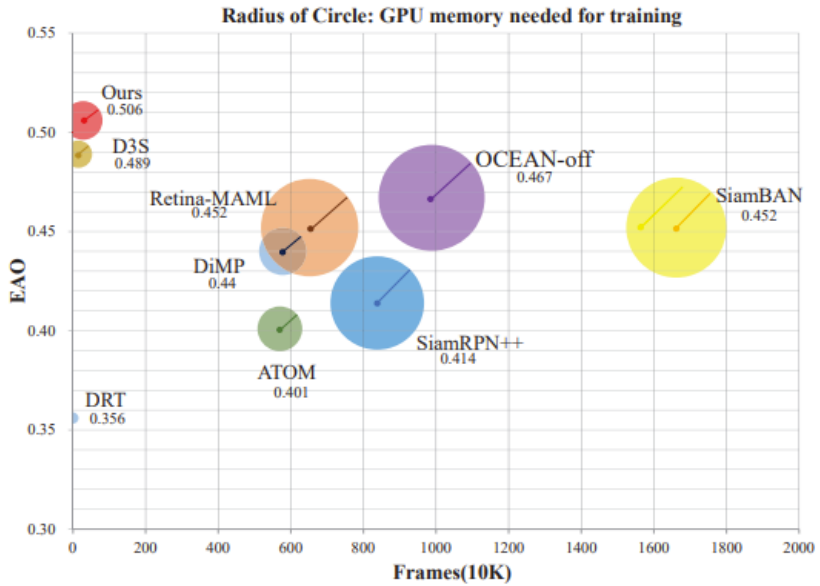
Learning Spatio-Appearance Memory Network for High-Performance Visual Tracking

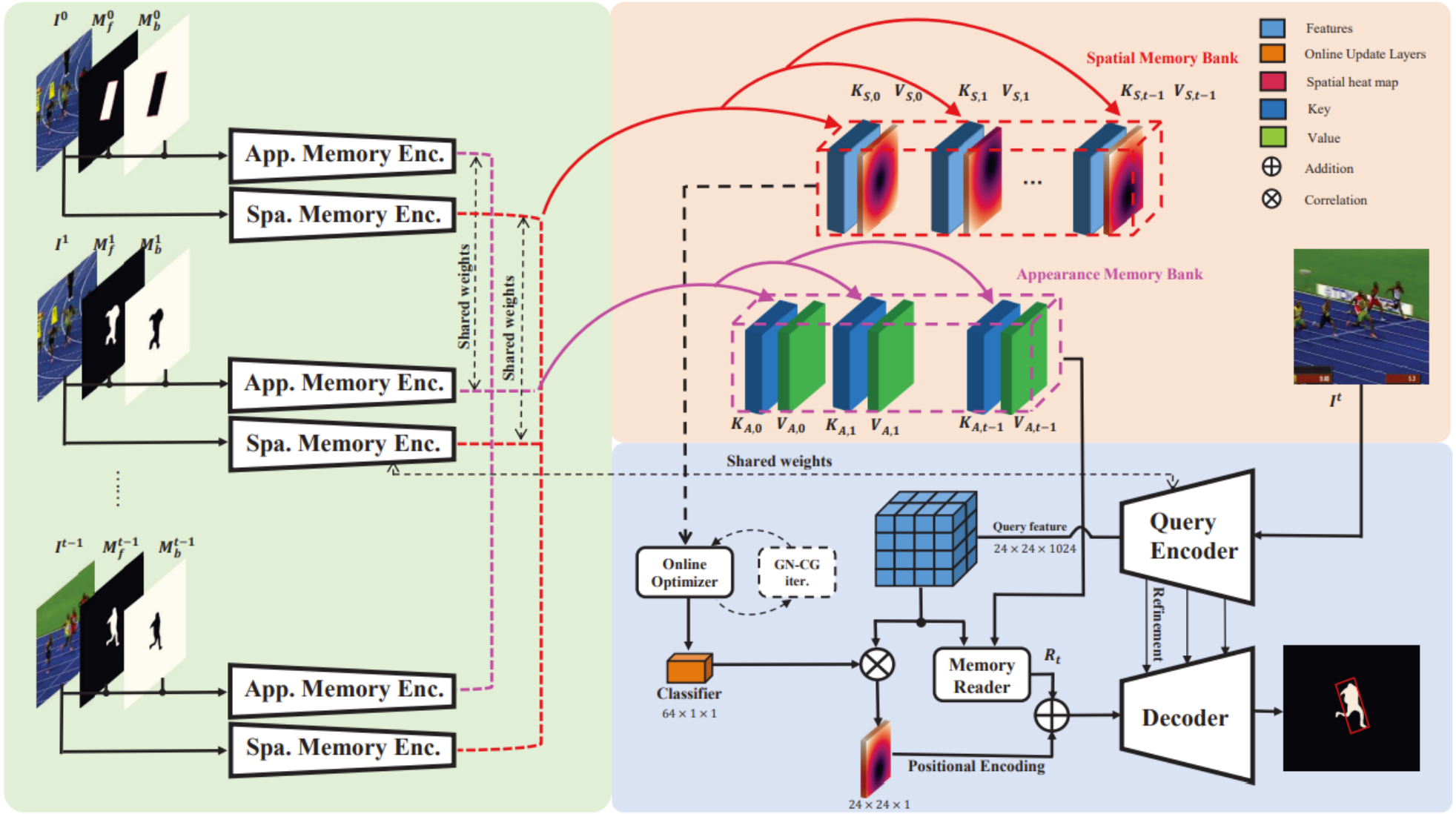
Fei Xie[1], Wankou Yang[1], Kaihua Zhang[2], Bo Liu[3], Guangting Wang[4], Wangmeng Zuo[5]
[1]School of Automation, Southeast University, China
[2]Nanjing University of Information Science and Technology, China
[3]JD Finance America Corporation, Mountain View, CA, USA
[4]University of Science and Technology of China
[5]School of Computer Science and Technology, Harbin Institute of Technology
jaffe03@seu.edu.cn, wkyang@seu.edu.cn, zhkhua@gmail.com, kfliubo@gmail.com
flylight@ustc.edu.cn, cswmzuo@gmail.com

Contributions

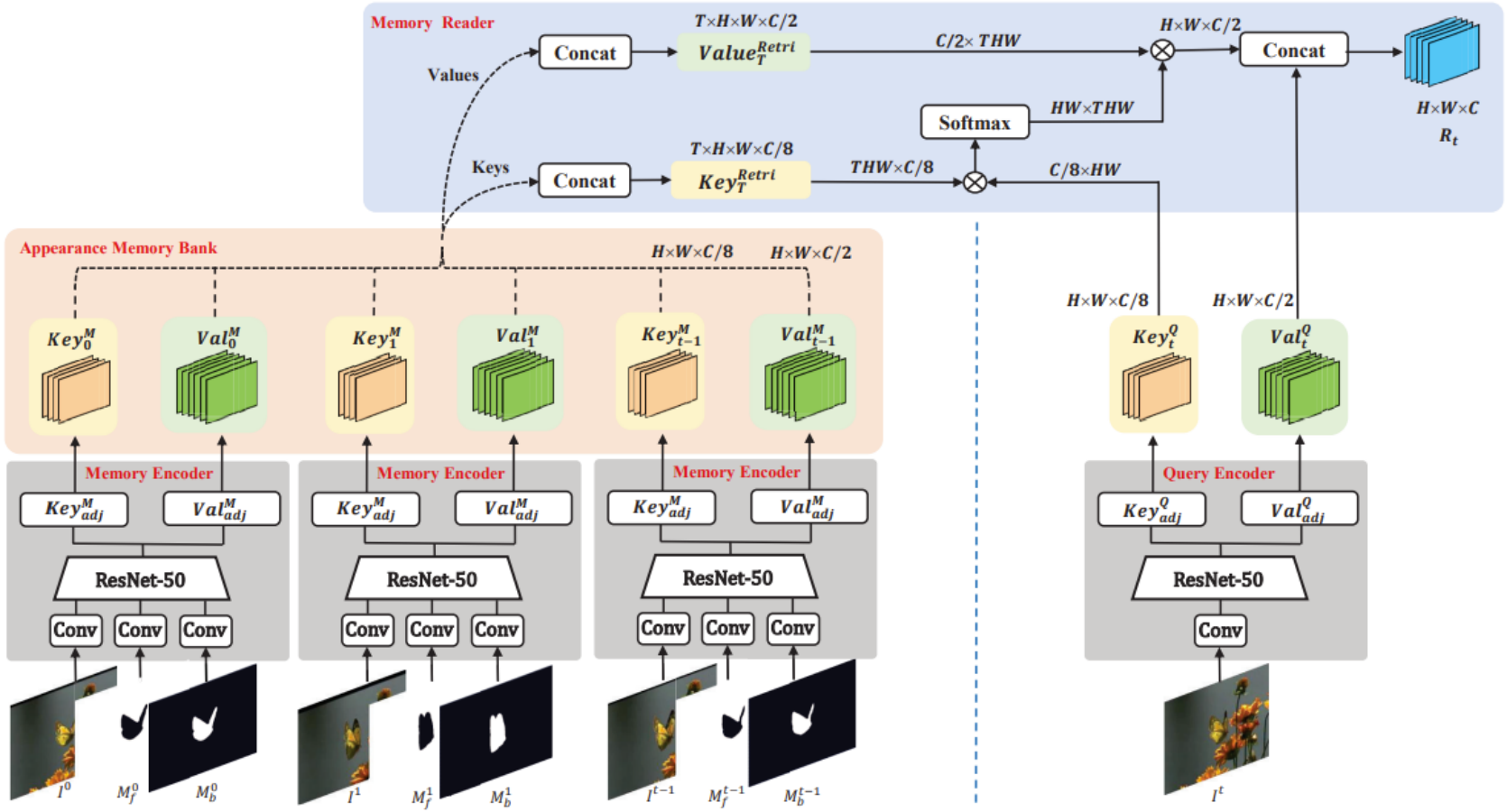
1. We adopt segmentation-based tracking model which is robust towards non-grid deformations.
2. Two memory banks for target regression and classification which both utilize temporal information.
3. A dynamic memory module to reweight memory samples.
4. Less training source to achieve high tracking performance.



Pipeline


I^0 M_f^0 M_b^0
App. Memory Enc.
Spa. Memory Enc.
I^1 M_f^1 M_b^1
App. Memory Enc.
Spa. Memory Enc.
I^{t-1} M_f^{t-1} M_b^{t-1}
App. Memory Enc.
Spa. Memory Enc.
Shared weights
Shared weights
Spatial Memory Bank
$K_{S,0}$ $V_{S,0}$ $K_{S,1}$ $V_{S,1}$ ... $K_{S,t-1}$ $V_{S,t-1}$
Appearance Memory Bank
$K_{A,0}$ $V_{A,0}$ $K_{A,1}$ $V_{A,1}$ $K_{A,t-1}$ $V_{A,t-1}$
Features
Online Update Layers
Spatial heat map
Key
Value
Addition
Correlation
I^t
Shared weights
Online Optimizer
GN-CG iter.
Query feature
24 × 24 × 1024
Query Encoder
Refinement
Classifier
64 × 1 × 1
Memory Reader
R_t
Positional Encoding
24 × 24 × 1
Decoder

Appearance memory network

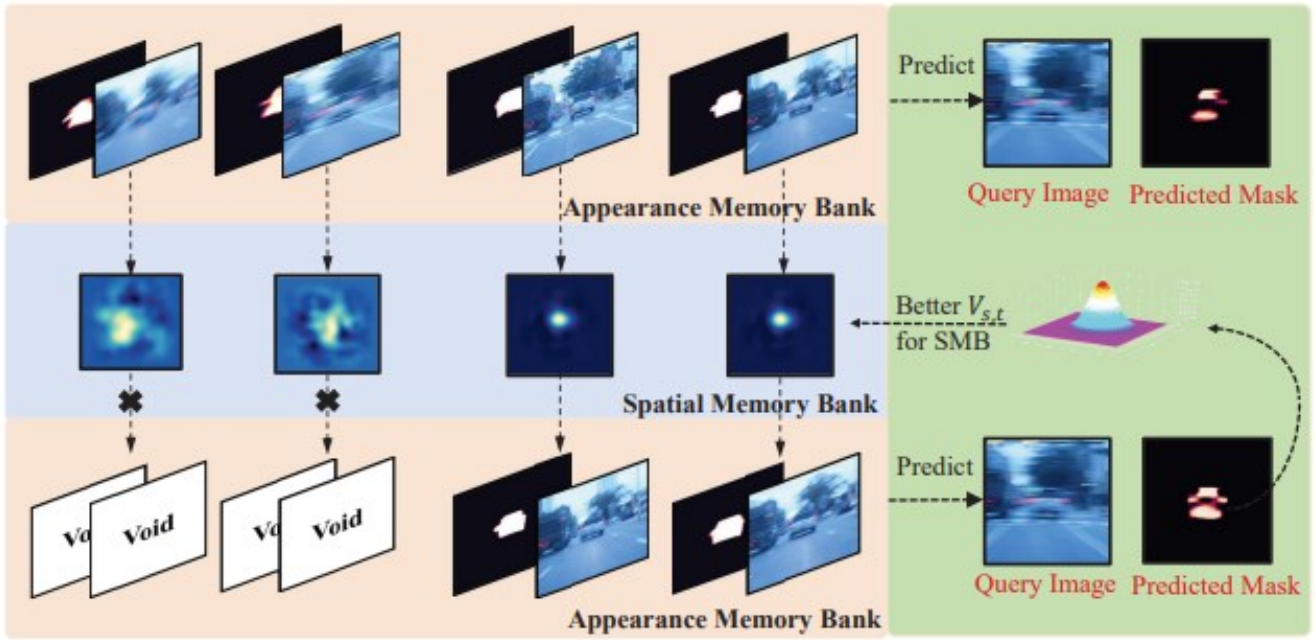


$$V_{A,t}^i = \sum_j \sum_{k=1}^{t-1} A_t^{i,j,k} V_{A,k}^j,$$

$$A_t^{i,j,k} = \frac{\exp\left\langle Q_t^i, K_{A,k}^j \right\rangle}{\sum_p \sum_{k=1}^{t-1} \exp\left\langle Q_t^i, K_{t-1}^p \right\rangle},$$

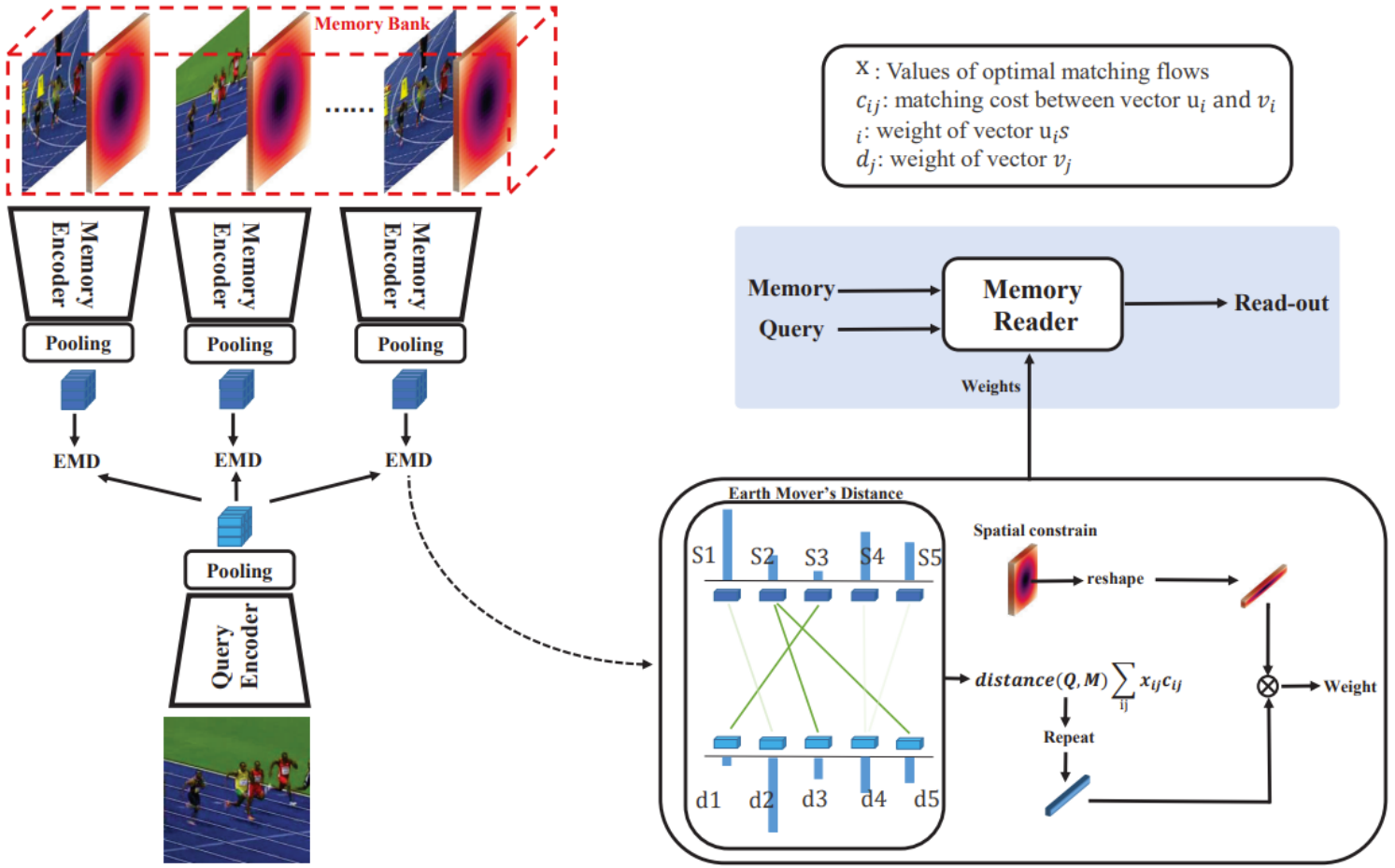
$$R_t = concat\left[V_{A,t}, V_{Q,t}\right],$$

Mutual Promotion Between Dual Memory Banks



Visualization on benefits from filtered samples.

Dynamic Memory Machine



$$s_a = \max\{\mathbf{u}_a^T \cdot \frac{\sum_{b=1}^{HW} \mathbf{v}_b}{HW}, 0\},$$

$$s_a := s_a \frac{HW}{\sum_{b=1}^{HW} s_b}.$$

Dynamic Memory Machine

Visualization of spatio-temporal weights in DRM. Transparent regions are more likely to be activated during query operation.

Comparisons on VOT benchmarks

Trackers		SPM [43]	SiamMask-opt [44]	SaimRPN++ [21]	ATOM [7]	D3S [27]	Ours (SAMN)
VOT-16	Acc.↑	0.62	0.67	0.64	0.61	0.66	0.684
	Rob.↓	0.21	0.23	0.20	0.18	0.131	0.121
	EAO↑	0.434	0.442	0.464	0.430	0.493	0.535

Table 1. Results on VOT2016. Top-3 results of each dimension (row) are colored in red, blue and green, respectively.

Trackers		DiMP-50 [1]	SiamBAN [4]	D3S [27]	Ocean-off [54]	DCFST [56]	Ours (SAMN)
VOT-18	Acc.↑	0.590	0.597	0.597	0.64	0.598	0.652
	Rob.↓	0.203	0.152	0.178	0.150	0.169	0.145
	EAO↑	0.440	0.452	0.489	0.467	0.452	0.521

Table 2. Results on VOT2018. Top-3 results of each dimension (row) are colored in red, blue and green, respectively.

Trackers		SiamRPN++	ATOM	Retina-MAML [42]	SiamFCOT [19]	Ocean-off	Ours (SAMN)
VOT-19	Acc.↑	0.580	0.603	0.570	0.601	0.590	0.639
	Rob.↓	0.446	0.411	0.366	0.386	0.376	0.231
	EAO↑	0.292	0.292	0.313	0.350	0.327	0.408

Table 3. Results on VOT2019. Top-3 results of each dimension (row) are colored in red, blue and green, respectively.

Trackers		SiamRPN++	ATOM	DiMP-18	DiMP-50	D3S	Ocean-off	Ours (SAMN)
GOT-10K	$SR_{.75}$↑	32.5	40.2	44.6	49.2	46.2	-	52.2
	AO↑	51.8	55.6	57.9	61.1	59.7	59.2	61.5

Table 4. Results on GOT-10K. Top-3 results of each dimension (row) are colored in red, blue and green, respectively.

Trackers		SiamRPN++	ATOM	DiMP-50	Retina-MAML	D3S	Ours (SAMN)
TrackingNet	Prec.↑	69.4	64.8	68.7	-	66.4	69.7
	Norm. Prec.↑	80.0	77.1	80.1	78.6	76.8	79.4
	Succ.↑	73.3	70.3	74.0	69.8	72.8	74.2

Table 5. Results on TrackingNet. Top-3 results of each dimension (row) are colored in red, blue and green, respectively.

Trackers		SiamMask	STM	DET50 [19]	Ocean	D3S	Ours (SAMN)
VOT-20	Mask	✓	✓	✓	✓	✓	✓
	Acc.↑	0.624	0.751	0.679	0.693	0.699	0.720
	Rob.↓	0.648	0.574	0.787	0.754	0.769	0.794
	EAO↑	0.321	0.308	0.441	0.430	0.439	0.461

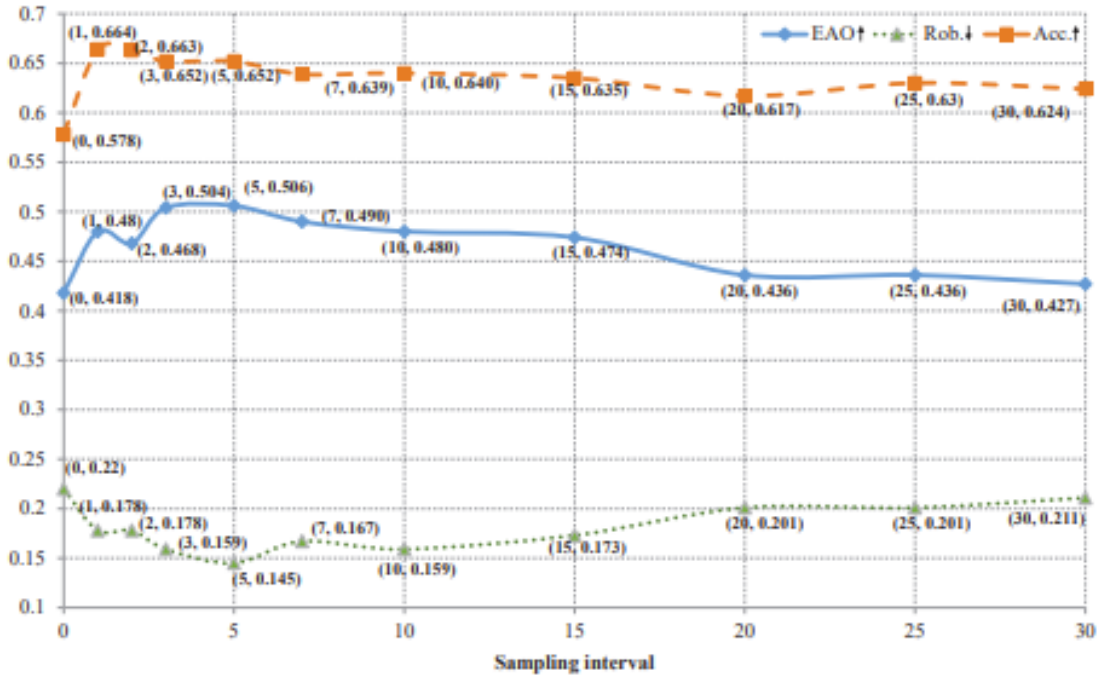
Table 6. Results on VOT2020. "Mask" denotes that prediction format is mask. Top-3 results of each dimension (row) are colored in red, blue and green, respectively.

Comparisons on VOS benchmarks

	$\mathcal{J}_{\mathcal{M}}^{16}$	$\mathcal{F}_{\mathcal{M}}^{16}$	$\mathcal{J}\&\mathcal{F}^{16}$	$\mathcal{J}_{\mathcal{M}}^{17}$	$\mathcal{F}_{\mathcal{M}}^{17}$	$\mathcal{J}\&\mathcal{F}^{17}$
Ours(SAMN)	79.0	75.5	77.3	64.8	67.7	66.3
D3S [27]	75.4	72.6	74.0	57.8	63.8	60.8
SiamMask [44]	71.7	67.8	69.8	54.3	58.5	56.4
OnAVOS [40]	86.1	84.9	85.5	61.6	69.1	65.4
STM [31]	84.8	88.1	86.4	69.2	74.0	71.6
MAST [20]	-	-	-	63.3	67.6	65.5
FAVOS [5]	82.4	79.5	80.9	54.6	61.8	58.2
VM [14]	81.0	-	-	56.6	-	-
OSVOS [3]	79.8	80.6	80.2	56.6	63.9	60.3
PLM [38]	75.5	79.3	77.4	-	-	-
OSMN [51]	74.0	72.9	73.5	52.5	57.1	54.8

Comparison with segmentation-based trackers and VOS methods on DAVIS16 and DAVIS17.

Ablation studies



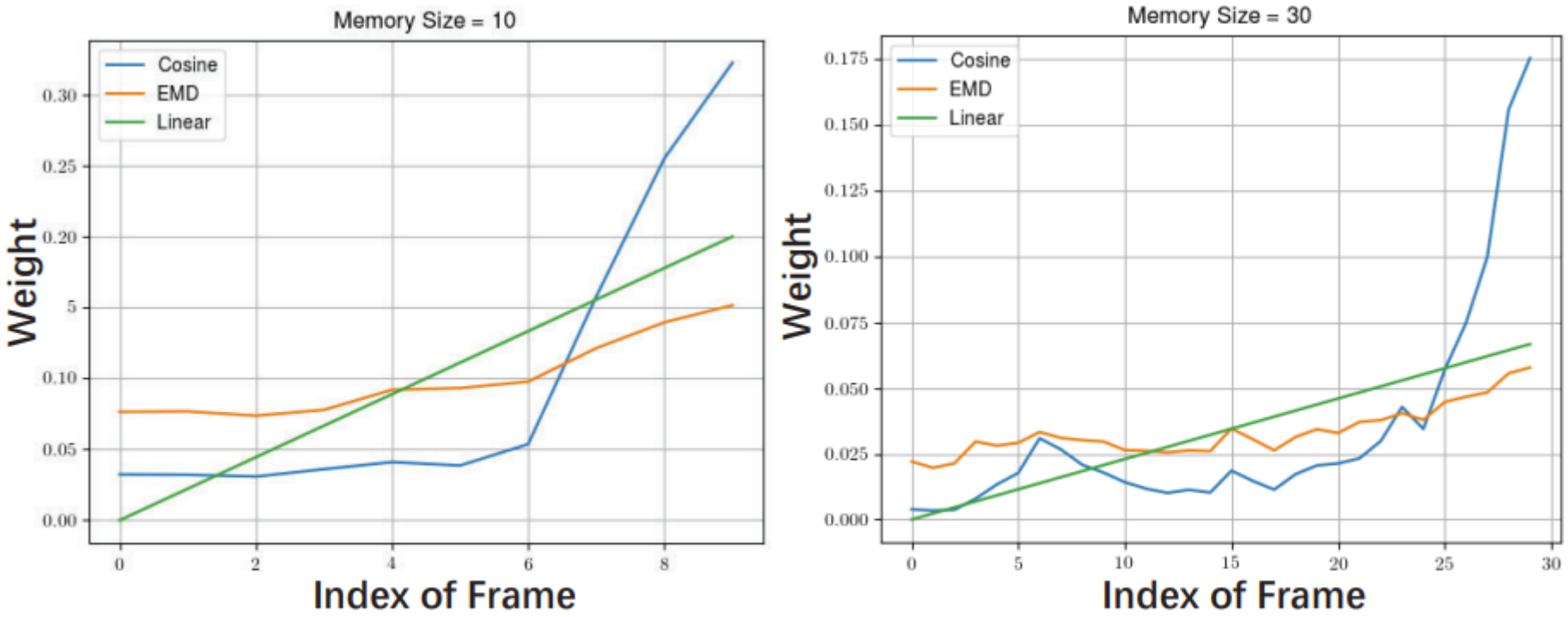
Time interval indicates the sampling interval of memory bank. Zero interval indicates that only the first frame and its ground truth is stored. Up-arrow (down-arrow) indicates higher (lower) is better.

Ablation studies

Ablation study on VOT2018 and DAVIS16.

Last Add.	✓		✓	✓	✓	✓	✓	✓	✓
Interv.	5	5	5	10	5	5	15	20	5
Filter Samp.	✓	✓	✓	✓		✓			✓
Pos. Encod.	sum	sum	sum	sum	sum	cat	sum	sum	sum
DRM	✓	✓	✓	✓	✓	✓	✓	✓	
A ↑	65.2	66.5	63.5	64.0	62.7	65.0	62.2	62.0	63.4
R ↓	0.145	0.173	0.164	0.159	0.210	0.150	0.225	0.227	0.173
EAO ↑	50.6	46.7	49.2	48.0	42.1	48.6	41.0	40.2	45.1
$\mathcal{J}\&\mathcal{F}^{16}$ ↑	77.3	70.3	-	67.8	69.1	-	66.4	63.6	67.2

Ablation studies



Comparisons of weights generation method.

Qualitative results

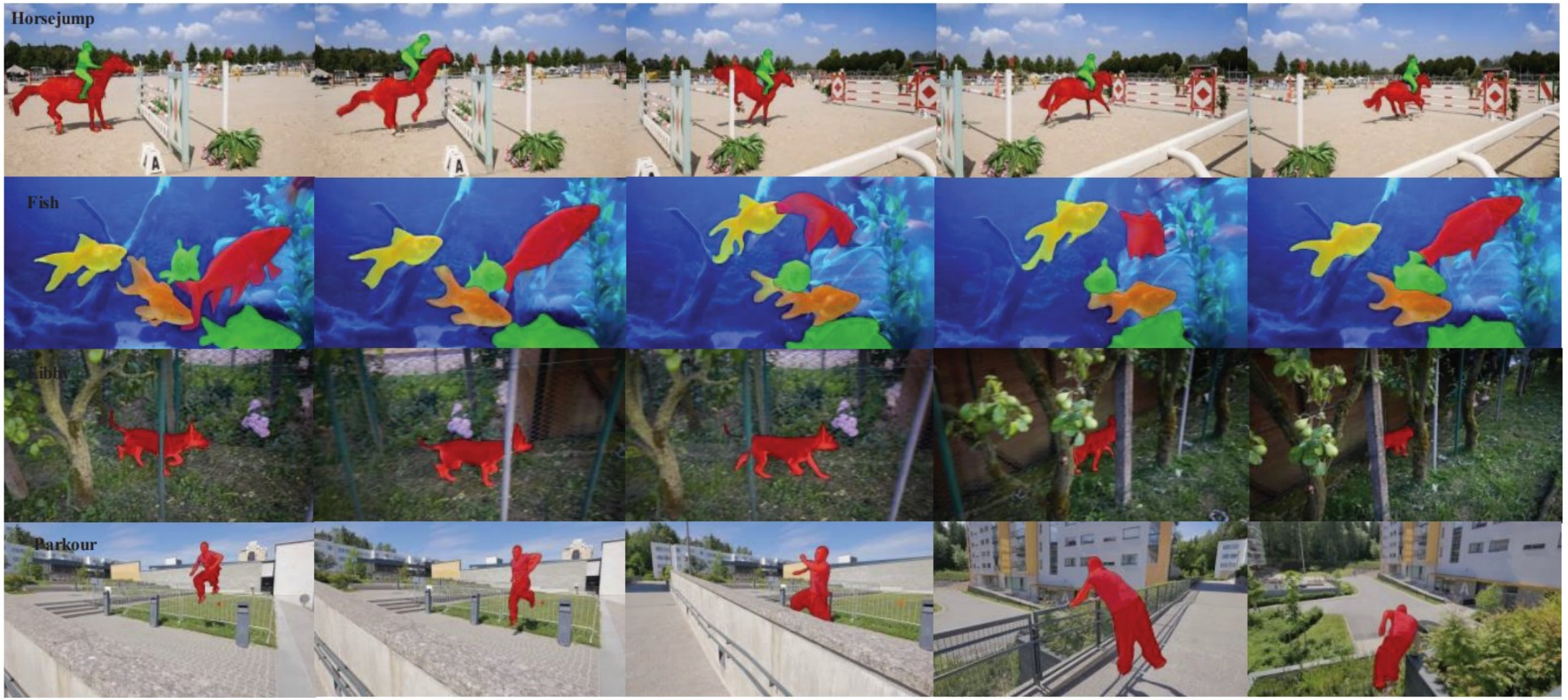


Figure 1: Video object segmentation on DAVIS16 or DAVIS17 datasets. Our tracker outperforms the leading segmentation-based trackers and predicts accurate masks under challenging scenes.

Qualitative results

Figure 2: Examples of sequences with similar distractors and extreme complex background. It shows our tracker is robust to the similar objects even the distractors are closed to the tragets. Our tracker can also handle with the complex background such as dark, raining and blur.

Qualitative results

Figure 3: Examples of sequences with dim targets and appearance variations. As is shown in the sequence Bird, our tracker can predict accurate segmentation results even when target is extremly small. Moreover, our tracker achieves marvelous performance when fast-moving targets have extrme appreance deformation.

Qualitative results

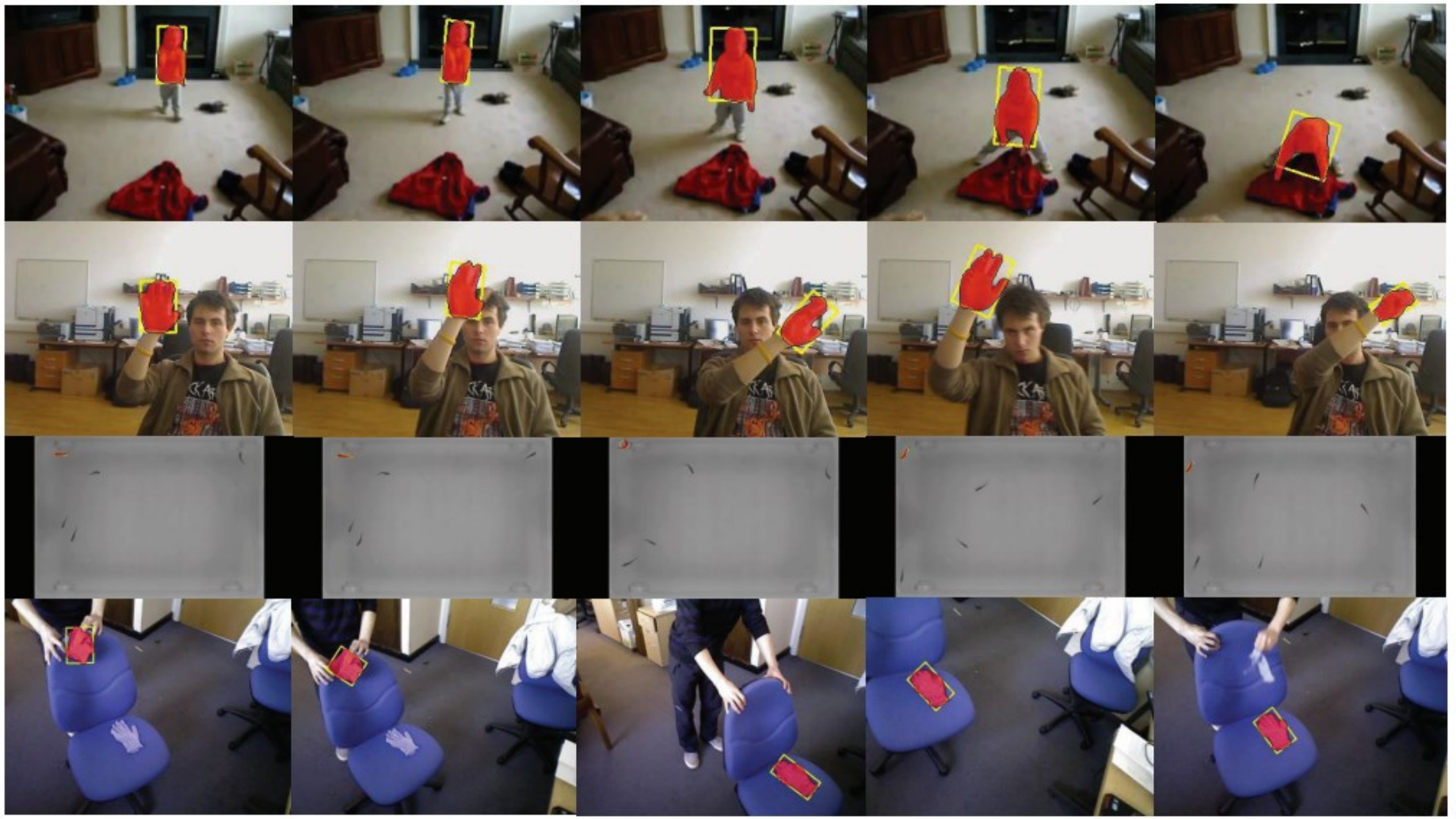


Figure 4: Examples of sequences with unseen targets. DNN-based trackers heavily rely on the pre-trained knowledge of targets. Our tracker utilizes the spatiotemporal information of every frame to enhance the generalizaion ability. From the examples of hand, blanket and glove, our tracker still tackles with those unseen targets well.

For more details, please refer to our paper.

Thanks!

https://github.com/phiphiphi31/DMB

Learning Spatio-Appearance Memory Network for High-Performance Visual Tracking

Fei Xie, Wankou Yang , Kaihua Zhang, Bo Liu, Guangting Wang, Wangmeng Zuo