READMem: Robust Embedding Association for a Diverse Memory in Unconstrained Video Object Segmentation

Stéphane Vujasinović[1], Sebastian Bullinger[1], Stefan Becker[1], Norbert Scherer-Negenborn[1], Michael Arens[1] and Rainer Stiefelhagen[2]

[1]Fraunhofer IOSB* (Ettlingen, Germany)
[2]Karlsruhe Institute of Technology (KIT) (Karlsruhe, Germany)

Fraunhofer IOSB — KIT Karlsruher Institut für Technologie — BMVC 2023

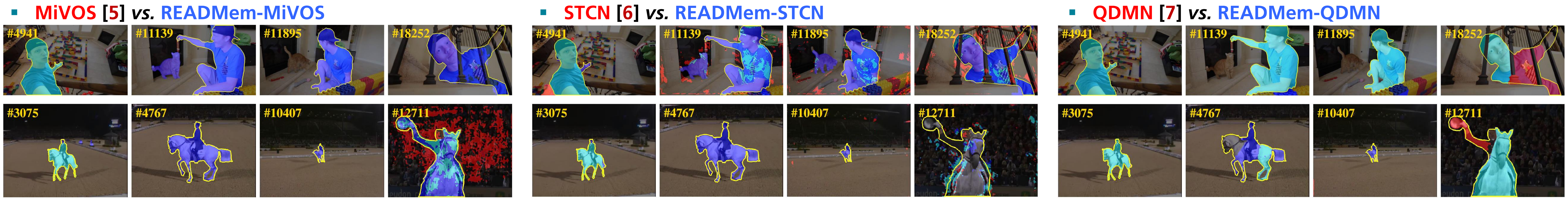
Scan me

Qualitative Results

Ground-Truth | Baseline | Baseline + READMem | Intersection

(Predicted results on LV1 [3])

- MiVOS [5] *vs.* READMem-MiVOS
- STCN [6] *vs.* READMem-STCN
- QDMN [7] *vs.* READMem-QDMN

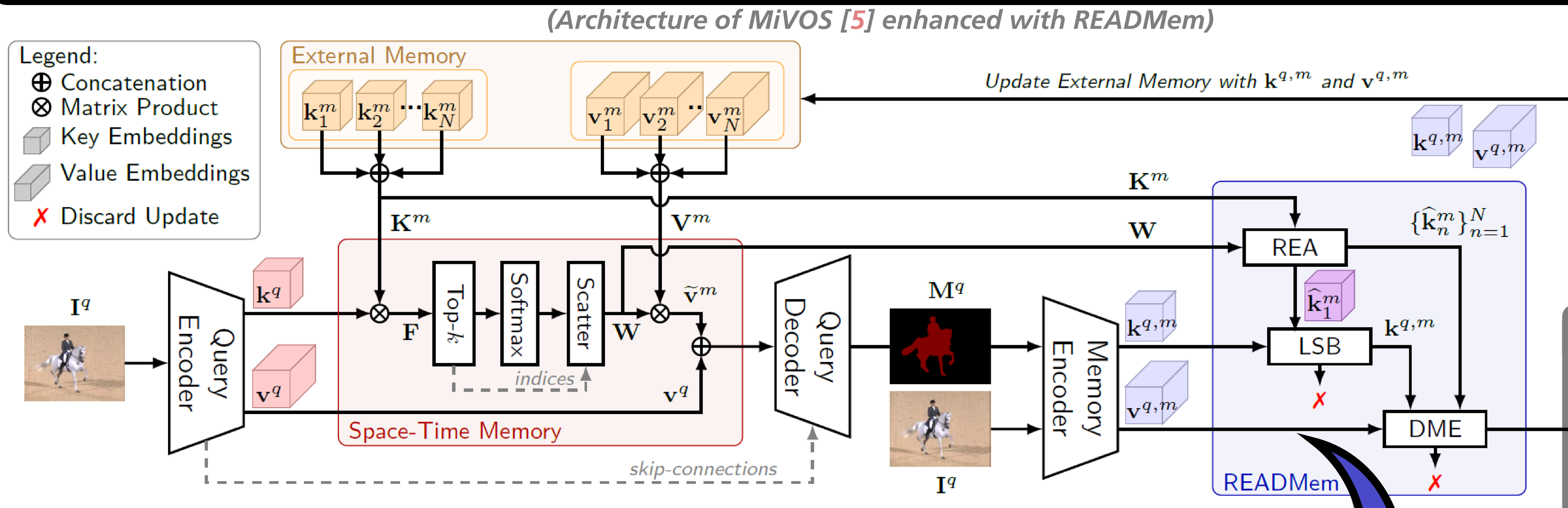


TL;DR

- *Task*: Semi-automatic Video Object Segmentation (sVOS).
- *Objective*: Handle unconstrained videos (arbitrary frame rate, length, object motion and camera motion).
- *Approach*: Increase the inter-frame diversity within the memory.
- *Implementation*: An extension to manage the memory of STM-like networks [1] available at `https://github.com/Vujas-Eteph/READMem`

Overview

(Architecture of MiVOS [5] enhanced with READMem)



- Rely on the Space-Time Memory operation (STM).

$$\widetilde{\mathbf{v}}^m = \mathbf{V}^m \sigma_{col}\left((\mathbf{K}^m)^\top \mathbf{k}^q\right)$$

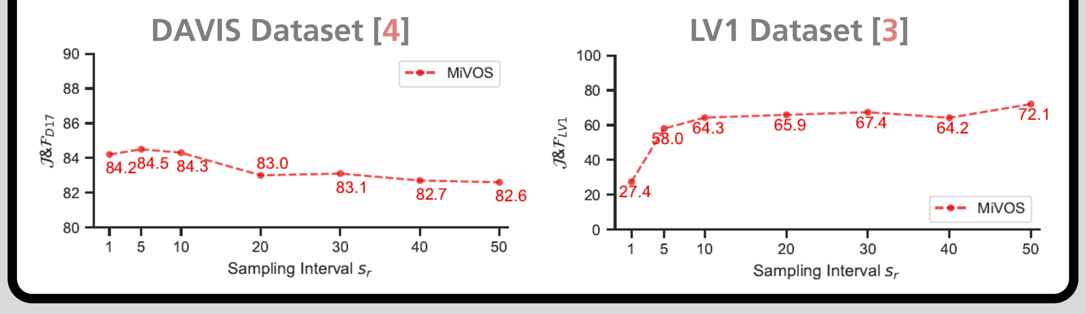
- READMem manages the external memory.

Legend

- $\mathbf{I}^q$: *Query image*
- $\mathbf{M}^q$: *Predicted mask*
- $\{\mathbf{k}^q, \mathbf{v}^q\}$: *Query key-value embeddings*
- $\{\mathbf{k}_n^m, \mathbf{v}_n^m\}$: *Memory key-value embeddings*
- n : *Index of the memory slot*
- N : *Total number of memory slots*
- $\{\mathbf{K}^m, \mathbf{V}^m\}$: *Concatenated embeddings*
- $\mathbf{F}$: *Similarity Matrix*
- $\mathbf{W}$: *Affinity Matrix*
- σ_{col} : *Softmax along the column*
- $\{\mathbf{k}^{q,m}, \mathbf{v}^{q,m}\}$: *Memory key-value embeddings of the query image*
- $\widehat{\mathbf{k}}_n^m$: *Projected key embeddings*
- $\widetilde{\mathbf{v}}^m$: *Aggregated memory values*
- $\mathbf{T}_n$: *Transformation Matrix*

Problem & Motivation

- Existing sVOS methods are tailored for short-video object segmentation:
  - Rely on an ever-expanding memory.
  - Define a dataset-specific sampling interval s_r, non-generalizable to unseen data.



Contributions

- *READMem* extends any sVOS method to deal with unconstrained sequences.
- Seamless integration without re-training or fine-tuning.
- *Generalizable* sampling interval s_r (no need to finetune on the validation set).
- Automatic memory embeddings diversity estimation via *Diversification of Memory Embeddings* (*DME*).
- Translation and scale invariant embedding association through *Robust Embedding Association* (*REA*).

READMem

Diversification of Memory Embeddings (DME):

- *Motivation*: Increase the diversity of the memory $\mathcal{M}^k = \{\mathbf{k}_n^m\}_{n=1}^N$, where $\mathbf{k}_n^m \in \mathbb{R}^{C_k \times HW}$.
- *Concept*: Represent the diversity of the memory $\mathcal{M}^k$ with the determinant of the associated Gram Matrix by:

$\mathbf{G} \in \mathbb{R}^{N \times N}$, where $\mathbf{X} \in \mathbb{R}^{C_k HW \times N}$ is the flatten representation of $\mathcal{M}^k = \{\mathbf{k}_n^m\}_{n=1}^N$

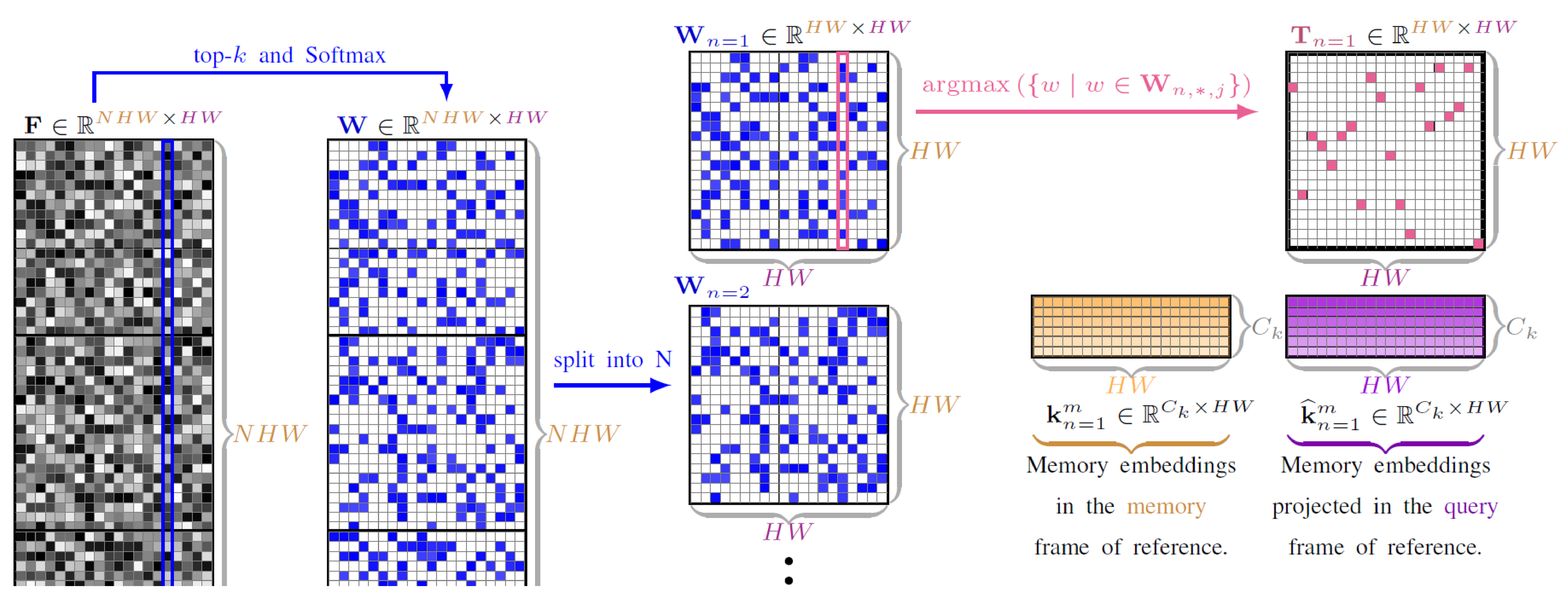
$$(volP(\{\mathbf{k}_n^m\}_{n=1}^N))^2 = \det(\mathbf{G}), \text{ where } \mathbf{G} = \mathbf{X}^\top\mathbf{X}$$

- *Process*: Memory keys updated if diversity increases.

Robust Embedding Association (REA):

- *Motivation*: Dampen translation and scale variations when computing the similarity.
- *Concept*: Projects the memory key embeddings $\mathbf{k}_n^m$ to the query's ($\mathbf{k}^{q,m}$) temporal frame of reference by:

$$\widehat{\mathbf{k}}_n^m = \mathbf{k}_n^m \mathbf{T}_n$$



- *Process*: Compute the similarity between $\widehat{\mathbf{k}}_n^m$ and $\mathbf{k}^{q,m}$ instead of $\mathbf{k}_n^m$.

Quantitative Results

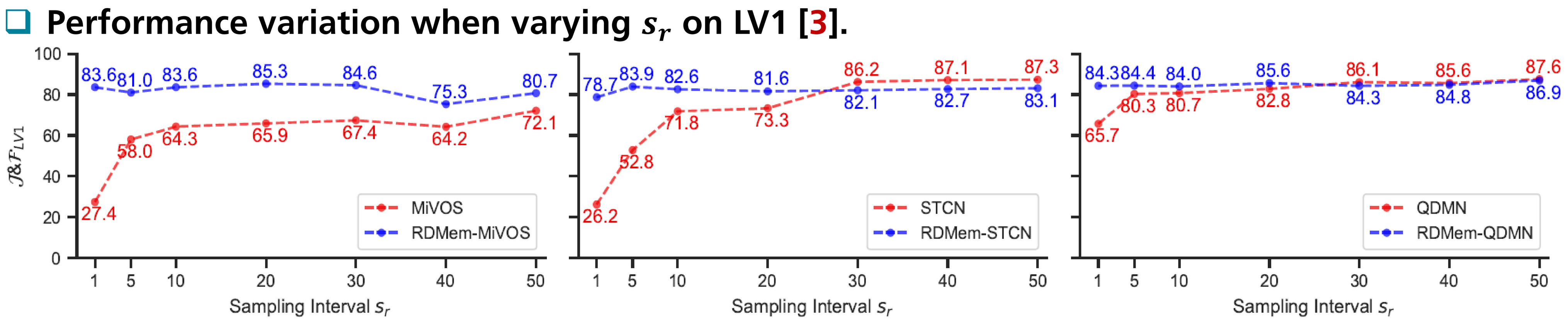
- Results on Long-Video [3] and DAVIS [4] with $s_r = 1$.

Method	LV1 [3] (Long Sequences) $\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	D17 [4] (Short Sequences) $\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
MiVOS [5] (CVPR 21)	27.4	27.4	27.4	84.3	81.3	87.3
READMem-MiVOS (ours)	83.6$^{\uparrow 56.2}$	**82.3**$^{\uparrow 54.9}$	85.0$^{\uparrow 57.6}$	84.3	81.4$^{\uparrow 0.1}$	87.1$^{\downarrow 0.2}$
STCN [6]† (NIPS 21)	26.2	22.9	29.4	83.2	79.9	86.5
READMem-STCN† (ours)	80.8$^{\uparrow 54.6}$	78.4$^{\uparrow 55.5}$	83.2$^{\uparrow 53.8}$	83.8$^{\uparrow 0.6}$	80.4$^{\uparrow 0.5}$	87.2$^{\uparrow 0.7}$
QDMN [7] (ECCV 22)	65.7	63.5	67.9	85.1	82.2	88.0
READMem-QDMN (ours)	**84.3**$^{\uparrow 18.6}$	81.9$^{\uparrow 18.4}$	**86.7**$^{\uparrow 18.8}$	**85.3**$^{\uparrow 0.2}$	**82.4**$^{\uparrow 0.2}$	88.1$^{\uparrow 0.1}$

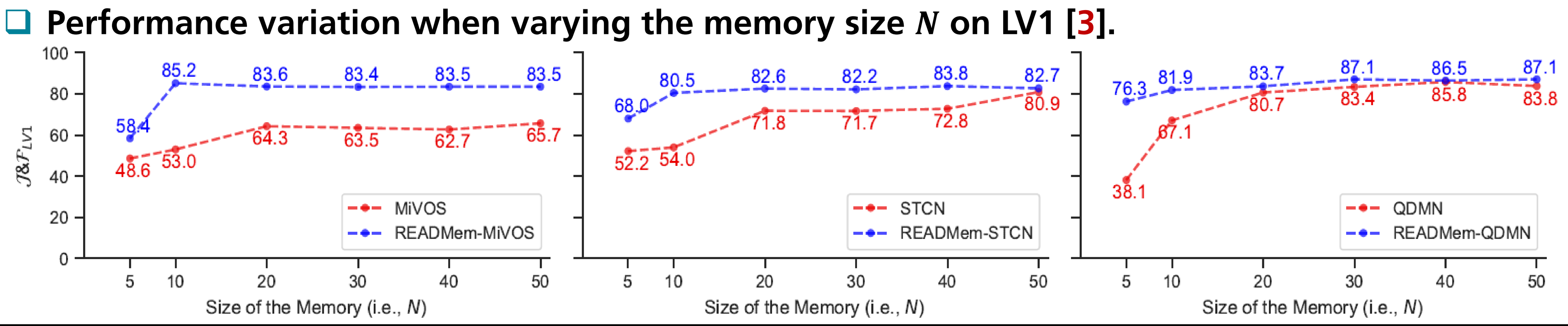
- Ablation Study.

Configuration	$\mathcal{J}\&\mathcal{F}_{\text{LV1}}$	$\mathcal{J}\&\mathcal{F}_{\text{D17}}$
MiVOS [5] + adj. frame (baseline)	64.3	84.3
MiVOS [5] + DME (ours)	69.5$^{\uparrow 5.2}$	81.3$^{\downarrow 3.0}$
MiVOS [5] + DME + LSB (ours)	75.0$^{\uparrow 10.7}$	81.3$^{\downarrow 3.0}$
MiVOS [5] + DME + LSB + adj. frame (ours)	77.4$^{\uparrow 13.1}$	84.3$^{\uparrow 0.0}$
MiVOS [5] + DME + LSB + adj. frame + REA (ours)	**86.0**$^{\uparrow 21.7}$	**84.6**$^{\uparrow 0.3}$

- Performance variation when varying s_r on LV1 [3].



- Performance variation when varying the memory size N on LV1 [3].



[1] Seoung Wug Oh, et al. *Video Object Segmentation using Space-Time Memory Networks*. ICCV, 2019.
[2] Tianfei Zhou, et al. *A Survey on Deep Learning Technique for Video Segmentation*. TPAMI, 2023.
[3] Yongqing Liang, et al. *Video Object Segmentation with Adaptive Feature Bank and Uncertain-Region Refinement*. NeurIPS, 2020.
[4] Jordi Pont-Tuset, et al. *The 2017 DAVIS Challenge on Video Object Segmentation*. arXiv, 2017.
[5] Ho Kei Cheng, et al. *Modular Interactive Video Object Segmentation: Interaction-to-Mask, Propagation and Difference-Aware Fusion*. CVPR, 2021.
6] Ho Kei Cheng, et al. *Rethinking Space-Time Networks with Improved Memory Coverage for Efficient Video Object Segmentation*. NeurIPS, 2021.
[7] Yong Liu, Ra et al. *Learning Quality-Aware Dynamic Memory for Video Object Segmentation*. ECCV, 2022.

Contact:
stephane.vujasinovic@iosb.fraunhofer.de