Samsung Research

MobileVOS: Real-Time Video Object Segmentation

Contrastive Learning meets Knowledge Distillation

Roy Miles, Mehmet Kerim Yucel, Bruno Manganelli, Albert Saa-Garriga

Samsung Research UK

May 22, 2023

Overview

1. Background
2. Method
3. Experiments
4. Conclusion

Background

Figure: Video object segmentation deals with tracking the location of objects over a video sequence. The goal is to generate accurate and consistent pixel-level object masks across the frames.

- This task serves as a foundation for many video understanding tasks such as action recognition, video editing, and augmented reality.

Background - Limitations

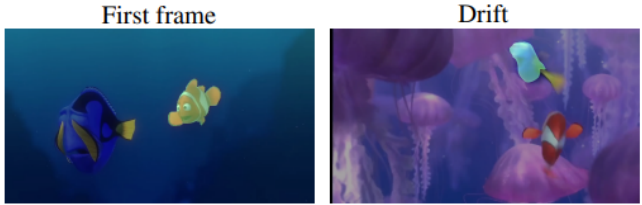


Figure: The STM memory model scales poorly for longer video sequences and introduces problems, such as drift, where the model can catastrophically degrade in performance over time.

- The current state-of-the-art relies on **space-time-memory networks** (STM), which relies on densely matching features from previous frames.

Method - Modified Architecture

Samsung Research

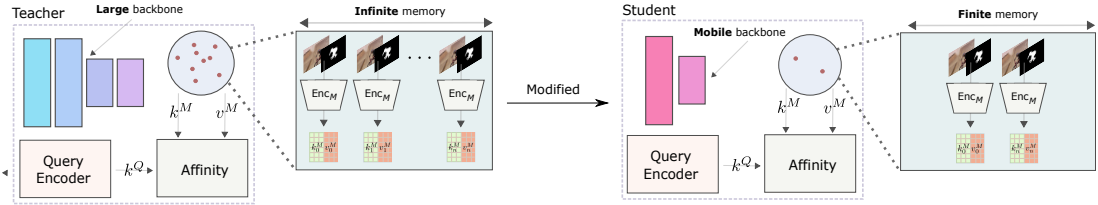


Figure: We restrict our attention to a finite STM network, whereby only the most recent and first frame features are stored in memory.

- To enable real-time performance, we use a much smaller MobileNet backbone.
- To encourage temporally consistent features, we **distill knowledge** a pre-trained infinite memory teacher.

Method - Knowledge Distillation

Samsung Research

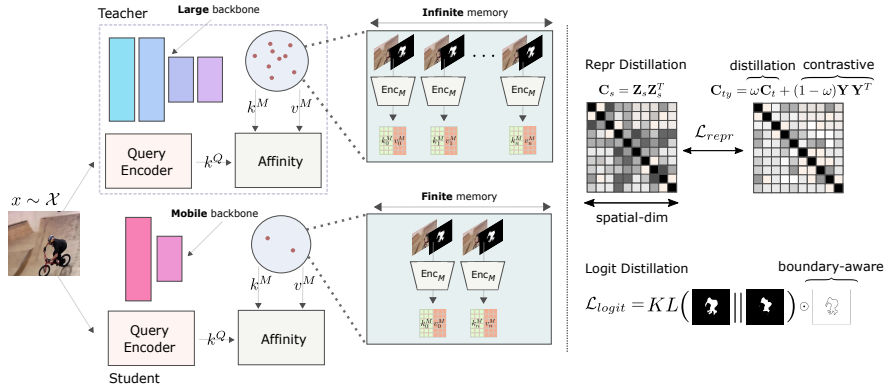


Figure: We jointly introduce **boundary aware sampling** to improve model convergence, and a natural **unification with supervised-contrastive learning** for varying student-teacher capacity gaps.

Method - Knowledge Distillation

The proposed distillation loss is given as follows.

$$\mathcal{L}_{repr} = \frac{1}{|\ \mathbf{C}_s\ |}\left(\log_2 \|\mathbf{C}_s\|^2 - \log_2 \|\mathbf{C}_s \odot \mathbf{C}_t\|^2\right) \quad (1)$$

where $\mathbf{C}_s, \mathbf{C}_t \in \mathbb{R}^{HW \times HW}$ capture the relationship between all pairs of pixels in the student and teacher feature space respectively.

Boundary-aware sampling

Sampling the boundary pixels not only improves model convergence and addresses observed limitations of SVOS models, but also significantly reduces the memory constraints in constructing these matrices.

Method - Knowledge Distillation

By introducing known relationships between the pixel-wise features, we can provide a natural scheme to interpolate between knowledge distillation and supervised constrastive learning.

$$\mathbf{C}_{ty} = \omega \mathbf{C}_t + (1 - \omega)\mathbf{Y}\mathbf{Y}^T \tag{2}$$

In the case where $\omega = 0$, we arrive at a familiar supervised constrastive setting.

$$\mathcal{L}_{repr} \rightarrow \mathcal{L}_{SupCon} = -\frac{1}{\mid \mathbf{C}_s \mid} \log_2 \sum_i \frac{\sum_{j \in \mathcal{P}_i} sim(\mathbf{Z}_i, \mathbf{Z}_j)}{\sum_k sim(\mathbf{Z}_i, \mathbf{Z}_k)} \tag{3}$$

Experiments - DAVIS16

Method	CC	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	FPS
STM†	✗	89.3	88.7	89.9	6.3
MiVOS†*	✗	91.0	89.7	92.4	16.9
STCN†*	✗	91.7	90.4	93.0	26.9
BATMAN	✗	**92.5**	**90.7**	**94.2**	-
XMem†*	✗	92.0	**90.7**	93.2	**29.6**
SwiftNet†	✓	90.4	**90.5**	90.3	25.0
RDE-VOS†*	✓	**91.6**	90.0	**93.2**	35.0
MobileVOS					
ResNet18†*	✓	91.4	90.3	92.6	**100.1**
MobileNetV2†	✓	90.5	89.5	91.5	81.8
↳ wo/ ASPP†	✓	90.1	89.0	91.1	86.0

Experiments - DAVIS17

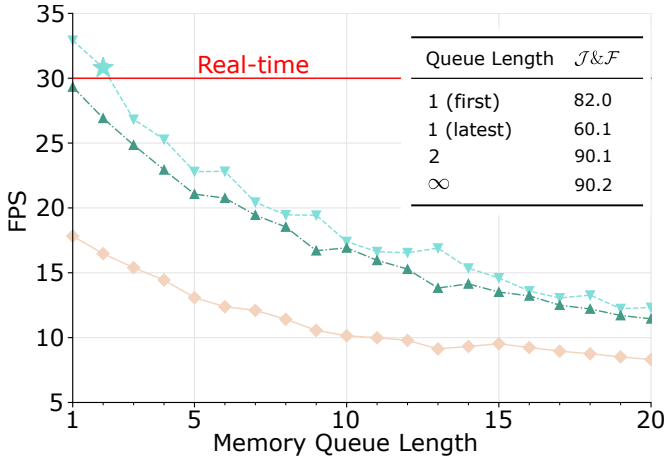
Method	CC	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	FPS
STM†	✗	81.8	79.2	84.3	10.2
STCN†*	✗	85.3	82.0	88.6	20.2
BATMAN	✗	86.2	83.2	89.4	-
XMem†*	✗	**87.7**	**84.0**	**91.4**	**22.6**
SwiftNet†	✓	81.1	78.3	83.9	<25.0
RDE-VOS†*	✓	**86.1**	82.1	**90.0**	27.0
MobileVOS					
ResNet18†*	✓	85.0	81.7	88.3	**90.6**
MobileNetV2†	✓	82.2	78.7	85.7	79.1
↳ wo/ ASPP†	✓	81.8	78.3	85.3	81.3

Experiments - Mobile Performance

Method	Params(M)	FPS *NVIDIA A40*		FPS *NVIDIA 1080Ti*	
		short	long (10×)	short	long (10×)
STM	38.9	8.9	4.3	6.8	✗
GSFM	67.0	18.4	4.2	7.6	✗
STCN	54.4	37.4	8.3	18.1	✗
RDE-VOS	64.0	32.0	34.2	14.4	14.1
XMem	62.2	38.6	39.9	12.6	12.7
MobileVOS					
ResNet18	8.1	**144.7**	**145.4**	**76.0**	**76.3**
MobileNetV2	2.5	99.9	99.1	61.6	60.6
↳ wo/ ASPP	**1.9**	105.1	103.4	66.8	67.4

Experiments - Mobile Performance

Samsung Research



Queue Length	$\mathcal{J}\&\mathcal{F}$
1 (first)	82.0
1 (latest)	60.1
2	90.1
∞	90.2

Conclusion

- Shown that finite memory STM networks are an efficient class of models for mobile device inference.
- Provide a natural unification of supervised contrastive learning and KD.
- Introduce boundary-aware sampling as a task specific trick for improving model convergence and memory constraints for SVOS distillation.