Video Object Segmentation with Episodic Graph Memory Networks *Supplemental Material*

Xiankai Lu[1], Wenguan Wang[2], Martin Danelljan[2]
Tianfei Zhou[1], Jianbing Shen[1], and Luc Van Gool[2]

[1] Inception Institute of Artificial Intelligence [2] ETH Zurich

In the supplemenatry file, we provide more details of the proposed *graph memory network*. We further report additional material including detailed experiment results. Specifically,

- in §1, we present more implementation details of the proposed *graph memory network*, including its overall structure and the memory module.
- in §2, we provide quantitative experimental results on one-shot video object segmentation (O-VOS) setting (*i.e.*, DAVIS$_{17}$ `val` set [7] and Youtube-VOS [15]).
- in §3, we provide quantitative experimental results on zero-shot video object segmentation (Z-VOS) setting (*i.e.*, DAVIS$_{16}$ `val` set [6] and Youtube-objects dataset [8]).
- in §4, we report more visualization results.

1 More Implementation Details of Graph Memory Network

Here, we present more details of the proposed graph memory network which consist the encoder-decoder segmentation network and the episodic memory network.

Segmentation network. The segmentation network is built based on Encoder-Decoder architecture. Among them, the encoder is initialized of pre-trained ResNet50 [2] on ImageNet. The input tensor of encoder is 4-channel by implanting additional single channel fiters at the first convolution layer. The first three channels are used for RGB input and the last channel is for mask input. During the pre-training on the syhthesis videos from the static images, the input frame size is 384×384. The feature map size of the fourth block of encoder is $24 \times 24 \times 512$. During the main training on the video, the input frame size is 384×640, the corresponding feature map size is $24 \times 40 \times 512$. As shown in Fig. 1, similar to RGMP [14], the decoder is consists of three blocks that each block contains a refinement module. To efficiently merge features in different scales, we employ the refinement module to take both the previous block feature as well as the features from the encoder with same scale as input. Each refinement module produces a feature map with 256 channels and the last one produces a two-channel mask map.

Episodic Graph memory module. Graph memory is a fully connected graph structure $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ where each node $v_i \in \mathcal{V}$ is represented by a feature map $\mathbf{m}_i^k$ from the encoder. k means the k-th step in the episodic graph memory. The edge function $e_{i,j} \in \mathcal{E}$ which is used for message passing is implemented by a matrix inner-product:

$$e_{i,j} = f_s(\mathbf{m}_i^k, \mathbf{m}_j^k) = softmax(\mathbf{m}_i^{k\top} \cdot \mathbf{m}_j^k), \tag{1}$$

where $softmax$ denotes the softmax normalization.

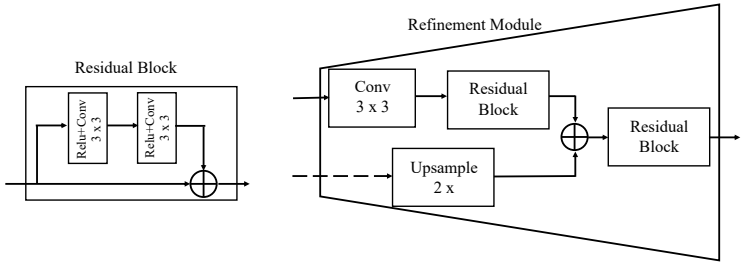

Fig. 1: The detailed architecture of decoder in our graph memory network. The refinement module takes two features as input. One feature comes from the previous block (solid line), another feature comes from encoder layer with skip connection (dashed line).

2 Additional Quantitative Results of O-VOS

DAVIS$_{17}$ Dataset. We maily compare our method with representative O-VOS methods including OSMN [16], OSVOS [1], RVOS [12], RGMP [14], AGAME [3] and STM [5]. Table 1 reports the per-sequence evaluation results in terms of region similarity $\mathcal{J}$ and boundary accuracy $\mathcal{F}$.

3 Additional Quantitative Results of Z-VOS

DAVIS$_{16}$ Dataset. We compare our MuG with representative Z-VOS methods including PDB [10], MotAdapt [9], LSMO [11], AGS [13], COSNet [4], and AnDiff [17].

Table 2 gives per-sequence evaluation in terms of region similarity $\mathcal{J}$ and boundary accuracy $\mathcal{F}$. As shown in Table 2, our model outperforms previous methods across the vast majority of sequences and on average.

4 Additional Qualitative Results

In this section, we present a qualitative evaluation of the proposed graph memory network on the sequences of O-VOS datasets: DAVIS$_{17}$ [7], Youtube-VOS [15] and Z-VOS datasets: DAVIS$_{16}$ [6], Youtube-objects [8]. Specifically, Fig. 2 shows the visualization results of O-VOS while Fig. 3 shows the visualization results of Z-VOS.

Dataset	Video	OSMN [16]		OSVOS [1]		RVOS [12]		RGMP [14]		AGAME [3]		STM [5]		**Ours**	
		$\mathcal{J}\uparrow$	$\mathcal{F}\uparrow$	$\mathcal{J}\uparrow$	$\mathcal{F}\uparrow$	$\mathcal{J}\uparrow$	$\mathcal{F}\uparrow$	$\mathcal{J}\uparrow$	$\mathcal{F}\uparrow$	$\mathcal{J}\uparrow$	$\mathcal{F}\uparrow$	$\mathcal{J}\uparrow$	$\mathcal{F}\uparrow$	$\mathcal{J}\uparrow$	$\mathcal{F}\uparrow$
	bike-packing	51.5	48.8	62.1	70.0	55.5	58.5	48.6	56.6	69.2	74.2	79.8	85.2	80.7	86.6
	blackswan	89.9	92.7	94.3	97.4	93.9	96.5	96.0	98.6	79.0	80.1	96.2	99.8	96.2	99.7
	bmx-tress	43.0	66.0	47.6	73.0	30.8	56.8	44.4	65.6	43.4	66.2	58.0	88.6	59.1	89.2
	breakdance	71.1	68.2	72.7	75.6	42.2	45.5	59.2	63.1	57.8	64.0	89.7	91.5	89.5	91.2
	camel	88.4	92.2	85.9	89.1	73.3	80.0	74.7	85.4	83.0	87.9	96.4	98.5	96.2	98.3
	car-roundabout	93.4	92.0	89.0	82.7	92.6	87.2	95.0	90.4	97.5	96.2	98.2	97.1	98.6	97.2
	car-shadow	90.5	92.2	92.8	91.7	93.3	98.8	96.3	99.7	95.9	99.4	96.8	99.7	96.8	99.7
	cow	87.2	86.5	95.2	95.7	91.2	92.5	93.7	93.0	93.9	96.3	95.5	98.0	95.8	98.7
	dance-twirl	75.8	72.9	64.1	71.6	62.3	61.5	83.8	82.5	84.9	86.5	86.2	87.0	86.3	87.4
	dog	87.7	84.6	71.1	69.1	93.3	94.6	95.5	96.2	94.5	97.6	95.2	98.2	95.9	98.2
DAVIS$_{17}$	dogs-jump	38.8	45.2	58.8	68.5	69.5	69.4	68.4	62.8	85.5	90.1	89.8	88.5	92.1	96.6
	drift-chicane	4.9	8.4	77.4	82.9	57.0	67.6	79.5	79.4	82.2	92.0	90.3	98.5	92.4	97.5
	drift-straight	66.4	57.7	66.5	70.7	89.4	85.4	91.3	86.7	92.5	89.8	94.5	95.2	93.6	94.2
	goat	80.4	74.7	86.9	87.6	84.4	83.2	86.4	85.1	88.4	89.0	90.7	93.7	91.0	94.0
	gold-fish	50.3	49.5	53.7	56.6	60.6	62.5	69.2	69.9	58.7	62.3	74.8	70.1	84.8	86.6
	horsejump-high	39.5	49.5	70.0	83.7	29.2	40.0	78.5	91.1	72.5	87.8	84.0	97.0	84.1	97.0
	india	59.4	55.1	28.8	31.5	34.5	43.6	41.1	37.0	55.1	58.4	80.7	79.4	81.2	79.6
	judo	46.0	52.2	44.1	55.0	74.3	62.5	64.0	76.4	67.6	74.3	86.9	89.6	87.4	90.0
	kite-surf	23.6	46.3	43.1	61.1	27.7	49.4	32.0	41.8	41.8	49.4	54.1	73.8	54.4	73.8
	lab-coat	41.3	38.1	21.1	29.0	63.4	48.0	51.3	77.2	53.3	66.3	56.7	56.1	59.3	56.5
	libby	44.7	63.6	62.3	75.1	56.5	75.4	43.9	50.7	85.8	96.3	90.2	98.4	90.1	98.2
	loading	60.4	66.2	57.8	59.3	56.3	57.1	60.1	62.4	78.6	78.8	89.5	91.8	89.8	91.8
	mbike-trick	72.5	76.6	64.7	72.1	35.3	57.4	69.6	70.5	67.7	73.6	81.3	84.0	81.6	84.7
	motorcross-jump	39.4	39.0	57.9	56.8	73.7	75.4	33.0	31.9	72.2	69.1	87.8	84.0	88.0	87.3
	paragliding-launch	38.0	58.1	47.8	67.9	29.0	34.7	42.3	48.4	38.4	49.6	53.1	68.5	53.7	68.0
	parkour	86.2	93.2	77.3	74.1	87.5	90.4	91.8	95.3	92.9	96.9	94.5	97.6	94.4	97.4
	pigs	64.4	63.7	52.7	58.5	76.5	76.4	67.3	68.0	63.1	65.7	83.3	85.8	83.4	86.1
	scooter-black	62.2	63.2	38.4	50.5	37.8	43.6	45.6	56.4	50.8	63.4	84.2	89.9	84.1	92.1
	shooting	42.3	53.7	61.3	65.7	46.6	55.2	61.6	65.5	57.5	67.6	68.4	70.6	68.3	70.3
	soapbox	45.8	50.3	44.7	58.5	48.9	56.7	70.6	74.5	66.5	74.6	72.5	77.7	72.4	81.0
	Average	52.5	57.1	56.7	63.9	68.5	73.6	70.5	74.7	68.5	73.6	79.5	84.5	**80.0**	**85.9**

Table 1: Evaluation of O-VOS on DAVIS$_{17}$ `val` set [7], with the region similarity $\mathcal{J}$ and boundary accuracy $\mathcal{F}$. For both two measure metrics, higher values are better.

Dataset	Video	PDB [10]		MotAdapt [9]		LSMO [11]		AGS [13]		COSNet [4]		AnDiff [17]		**Ours**	
		$\mathcal{J}\uparrow$	$\mathcal{F}\uparrow$	$\mathcal{J}\uparrow$	$\mathcal{F}\uparrow$	$\mathcal{J}\uparrow$	$\mathcal{F}\uparrow$	$\mathcal{J}\uparrow$	$\mathcal{F}\uparrow$	$\mathcal{J}\uparrow$	$\mathcal{F}\uparrow$	$\mathcal{J}\uparrow$	$\mathcal{F}\uparrow$	$\mathcal{J}\uparrow$	$\mathcal{F}\uparrow$
DAVIS$_{16}$	Blackswan	90.8	93.2	93.9	91.9	92.9	93.9	94.4	96.7	88.0	89.9	94.4	96.2	94.5	95.5
	Bmx-Trees	49.9	61.3	46.2	45.8	49.9	66.8	51.4	66.4	46.5	63.3	56.6	77.1	50.2	67.1
	Breakdance	59.0	55.1	35.2	36.2	45.9	43.6	60.7	58.1	68.3	63.8	41.9	35.6	73.1	71.5
	Camel	82.4	84.7	84.6	82.9	88.6	89.4	85.7	85.1	89.4	90.8	89.6	91.9	90.5	90.6
	Car-R-about	85.9	79.7	87.1	67.8	85.9	79.3	94.9	91.7	94.7	92.7	94.3	90.7	95.1	98.1
	Car-Shadow	91.8	92.8	75.9	94.9	88.0	85.8	91.8	95.5	93.5	97.7	95.8	98.7	95.7	98.3
	Cows	91.8	90.2	97.5	94.4	90.9	90.0	92.2	93.7	91.4	93.6	94.7	96.3	91.8	93.0
	Dance-Twirl	65.8	60.3	68.1	67.0	83.1	81.7	78.7	76.2	77.7	77.2	71.6	69.3	79.9	79.8
	Dog	92.4	91.1	96.0	93.9	92.9	94.5	93.5	93.4	93.7	95.5	95.6	97.6	92.9	93.8
	Drift-Chicane	60.7	65.4	85.1	70.0	69.6	79.1	69.9	77.1	77.7	77.1	71.1	80.4	82.5	91.3
	Drift-Straight	86.8	79.9	90.9	90.0	82.6	67.0	90.0	88.6	93.7	95.5	90.7	87.1	91.4	89.3
	Goat	83.7	80.8	88.4	88.3	84.4	82.3	84.7	82.8	70.5	78.8	88.8	90.4	84.7	81.8
	Horsejump-H	85.7	91.6	93.9	87.8	86.2	92.6	73.4	74.9	91.7	93.5	88.5	95.3	84.9	90.9
	Kite-Surf	67.4	49.8	52.4	68.9	50.3	45.4	68.7	49.3	67.5	55.1	67.6	52.5	66.7	53.0
	Libby	73.1	82.6	93.3	83.3	78.0	87.3	66.5	78.1	68.9	81.9	85.9	95.1	75.6	84.2
	Motocross-J	85.4	74.1	77.5	85.1	82.3	70.9	81.8	69.0	82.5	72.5	86.7	79.2	72.3	66.6
	Paragliding-L	63.5	23.2	28.5	64.1	63.3	23.2	63.1	21.6	61.2	19.9	63.4	23.5	63.1	21.8
	Parkour	90.1	92.9	93.6	90.6	89.2	93.4	90.8	93.7	87.7	92.1	93.3	96.3	91.5	93.6
	Scooter-Black	68.5	63.1	58.8	53.7	70.9	65.1	75.1	66.1	83.8	75.6	81.2	73.6	85.8	79.1
	Soapbox	73.4	73.0	76.2	71.6	88.1	87.5	76.2	75.8	87.3	86.7	82.5	82.6	89.6	86.3
	Average	77.2	77.4	77.2	74.5	78.2	75.9	79.7	77.3	80.5	79.5	81.7	80.4	82.5	81.2

Table 2: Evaluation of object-level Z-VOS on DAVIS$_{16}$ `val` set [6], with region similarity $\mathcal{J}$ and boundary accuracy $\mathcal{F}$. For both two measure metrics, higher values are better.

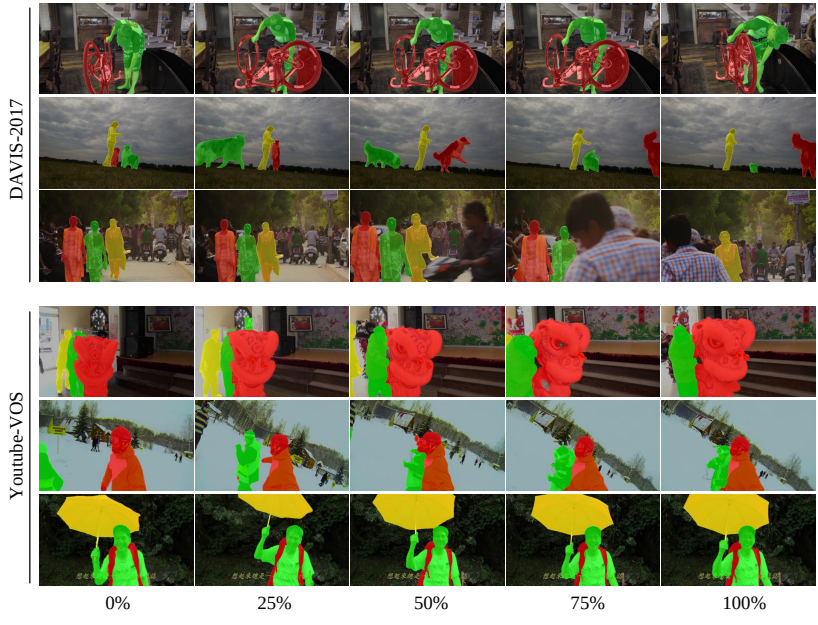

Fig. 2: Qualitative results on O-VOS datasets. From top to bottom are bikepacking, dog-jump, india from DAVIS$_{17}$ and 0788b4033d, 2caa2b45c7, 03deb7ad95 from Youtube-VOS.

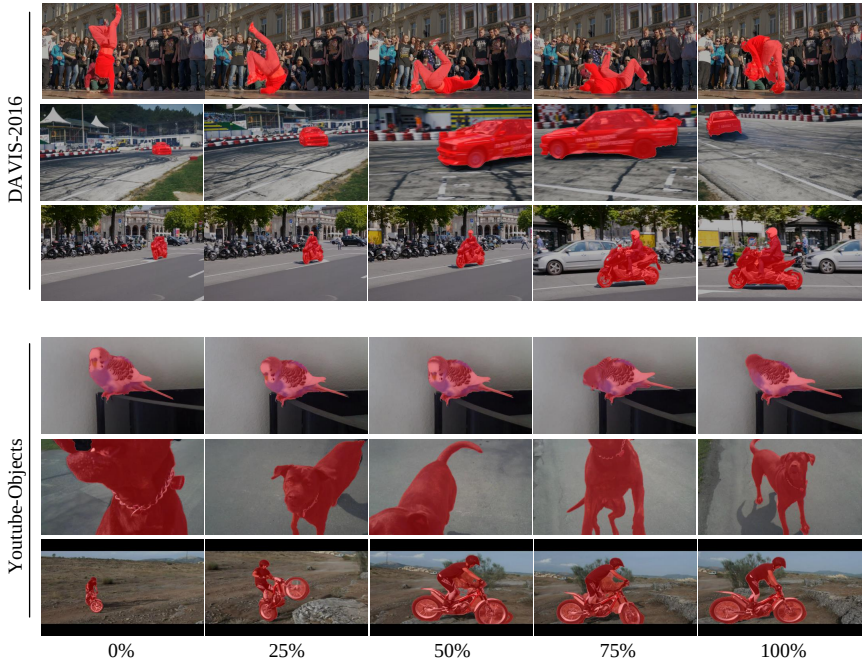

Fig. 3: Qualitative results on Z-VOS datasets. From top to bottom are breakdance, car-roundabout, scooter from DAVIS_{16} and bird, dog, motorbike from Youtube-objects.

References

1. Caelles, S., Maninis, K.K., Pont-Tuset, J., Leal-Taixé, L., Cremers, D., Van Gool, L.: One-shot video object segmentation. In: CVPR (2017)
2. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016)
3. Johnander, J., Danelljan, M., Brissman, E., Khan, F.S., Felsberg, M.: A generative appearance model for end-to-end video object segmentation. In: CVPR (2019)
4. Lu, X., Wang, W., Ma, C., Shen, J., Shao, L., Porikli, F.: See more, know more: Unsupervised video object segmentation with co-attention siamese networks. In: CVPR (2019)
5. Oh, S.W., Lee, J.Y., Xu, N., Kim, S.J.: Video object segmentation using space-time memory networks. In: ICCV (2019)
6. Perazzi, F., Pont-Tuset, J., McWilliams, B., Van Gool, L., Gross, M., Sorkine-Hornung, A.: A benchmark dataset and evaluation methodology for video object segmentation. In: CVPR (2016)
7. Pont-Tuset, J., Perazzi, F., Caelles, S., Arbeláez, P., Sorkine-Hornung, A., Van Gool, L.: The 2017 davis challenge on video object segmentation. arXiv preprint arXiv:1704.00675 (2017)
8. Prest, A., Leistner, C., Civera, J., Schmid, C., Ferrari, V.: Learning object class detectors from weakly annotated video. In: CVPR (2012)
9. Siam, M., Jiang, C., Lu, S., Petrich, L., Gamal, M., Elhoseiny, M., Jagersand, M.: Video segmentation using teacher-student adaptation in a human robot interaction (hri) setting. In: ICRA (2019)
10. Song, H., Wang, W., Zhao, S., Shen, J., Lam, K.M.: Pyramid dilated deeper convlstm for video salient object detection. In: ECCV (2018)
11. Tokmakov, P., Schmid, C., Alahari, K.: Learning to segment moving objects. IJCV **127**(3), 282–301 (2019)
12. Ventura, C., Bellver, M., Girbau, A., Salvador, A., Marques, F., Giro-i Nieto, X.: Rvos: End-to-end recurrent network for video object segmentation. In: CVPR (2019)
13. Wang, W., Song, H., Zhao, S., Shen, J., Zhao, S., Hoi, S.C., Ling, H.: Learning unsupervised video object segmentation through visual attention. In: CVPR (2019)
14. Wug Oh, S., Lee, J.Y., Sunkavalli, K., Joo Kim, S.: Fast video object segmentation by reference-guided mask propagation. In: CVPR (2018)
15. Xu, N., Yang, L., Fan, Y., Yang, J., Yue, D., Liang, Y., Price, B., Cohen, S., Huang, T.: Youtube-vos: Sequence-to-sequence video object segmentation. In: ECCV (2018)
16. Yang, L., Wang, Y., Xiong, X., Yang, J., Katsaggelos, A.K.: Efficient video object segmentation via network modulation. In: CVPR (2018)
17. Yang, Z., Wang, Q., Bertinetto, L., Bai, S., Hu, W., Torr, P.H.: Anchor diffusion for unsupervised video object segmentation. In: ICCV (2019)