Recurrent Dynamic Embedding for Video Object Segmentation

CVPR 2022

提纲

Authors

- Authors

Recurrent Dynamic Embedding for Video Object Segmentation

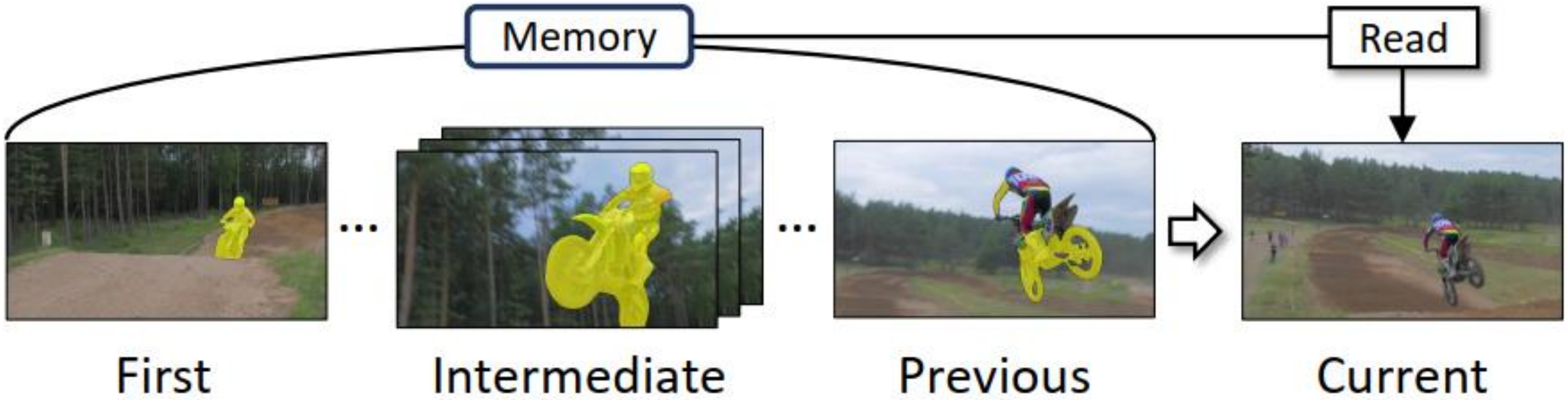
Mingxing Li[1*], Li Hu[2*], Zhiwei Xiong[1†], Bang Zhang[2], Pan Pan[2], Dong Liu[1]
[1]University of Science and Technology of China
[2]Alibaba DAMO Academy, Alibaba Group
mxli@mail.ustc.edu.cn {zwxiong, dongeliu}@ustc.edu.cn
{hooks.hl, zhangbang.zb, panpan.pp}@alibaba-inc.com

提纲

- About Article
- Introduction
- Methodology
- Experiments and Result
- Conclusion
- Inspiration

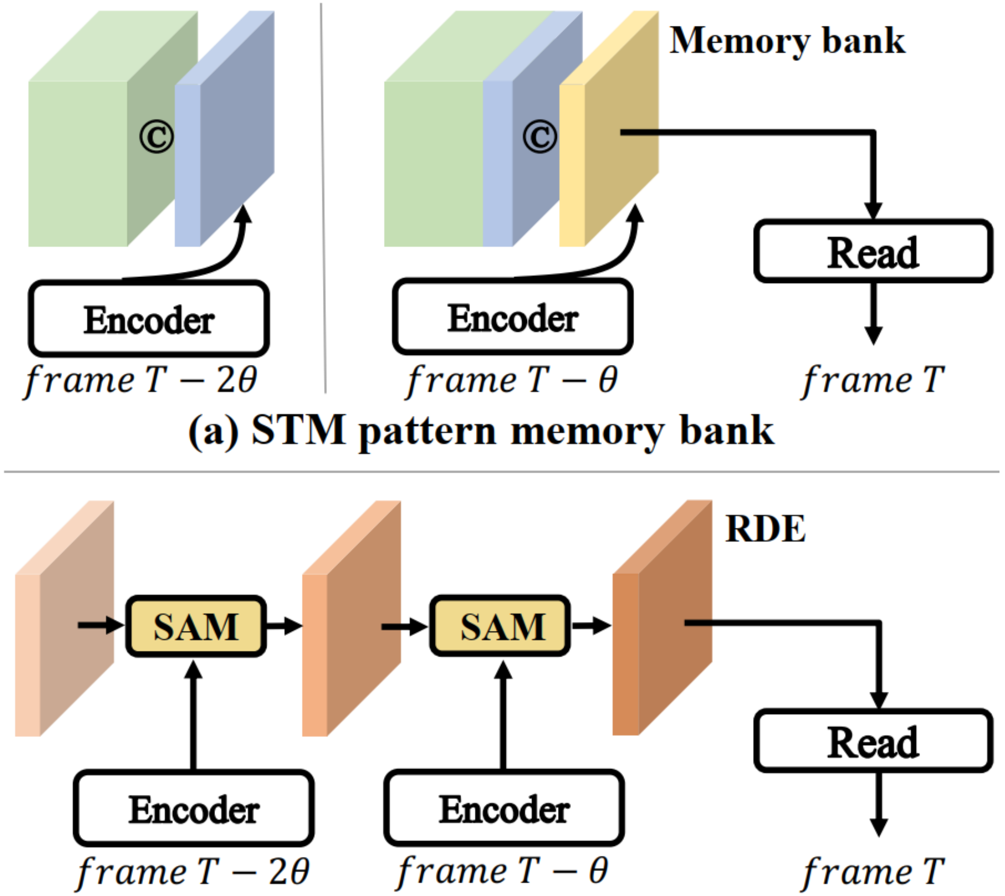
Introduction

- Space-Time Memory(STM)



Introduction

- Memory Bank



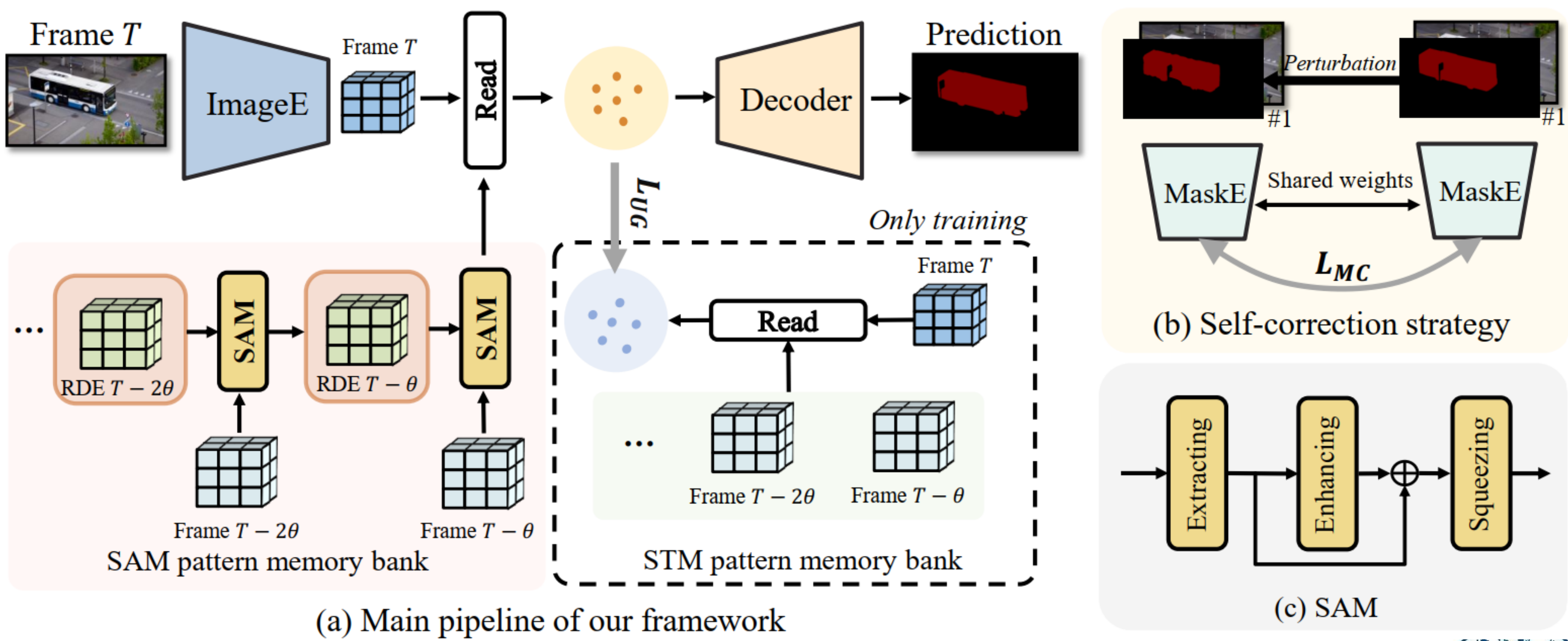
(a) STM pattern memory bank

(b) Our SAM pattern memory bank

提纲

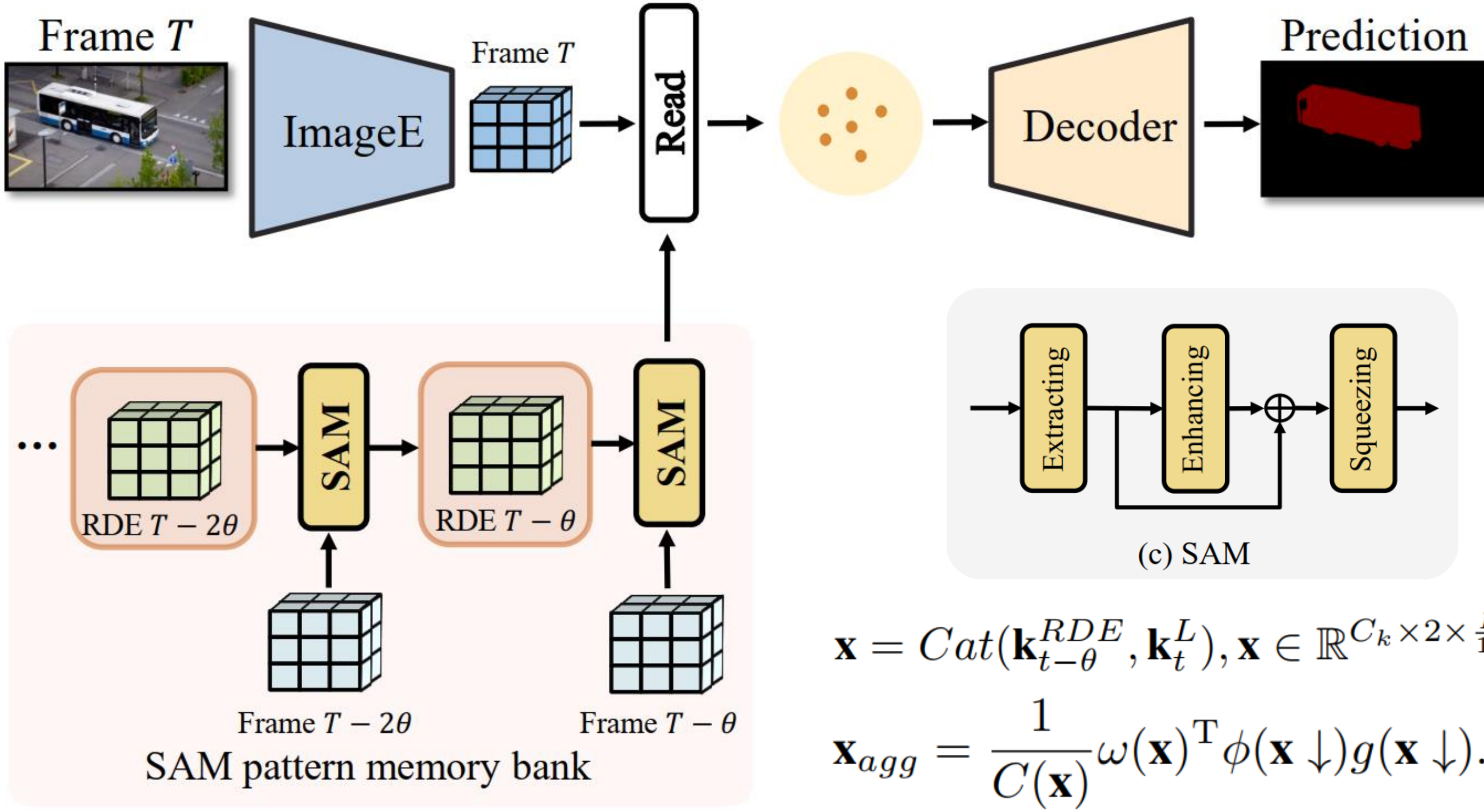
Methodology

Framework



Methodology

- Recurrent Dynamic Embedding



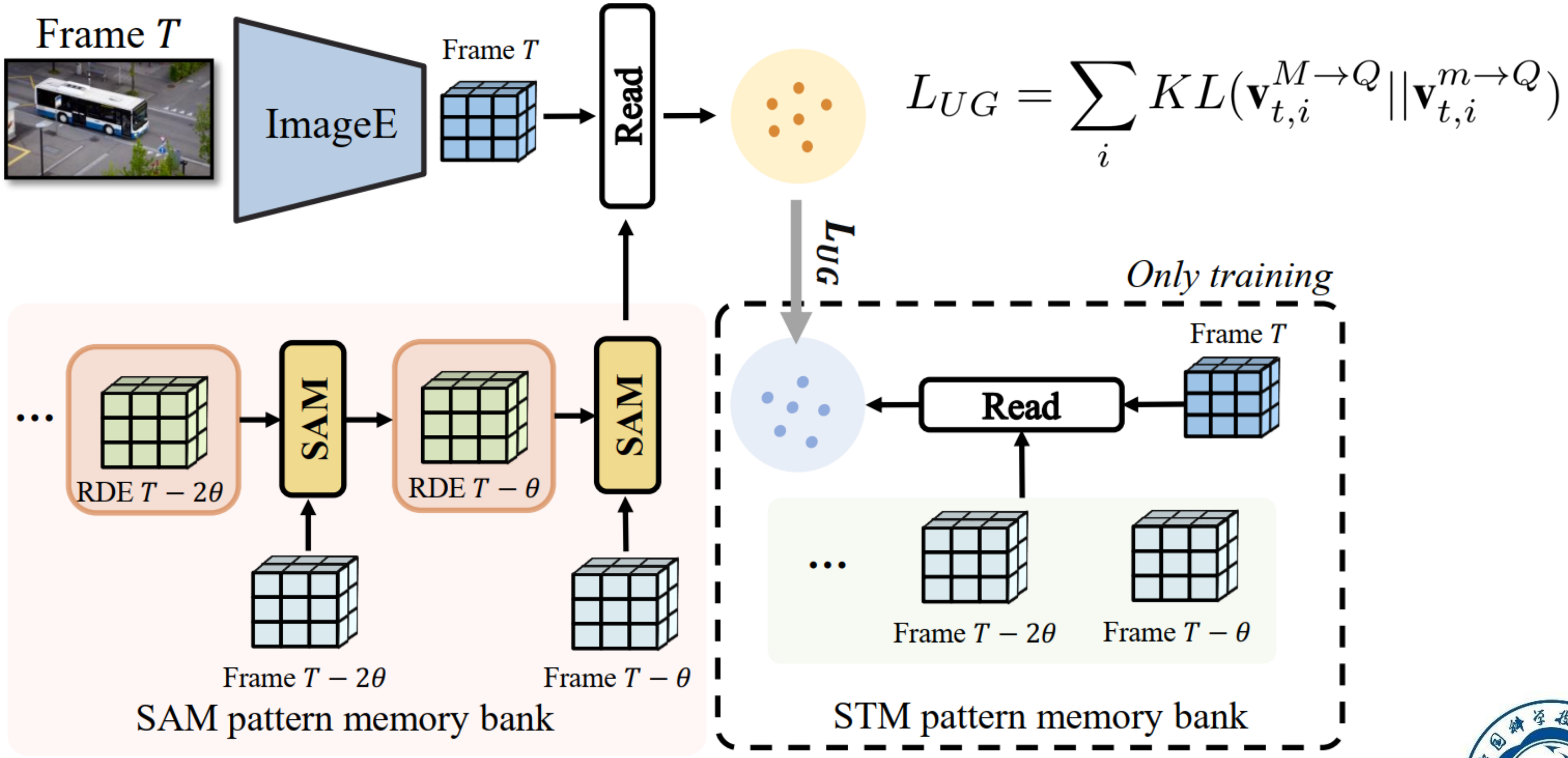
$$\mathbf{x} = Cat(\mathbf{k}_{t-\theta}^{RDE}, \mathbf{k}_t^L), \mathbf{x} \in \mathbb{R}^{C_k \times 2 \times \frac{H}{16} \times \frac{W}{16}}$$

$$\mathbf{x}_{agg} = \frac{1}{C(\mathbf{x})} \omega(\mathbf{x})^{\mathrm{T}} \phi(\mathbf{x}\downarrow) g(\mathbf{x}\downarrow).$$

$$\mathbf{k}_t^{RDE} = Squeeze(\mathbf{x}_{agg} + ASPP(\mathbf{x}_{agg})).$$

Methodology

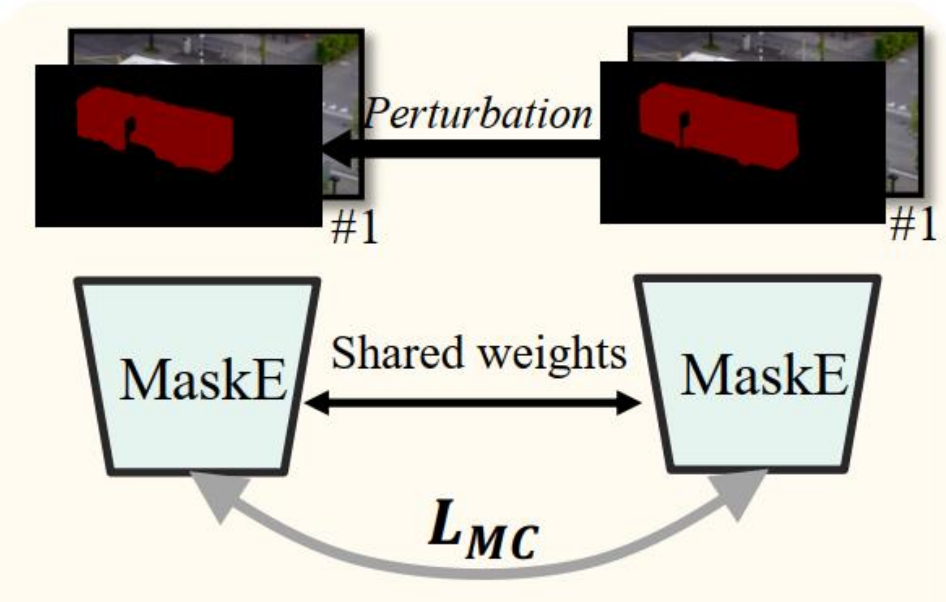
- Unbiased Guidance Loss



(a) Main pipeline of our framework

Methodology

- Self-correction Strategy



(b) Self-correction strategy

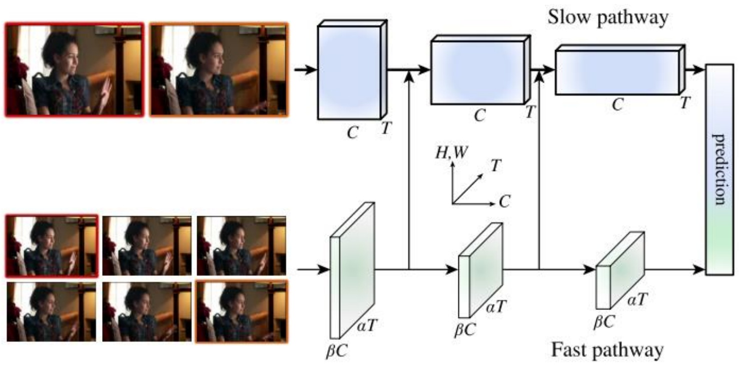
$$L_{MC} = KL(\mathbf{k}_1 || \ddot{\mathbf{k}}_1) + \sum_i KL(\mathbf{v}_{1,i} || \ddot{\mathbf{v}}_{1,i})$$

Methodology

□ Training Strategy

$$L_{Seg} = \frac{1}{2}\Big(\sum_{i}\sum_{t=2,4}\underbrace{BCE(\tilde{\mathbf{y}}^{M}_{t,i}, \mathbf{y}_{t,i})}_{STM\ pattern\ item} + \sum_{i}\sum_{t=3,5}\underbrace{BCE(\tilde{\mathbf{y}}^{m}_{t,i}, \mathbf{y}_{t,i})}_{SAM\ pattern\ item}\Big)$$

$$Loss = L_{Seg} + \mathbb{1}[t=3,5]\mu L_{UG} + \gamma L_{MC}$$



提纲

Experiments and Result

Experiments

Method	CC	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	FPS
RMNet† [37]	×	88.8	88.9	88.7	11.9
STM† [25]	×	89.3	88.7	89.9	6.3
KMN† [30]	×	90.5	89.5	91.5	8.4
LCM† [13]	×	90.7	89.9	91.4	8.5
HMMN† [31]	×	90.8	89.6	92.0	10.0
MiVOS†* [5]	×	91.0	89.7	92.4	16.9
STCN†* [6]	×	**91.7**	**90.4**	**93.0**	**26.9**
GCNet [17]	√	86.6	87.6	85.7	25.0
CFBI+† [41]	√	89.9	88.7	91.1	5.9
SwiftNet† [33]	√	90.4	**90.5**	90.3	25.0
RDE-VOS†	√	91.1	89.7	92.5	**35.0**
RDE-VOS†*	√	**91.6**	90.0	**93.2**	**35.0**

Table 3. Results on the DAVIS 2016 validation set. CC denotes constant cost during the inference.

Method	CC	Overall	$\mathcal{J}_{seen}$	$\mathcal{F}_{seen}$	$\mathcal{J}_{unseen}$	$\mathcal{F}_{unseen}$
STM† [25]	×	79.2	79.6	83.6	73.0	80.6
MiVOS†* [5]	×	82.4	80.6	84.7	78.2	85.9
STCN†* [6]	×	**84.2**	**82.6**	**87.0**	**79.4**	**87.7**
CFBI† [40]	√	81.0	80.6	85.1	75.2	83.0
SST† [8]	√	81.8	80.9	-	76.6	-
RDE-VOS †	√	81.9	81.1	85.5	76.2	84.8
RDE-VOS †*	√	**83.3**	**81.9**	**86.3**	**78.0**	**86.9**

Table 4. Results on the YouTube-VOS 2019 validation set.

Method	CC	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	FPS
STM† [25]	×	81.8	79.2	84.3	10.2
KMN† [30]	×	82.8	80.0	85.6	<8.4
JOINT† [23]	×	83.5	80.8	86.2	4.0
LCM† [13]	×	83.5	80.5	86.5	<8.5
RMNet† [37]	×	83.5	81.0	86.0	<11.9
MiVOS†* [5]	×	84.5	81.7	87.4	11.2
HMMN† [31]	×	84.7	81.9	87.5	<10.0
STCN†* [6]	×	**85.3**	**82.0**	**88.6**	**20.2**
GCNet [17]	√	71.4	69.3	73.5	<25.0
Liang *et al.* [19]	√	74.6	73.0	76.1	4.0
G-FRTM† [26]	√	76.4	-	-	18.2
PReMVOS [21]	√	77.8	73.9	81.7	0.01
SwiftNet† [33]	√	81.1	78.3	83.9	<25.0
SST† [8]	√	82.5	79.9	85.1	-
Ge *et al.*† [10]	√	82.7	80.2	85.3	6.7
RDE-VOS†	√	84.2	80.8	87.5	**27.0**
RDE-VOS†*	√	**86.1**	**82.1**	**90.0**	**27.0**

Method	CC	600p	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
STM† [25]	×	√	72.2	69.3	75.2
KMN† [30]	×	√	77.2	74.1	80.3
RMNet† [37]	×	×	75.0	71.9	78.1
Ge *et al.*† [10]	×	×	75.2	72.0	78.3
STCN†* [6]	×	×	77.8	74.3	81.3
MiVOS†* [5]	×	×	**78.6**	**74.9**	**82.2**
CFBI† [40]	√	×	74.8	71.1	78.5
Ge *et al.*† [10]	√	×	75.2	72.0	78.3
CFBI+† [41]	√	×	75.6	71.6	79.6
RDE-VOS†	√	×	77.4	73.6	81.2
RDE-VOS†*	√	×	**78.9**	**74.9**	**82.9**

Experiments and Result

- Experiments

Variants	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
Strategy permutation			
RDE	81.8	78.0	85.7
First frame	71.6	67.8	75.4
First frame & RDE	85.3	81.6	89.0
Latest frame	80.4	76.9	83.8
Latest frame & RDE	82.2	78.4	86.0
First frame & latest frame	84.6	81.0	88.2
F & L & RDE	85.4	81.6	89.2
First frame $\times 2$ & latest frame	85.1	81.5	88.7
First frame & latest frame $\times 2$	84.0	80.4	87.6
2F & L & RDE	**86.1**	**82.1**	**90.0**
Sampling interval θ			
2F & L & RDE ($\theta = 2$)	85.1	81.4	88.9
2F & L & RDE ($\theta = 3$)	**86.1**	**82.1**	**90.0**
2F & L & RDE ($\theta = 4$)	85.1	81.5	88.8
2F & L & RDE ($\theta = 5$)	84.2	80.5	87.9

Experiments and Result

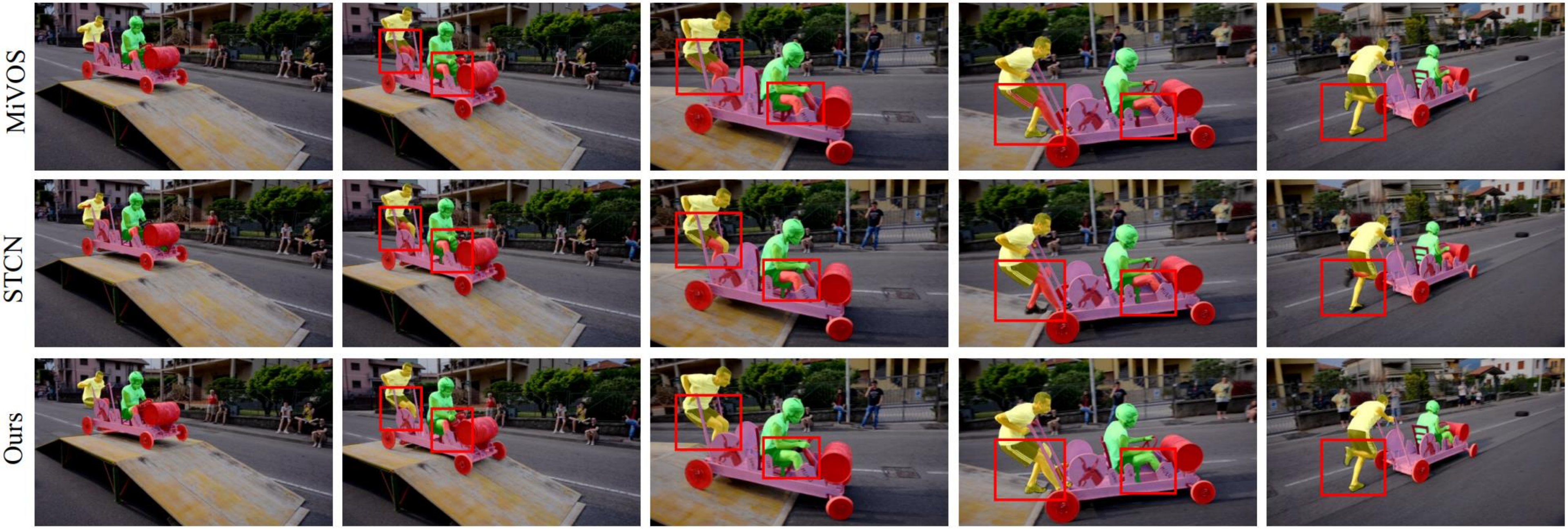
- Experiments

	Ablation Settings	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
Loss	w/o L_{MC}	83.7	80.5	86.9
	w/o L_{UG}	82.9	79.5	86.4
	w/o L_{MC} & L_{UG}	82.5	79.1	86.0
	L_{Seg} w/o STM pattern item	83.0	79.4	86.6
	Full	**84.2**	**80.8**	**87.5**

Table 6. Ablation of different loss functions without the BL30K [5] pre-training.

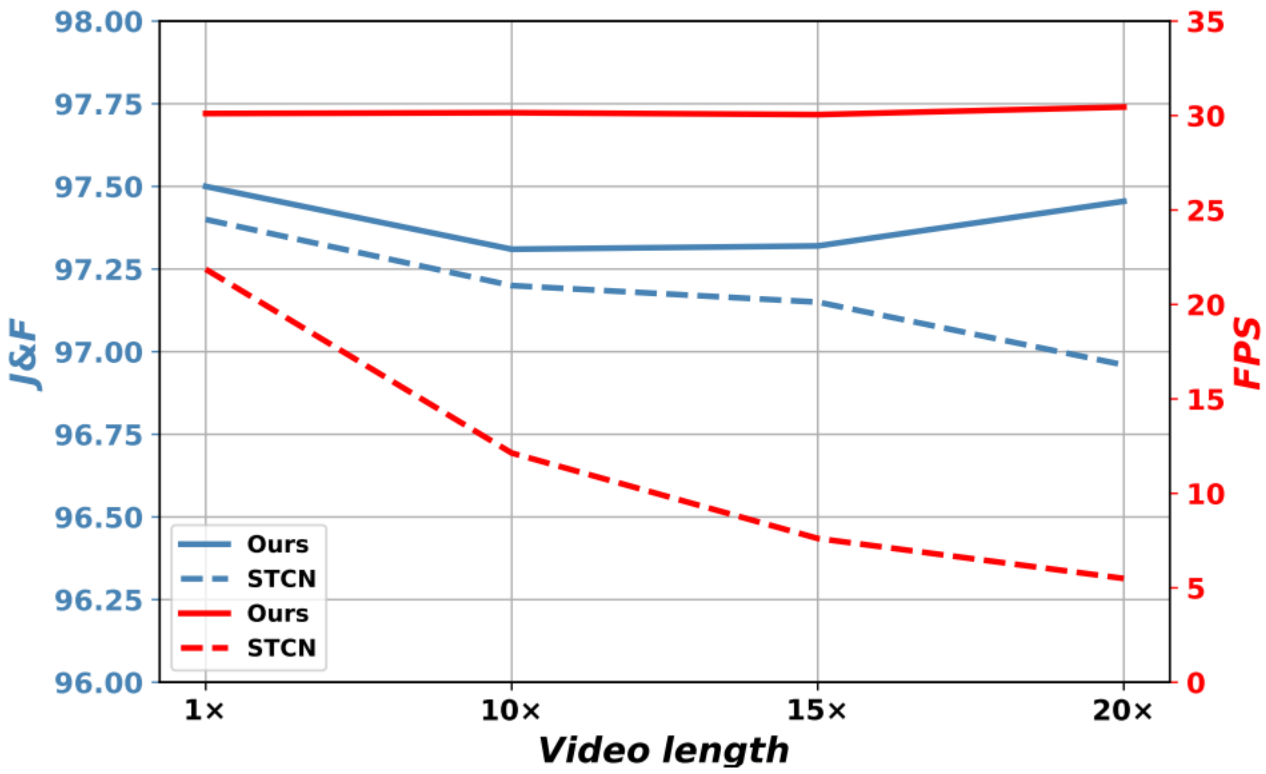
Experiments and Result

- Experiments



Experiments and Result

- Experiments



提纲

Conclusion

- Conclusion:

- Constant memory cost

- Lack of Intuitive Interpretation for memory update