_{vt}
v_{it} k_{it} q_{it}
v_{ht} k_{ht} q_{ht}
v_{sr} k_{sr} q_{sr}
Rearrange & Concatenate
v_{it} k_{it} q_{it}
$[v_{vt}, v_{it}, v_{ht}, v_{sr}]$
$[k_{vt}, k_{it}, k_{ht}, k_{sr}]$
$[q_{ht}, q_{sr}]$
Attention
Attention
A_{it}
$\begin{bmatrix} A_{ht} \\ A_{sr} \end{bmatrix}$
C
Linear Projection
Object Attention

(a) ROMTrack

(b) Object Encoder in ROMTrack

Figure 2: (a) Overview of the proposed ROMTrack framework. The template and search region images are split into patches, and then linearly projected, concatenated, and fed into stacked encoder layers for robust object modeling. it, ht, and sr denote the inherent template, the hybrid template, and the search region, respectively. (b) Architecture of the object encoder layer. vt denotes variation tokens.

- Variation tokens encode the contextual appearance changes to tackle the problem of object deformation and appearance variations, as shown in Figure 3.

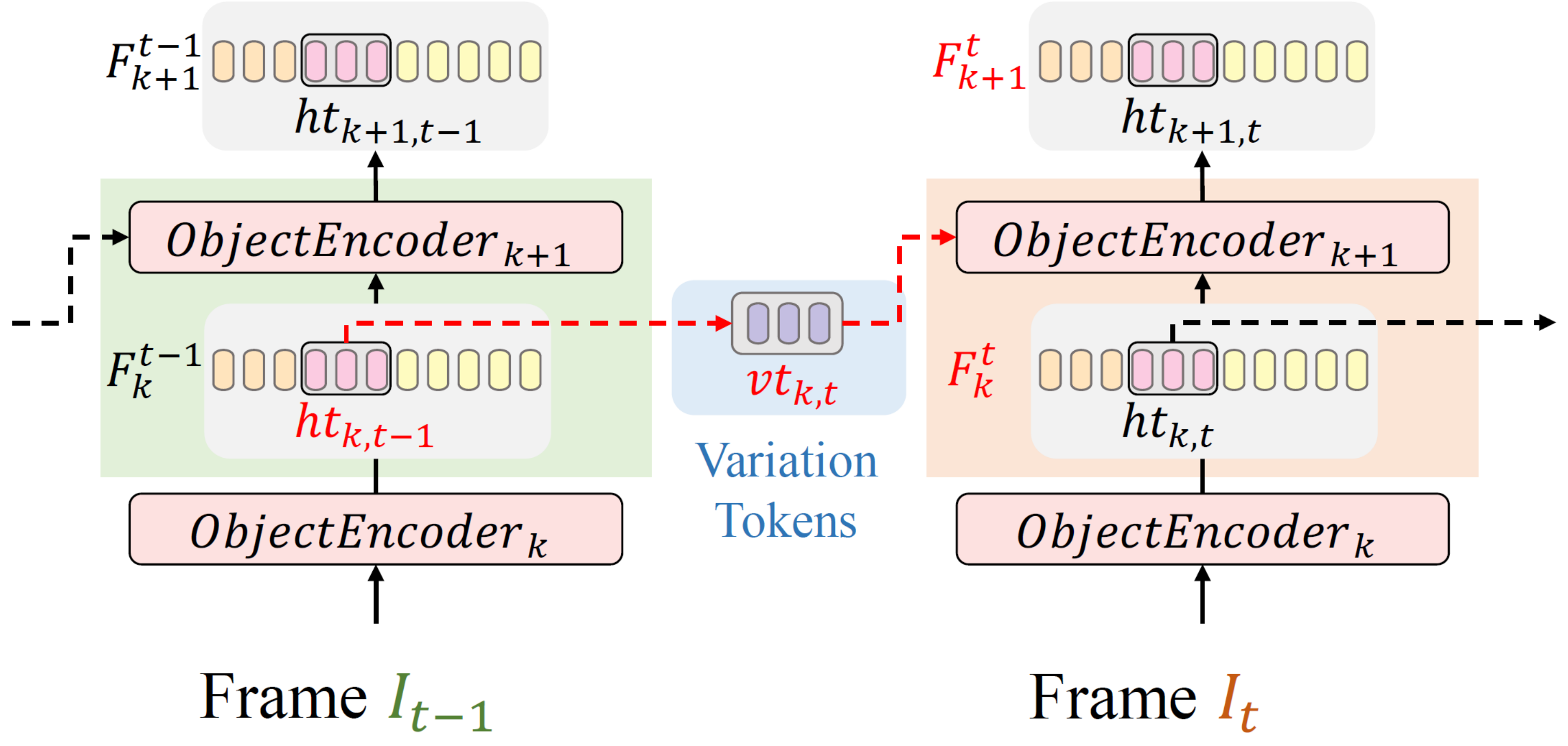


Figure 3: Schema of the proposed variation-token design.

- For inference, variation tokens are obtained per frame and employed for subsequent tracking procedure.

Experiments

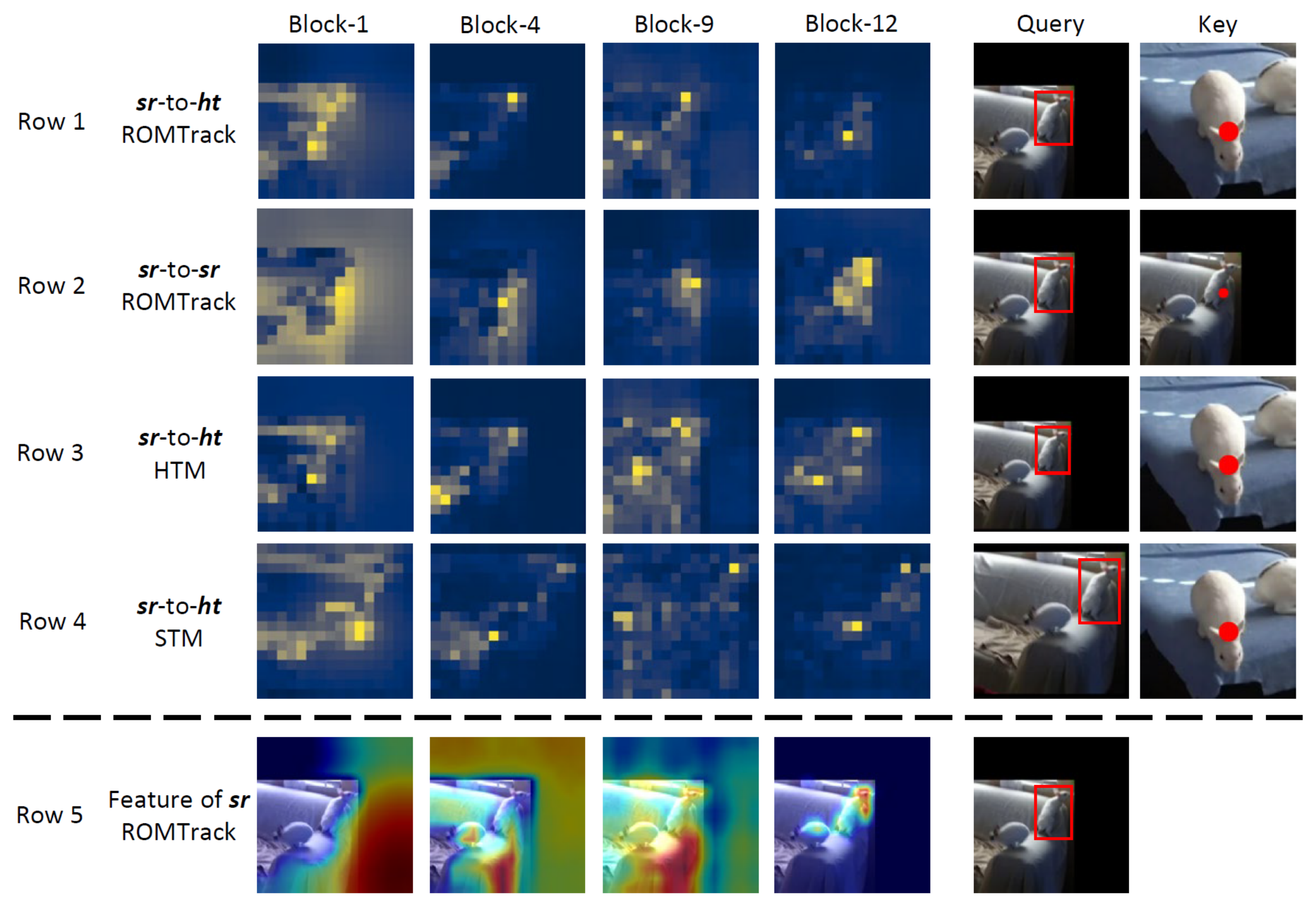
- Evaluation Metrics

Method	Source	GOT-10k*			LaSOT			TrackingNet			$\text{LaSOT}_{\text{ext}}$		
		$AO(\%)$	$SR_{0.5}(\%)$	$SR_{0.75}(\%)$	$AUC(\%)$	$P_{Norm}(\%)$	$P(\%)$	$AUC(\%)$	$P_{Norm}(\%)$	$P(\%)$	$AUC(\%)$	$P_{Norm}(\%)$	$P(\%)$
ROMTrack	Ours	**72.9**	**82.9**	**70.2**	**69.3**	78.8	**75.6**	**83.6**	**88.4**	**82.7**	**48.9**	**59.3**	**55.0**
SwinTrack-T-224 [31]	NIPS22	**71.3**	**81.9**	64.5	67.2	-	70.8	81.1	-	78.4	47.6	-	53.9
OSTrack-256 [54]	ECCV22	71.0	80.4	**68.2**	69.1	78.7	**75.2**	83.1	87.8	82.0	47.4	57.3	53.3
OSTrack-256(w/o CE) [54]	ECCV22	71.0	80.3	68.2	68.7	78.1	74.6	82.9	87.5	81.6	-	-	-
AiATrack [23]	ECCV22	69.6	80.0	63.2	69.0	**79.4**	73.8	82.7	87.8	80.4	46.8	54.4	**54.2**
SimTrack-B/16 [5]	ECCV22	68.6	78.9	62.4	**69.3**	78.5	74.0	82.3	86.5	-	-	-	-
Unicorn [52]	ECCV22	-	-	-	68.5	76.6	74.1	83.0	86.4	**82.2**	-	-	-
MixFormer-22k [10]	CVPR22	70.7	80.0	67.8	69.2	78.7	74.7	**83.1**	**88.1**	81.6	-	-	-
MixFormer-1k [10]	CVPR22	71.2	79.9	65.8	67.9	77.3	73.9	82.6	87.7	81.2	-	-	-
ToMP50 [36]	CVPR22	-	-	-	67.6	78.0	72.2	81.2	86.2	78.6	45.4	57.6	-
ToMP101 [36]	CVPR22	-	-	-	68.5	**79.2**	73.5	81.5	86.4	78.9	45.9	**58.1**	-
SBT-large [50]	CVPR22	70.4	80.8	64.7	66.7	-	71.1	-	-	-	-	-	-
KeepTrack [37]	ICCV21	-	-	-	67.1	77.2	70.2	-	-	-	**48.2**	58.0	-
STARK [53]	ICCV21	68.8	78.1	64.1	67.1	77.0	-	82.0	86.9	-	-	-	-
DTT [55]	ICCV21	63.4	74.9	51.4	60.1	-	-	79.6	85.0	78.9	-	-	-
TransT [6]	CVPR21	67.1	76.8	60.9	64.9	73.8	69.0	81.4	86.7	80.3	45.1	51.3	51.2
TrDiMP [44]	CVPR21	67.1	77.7	58.3	63.9	-	61.4	78.4	83.3	73.1	-	-	-
LTMU [11]	CVPR20	-	-	-	57.2	-	57.2	-	-	-	41.4	49.9	47.3
SiamR-CNN [43]	CVPR20	64.9	72.8	59.7	64.8	72.2	-	81.2	85.4	80.0	-	-	-
Ocean [58]	ECCV20	61.1	72.1	47.3	56.0	65.1	56.6	-	-	-	-	-	-
DiMP [3]	ICCV19	61.1	71.7	49.2	56.9	65.0	56.7	74.0	80.1	68.7	39.2	47.6	45.1
SiamRPN++ [29]	CVPR19	51.7	61.6	32.5	49.6	56.9	49.1	73.3	80.0	69.4	34.0	41.6	39.6
MDNet [39]	CVPR16	29.9	30.3	9.9	39.7	46.0	37.3	60.6	70.5	56.5	27.9	34.9	31.8
SiamFC [2]	ECCV16	34.8	35.3	9.8	33.6	42.0	33.9	57.1	66.3	53.3	23.0	31.1	26.9
Trackers with Higher Resolution or Larger Model													
ROMTrack-384	Ours	**74.2**	**84.3**	**72.4**	**71.4**	**81.4**	**78.2**	**84.1**	**89.0**	**83.7**	**51.3**	**62.4**	**58.6**
SwinTrack-B-384 [31]	NIPS22	72.4	80.5	67.8	**71.3**	-	76.5	**84.0**	-	82.8	49.1	-	55.6
OSTrack-384 [54]	ECCV22	**73.7**	**83.2**	**70.8**	71.1	**81.1**	**77.6**	83.9	88.5	**83.2**	**50.5**	**61.3**	**57.6**
SimTrack-L/14 [5]	ECCV22	69.8	78.8	66.0	70.5	79.7	76.2	83.4	87.4	-	-	-	-
MixFormer-L [10]	CVPR22	-	-	-	70.1	79.9	76.3	83.9	**88.9**	83.1	-	-	-

- Performance Comparison

Method	Speed (FPS)	MACs (G)	Params (M)	LaSOT AUC(%)	GOT-10k* AO(%)
OSTrack-256 (w/o CE) [57]	65	29.0	92.1	68.7	71.0
MixFormer-22k [11]	25	23.0	35.6	69.2	70.7
ROMTrack	62	34.5	92.1	**69.3**	**72.9**
OSTrack-384 (w/o CE)	29	65.3	92.1	71.0	73.7
MixFormer-L	18	127.8	183.9	70.1	-
ROMTrack-384	28	77.7	92.1	**71.4**	**74.2**

- Visualization



Paper

Code