SRNet: Spatial Relation Network for Efficient Single-stage Instance Segmentation in Videos

Xiaowen Ying Xin Li Mooi Choo Chuah

Lehigh University

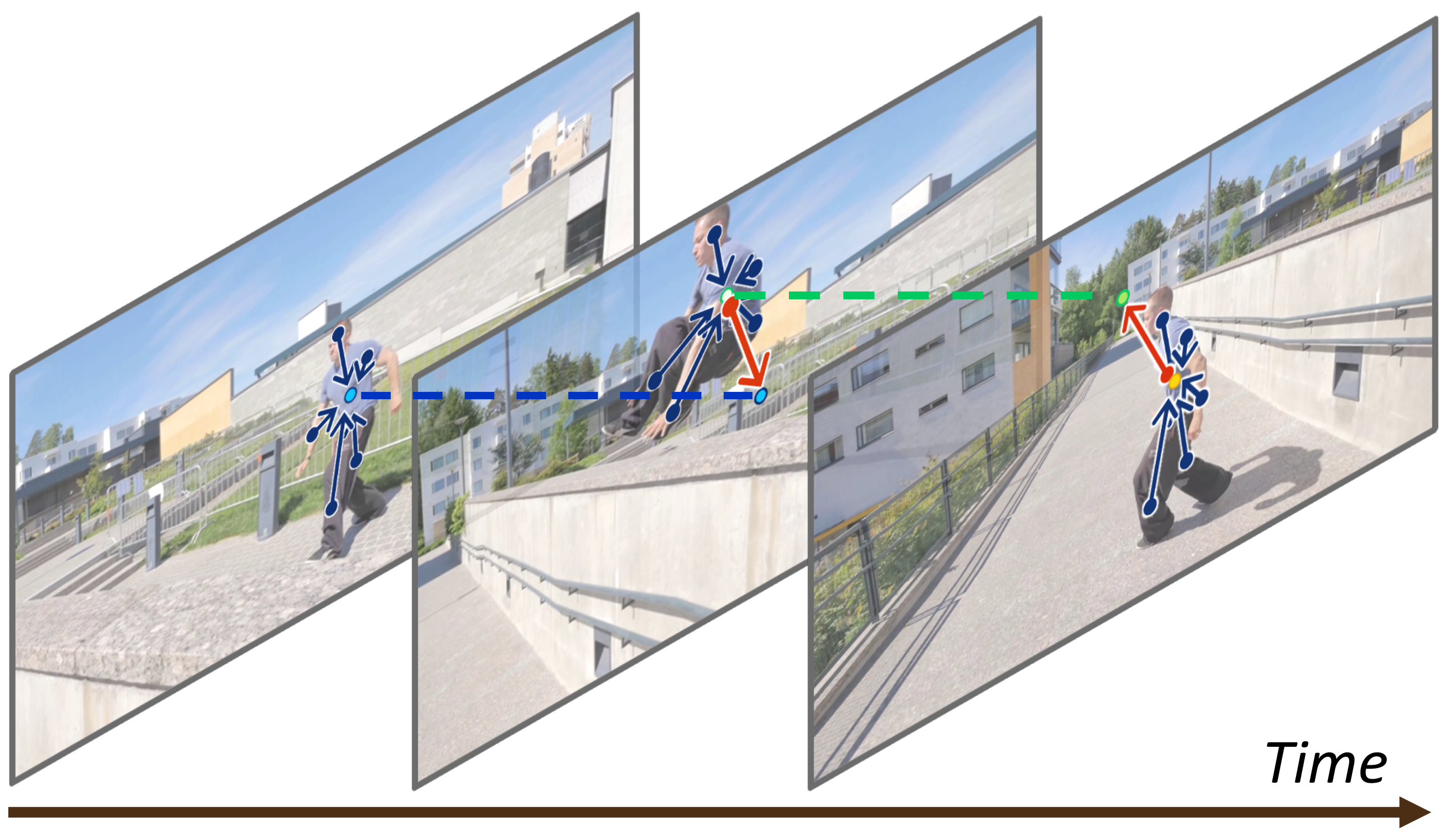
ACM multimedia

Chengdu, China OCT 20-24 2021

Paper ID: 2495

Overview

- We propose SRNet, a simple and efficient framework for joint segmentation and tracking of object instances in videos.
- We formulate the instance segmentation and tracking problem into a unified spatial-relation learning task where each pixel in the current frame relates to its object center, and each object center relates to its location in the previous frame.
- This unified learning framework allows our framework to perform join instance segmentation and tracking through a single stage while maintaining low overheads among different learning tasks.
- Our proposed framework can handle both UVOS and VIS tasks and demonstrates comparable performance with state-of-the-art methods on two different benchmarks while running significantly faster.

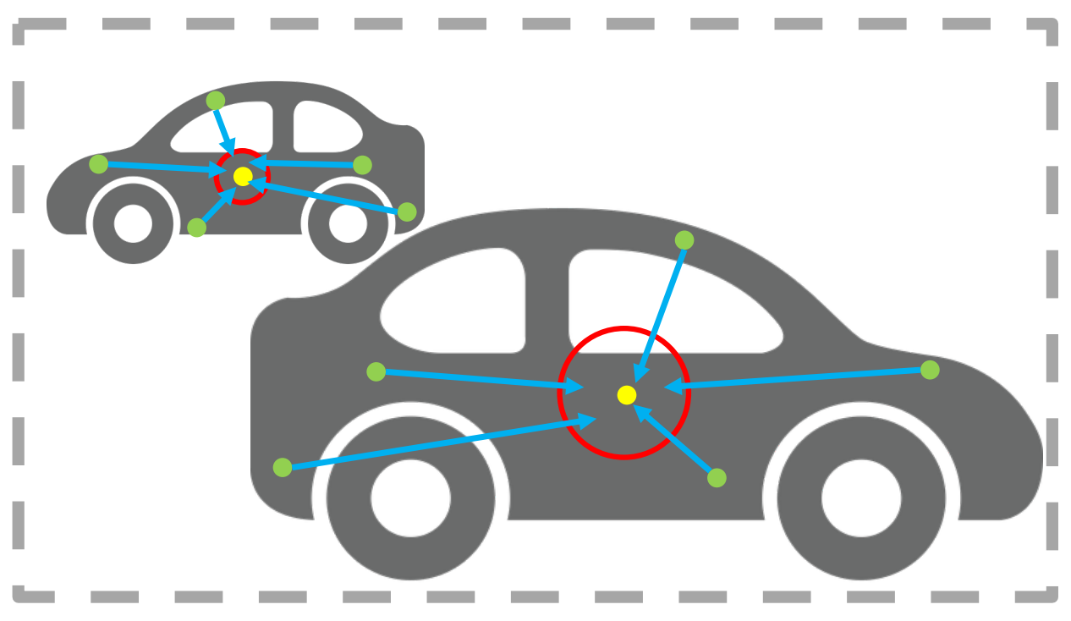


Blue Solid Arrows:
Pixels belonging to an instance point to their instance centers.

Red Solid Arrows:
Instance centers link to their previous locations.

Spatial Relation Learning

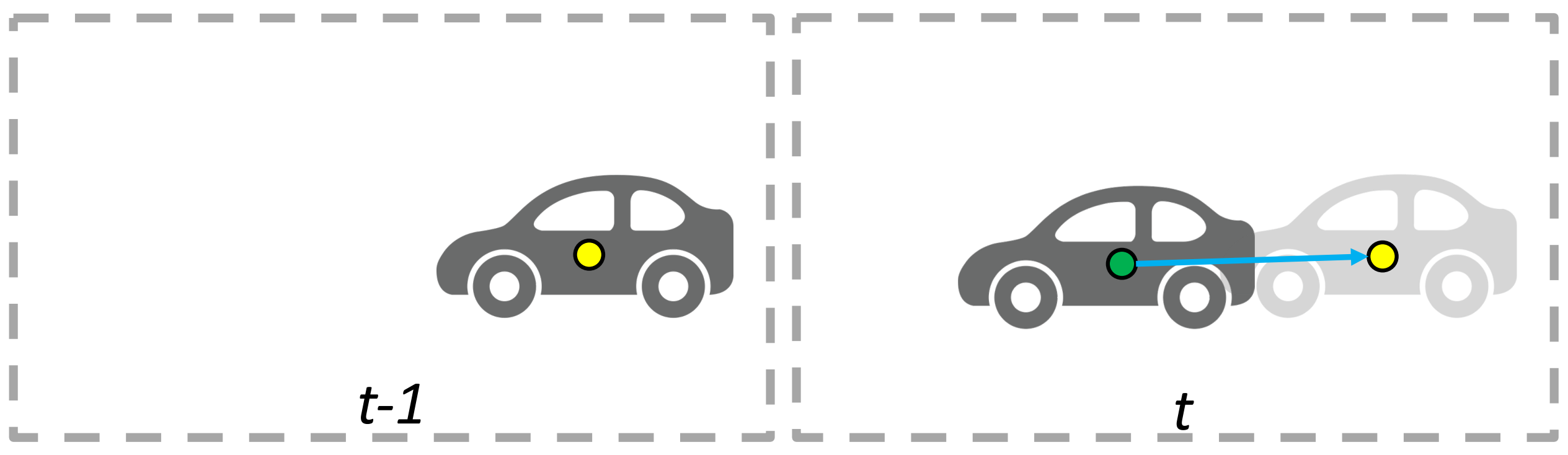
(1) Learning to generate instance mask



Learning Objective: $\|\mathcal{E}_{i,j} + \mathcal{S}_{i,j} - c_n\| \leq \Sigma_n$

(Color of the symbols matches the corresponding elements in the figure)

(2) Learning to associate instances across time

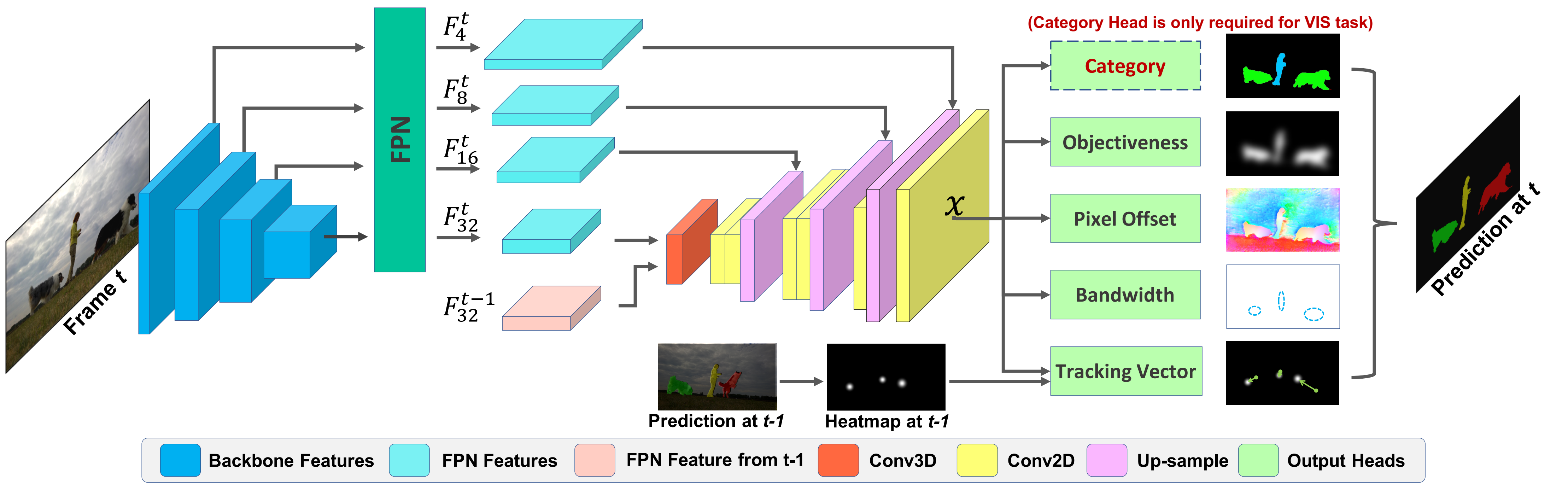


Learning Objective: $\mathcal{T}_{c_n^t} + \mathcal{S}_{c_n^t} = \mathcal{S}_{c_n^{t-1}}$

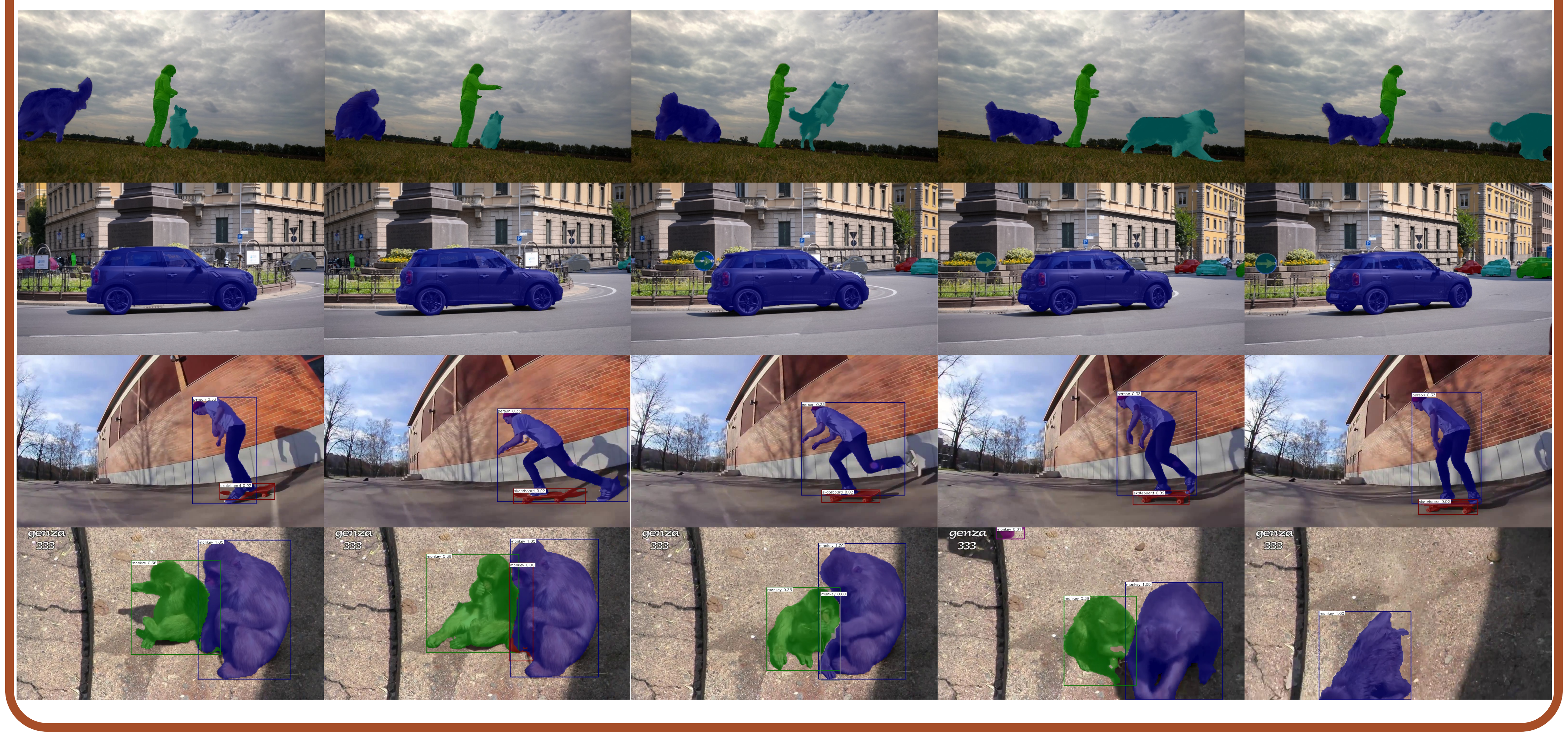
Acknowledgement

This work was partially supported by **National Science Foundation (NSF)** Grant CPS 1931867, and a gift from **Qualcomm Inc.**

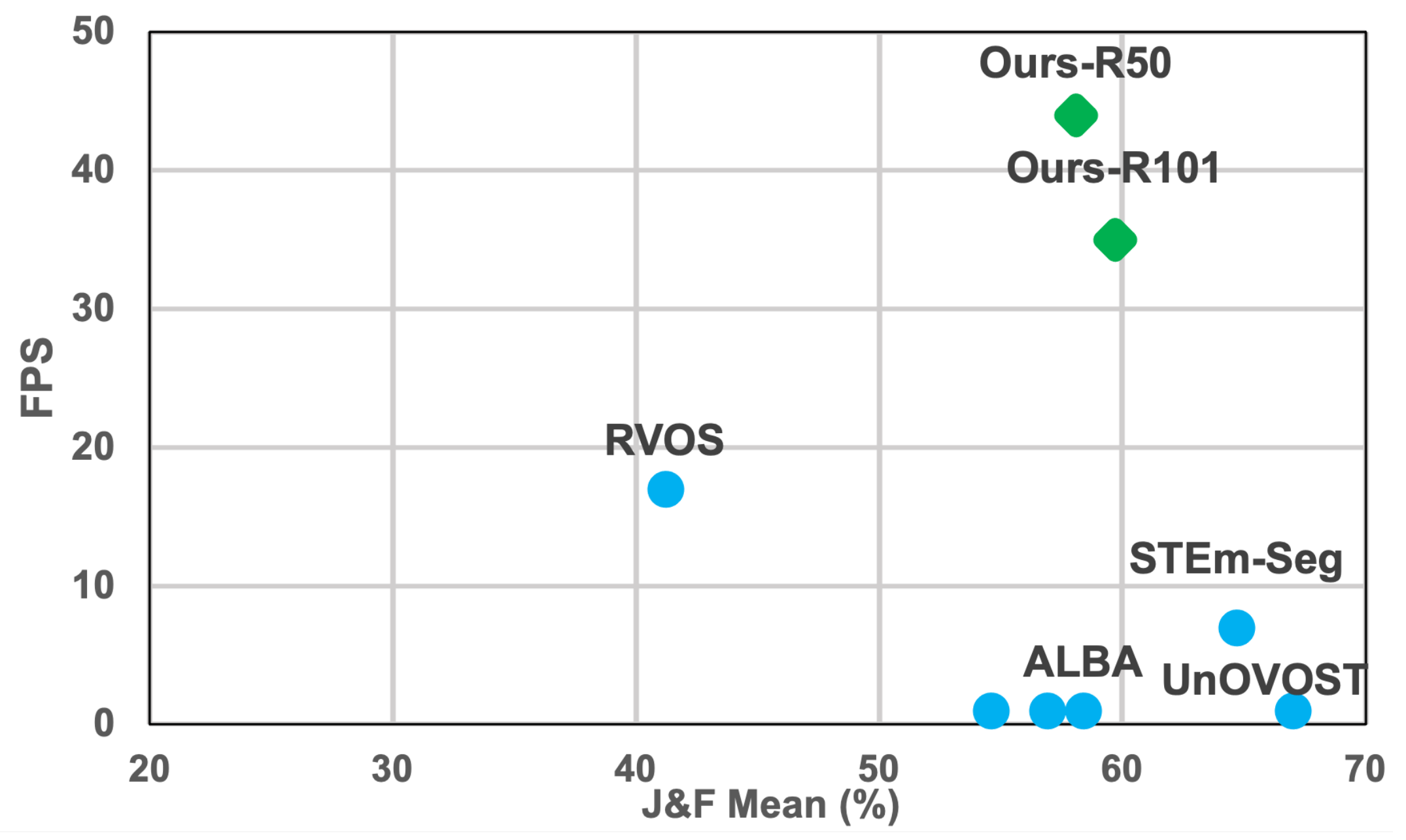
Architecture



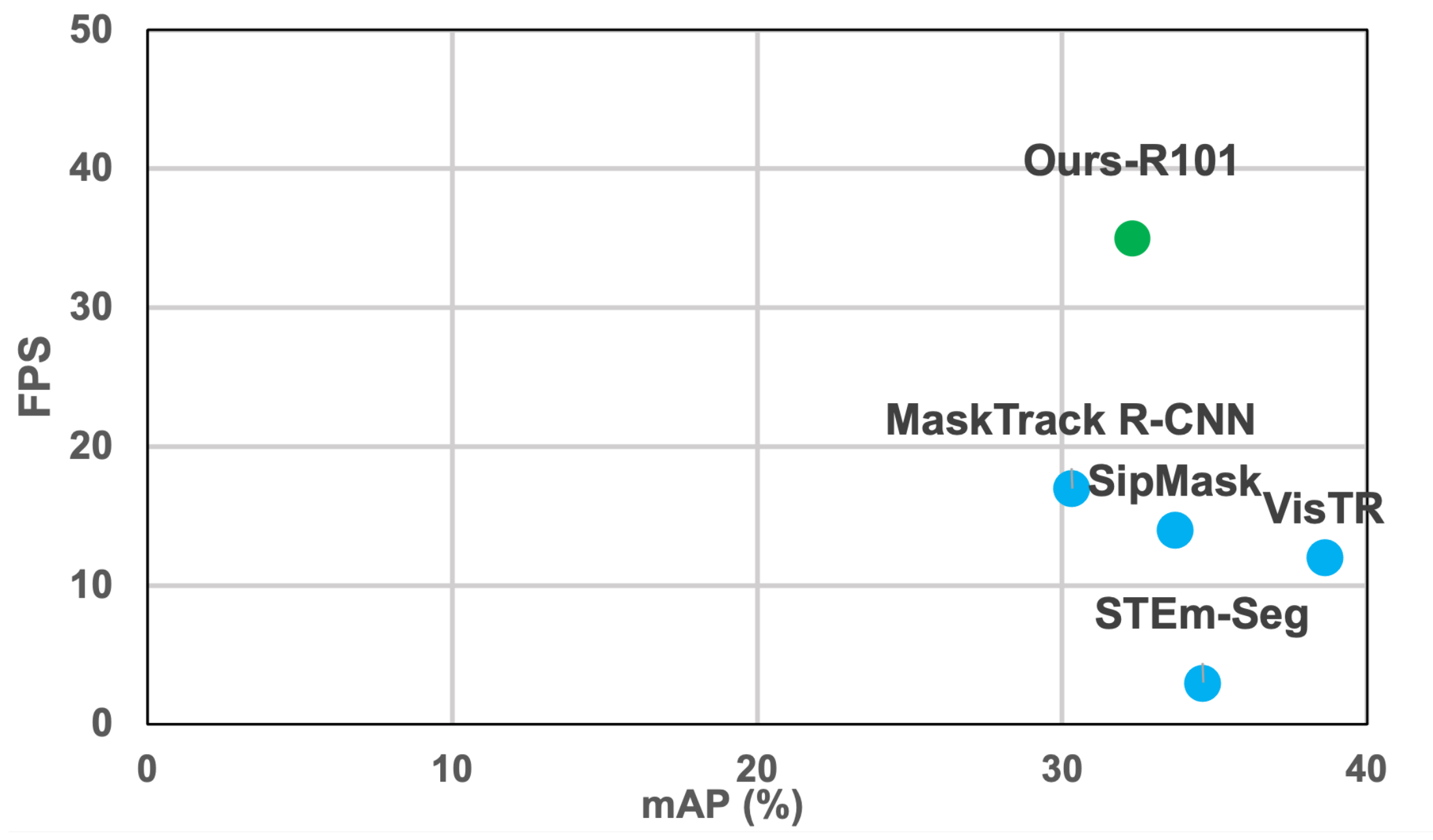
Qualitative Results



Quantitative Results



Speed-Accuracy Trade-off on DAVIS-2019



Speed-Accuracy Trade-off on Youtube-VIS

Methods	#Frames	Proposal	Flow	Re-ID	FPS	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$-Mean	$\mathcal{J}$-Recall	$\mathcal{J}$-Decay	$\mathcal{F}$-Mean	$\mathcal{F}$-Recall	$\mathcal{F}$-Decay
UnOVOST † [25]	1	✓	✓	✓	<1	67.0	65.6	75.5	0.3	68.4	75.9	3.7
STEm-Seg [1]	16				7	64.7	61.5	70.4	-4.0	67.8	75.5	1.2
OF-Tracker [1]	1	✓	✓		1	54.6	53.4	60.9	-1.3	55.9	63.0	1.1
RI-Tracker [1]	1	✓		✓	<1	56.9	55.5	63.3	2.7	58.2	64.4	6.4
ALBA [12]	1	✓	✓		<1	58.4	56.6	63.4	7.7	60.2	63.1	7.9
AGNN [42]	1	✓	✓		<1	61.1	58.9	65.7	11.7	63.2	67.1	14.3
RVOS [40]	1				17	41.2	36.8	40.2	0.5	45.7	46.4	1.7
Ours	1				35	59.7	58.2	66.6	-3.7	61.3	68.2	-0.9

Experimental Results on the Validation Set of **DAVIS-2019 UVOS** Track.

Methods	#Frames	Proposal	FPS	mAP	AP@50	AP@75	AR@1	AR@10
STEm-Seg [1]	16		3	34.6	55.8	37.9	34.4	41.6
VisTR [43]	36		12	38.6	61.3	42.3	37.6	44.2
IoUTracker+ [47]	1	✓	-	23.6	39.2	25.5	26.2	30.9
DeepSORT [45]	1	✓	-	26.1	42.9	26.1	27.8	31.3
OSMN [48]	1	✓	-	27.5	45.1	29.1	28.6	33.1
SeqTracker [47]	1		-	27.5	45.7	28.7	29.7	32.5
MaskTrack R-CNN [47]	1		17	30.3	51.1	32.6	31	35.5
SipMask [5]	1		14	33.7	54.1	35.8	35.4	40.1
Ours	1		35	32.3	50.2	34.8	32.3	40.1

Experimental Results on the Validation Set of **Youtube-VIS** Benchmark.