**Overview**

**Task:** Semi-automatic Video Object Segmentation (vSVOS).

**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/Vujasinovic-READMem

---

**Problem & Motivation**

- **Existing vSVOS methods** are tailored for short-video object segmentation:
  - Rely on an ever-expanding memory.
  - Define a dataset-specific sampling interval \( s \), non-generic to unseen data.
- **DAVIS Dataset** [4] and **LV1 Dataset** [5] show that memory diversity is crucial for accurate video segmentation.

**Contributions**

- **READMem extends any vSVOS method** to deal with unconstrained sequences.
- **Seamless integration** without re-training or fine-tuning.
- **Generalizable sampling interval** \( s \), 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).**

**Diversification of Memory Embeddings (DME):**

- **Motivation:** Increase the diversity of the memory \( M^t = \{k^n_{m,v} \} \), where \( k^n \in \mathbb{R}^{N \times s} \) with the determinant of the associated Gram Matrix by:

\[
G \in \mathbb{R}^{N \times N}, \quad \text{where } X \in \mathbb{R}^{H \times W \times N} \text{ is the flatten representation of } M^t = \{k^n_{m,v} \}
\]

- **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 \( k^n_{m,v} \) to the query’s \( (k^{q, t}_m) \) temporal frame of reference by:

\[
k^{q, t}_m = k^n_{m,v} T_{q, t}
\]

- **Process:** Compute the similarity between \( k^{q, t}_m \) and \( k^{q, t}_m \) instead of \( k^{q, t}_m \).

---

**Results on Long-Video** [3] and **DAVIS** [4] with \( s = 1 \):

- **Method**
  - MiVOS [1] (CVPR 21)
  - READMem-MiVOS (ours)
  - STCN [6] (IPS 21)
  - READMem-STCN (ours)
  - QDMN [7] (ECCV 22)
  - READMem-QDMN (ours)

- **Configuration**
  - \( \{\text{L1 [10], D17 [4] \text{ (Long Sequences) \text{ vs. Short Sequences)]} \}
  - \( \text{MiVOS [1]} \text{ [adj. frame (baseline)]} \text{ vs. READMem-MiVOS (ours)} \)
  - \( \text{STCN [6]} \text{ [DME (ours)]} \text{ vs. READMem-STCN (ours)} \)
  - \( \text{QDMN [7]} \text{ [DME (ours)]} \text{ vs. READMem-QDMN (ours)} \)

**Ablation Study.**

- **Performance variation when varying \( x \) on LV1 [3].**
- **Performance variation when varying the memory size \( N \) on LV1 [3].**

---

**Qualitative Results**

**TL;DR**

- **Motivation:** Increase the diversity of the memory \( M^t = \{k^n_{m,v} \} \), where \( k^n \in \mathbb{R}^{N \times s} \) with the determinant of the associated Gram Matrix by:

\[
G \in \mathbb{R}^{N \times N}, \quad \text{where } X \in \mathbb{R}^{H \times W \times N} \text{ is the flatten representation of } M^t = \{k^n_{m,v} \}
\]

- **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 \( k^n_{m,v} \) to the query’s \( (k^{q, t}_m) \) temporal frame of reference by:

\[
k^{q, t}_m = k^n_{m,v} T_{q, t}
\]

- **Process:** Compute the similarity between \( k^{q, t}_m \) and \( k^{q, t}_m \) instead of \( k^{q, t}_m \).