Supplementary Material: Video Object Segmentation using Space-Time Memory Networks

Seoung Wug Oh* Yonsei University
Joon-Young Lee Adobe Research
Ning Xu Adobe Research
Seon Joo Kim Yonsei University

1. Network Structure Details

We provide more detailed descriptions of our three sub-networks: the memory encoder, the query encoder, and the decoder. These sub-networks are illustrated in Fig. 1. The backbone network for the encoders is the ResNet50 [7]. We used the residual block [8] and the refinement module [22, 19] to build the decoder.

2. Multi-Object Segmentation Details

In this section, we add more in-depth details of our mask merging operation for multi-object segmentation. If there exist more than one object in the video, mask probability maps for every object are independently computed by running our model for each object. Then, the predicted maps are merged using a soft aggregation operation defined as:

$$p_{i,m} = \sigma\big(l(\hat{p}_{i,m})\big) = \frac{\hat{p}_{i,m}/(1-\hat{p}_{i,m})}{\sum_{j=0}^{M}\hat{p}_{i,j}/(1-\hat{p}_{i,j})}, \quad \text{s.t.} \quad \hat{p}_{i,0} = \Pi_{j=1}^{M}(1-\hat{p}_{i,j}), \quad (1)$$

where σ and l represent the softmax and the logit function respectively, $\hat{p}_{i,m}$ is the network output probability of the object m at the pixel location i, m=0 indicates the background, and M is the total number of objects.

The above operation is originally proposed in [19]. In [19], the mask merging is performed during the testing as a post-processing step. Different from the original work, we coin Equation (1) as a differential network module and apply it during both the training and the testing. Using the mask merging module, our network outputs per-pixel M+1 way classification results (similar to the semantic segmentation) and it can be trained *end-to-end* using the cross entropy loss. Besides, if there are multiple objects, we provide additional information to the memory encoder about other objects. Specifically, a probability mask for all other objects, computed as $o_{i,m} = \sum_{j\neq m}^{M} p_{i,j}$, is additionally given.

*This work was done during an internship at Adobe Research.

3. More Results

Full results on the DAVIS-2016. In Table 2, we provide a full table with results on the DAVIS-2016 benchmarks [21]. We include results omitted in the main paper due to the space limit, *e.g.* methods with the mean $\mathcal{J}$ score below 79 and variants of some methods evaluated without online learning.

Results on DAVIS-2017 test-dev set. In Table 2, we report the results of multi-object video segmentation on the DAVIS-2017 test-dev set. In case of test-dev set, we resize the test video to be 600p to handle small objects. And, for some videos with many objects (*e.g.* salsa), we reduced the memory saving frequency to avoid GPU memory overflow.

4. Video Comparisons on DAVIS.

We provide side-by-side comparisons on the DAVIS benchmark [20, 23] in the video file `Comparisons_DAVIS.mp4`. We compare our method against three state-of-the-art methods: OSVOS [2], RGMP [19], and PReMVOS [16]. We choose some challenging video sequences from both DAVIS 2016 [20] and 2017 [23]. The pre-computed results of other methods are downloaded from the DAVIS benchmark leaderboard[1]

References

[1] Linchao Bao, Baoyuan Wu, and Wei Liu. Cnn in mrf: Video object segmentation via inference in a cnn-based higher-order spatio-temporal mrf. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2, 3

[2] Sergi Caelles, Kevis-Kokitsi Maninis, Jordi Pont-Tuset, Laura Leal-Taixé, Daniel Cremers, and Luc Van Gool. One-shot video object segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017. 1, 2, 3

[3] Yuhua Chen, Jordi Pont-Tuset, Alberto Montes, and Luc Van Gool. Blazingly fast video object segmentation with pixel-wise metric learning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 3

[1] https://davischallenge.org/davis2017/soa_compare.html

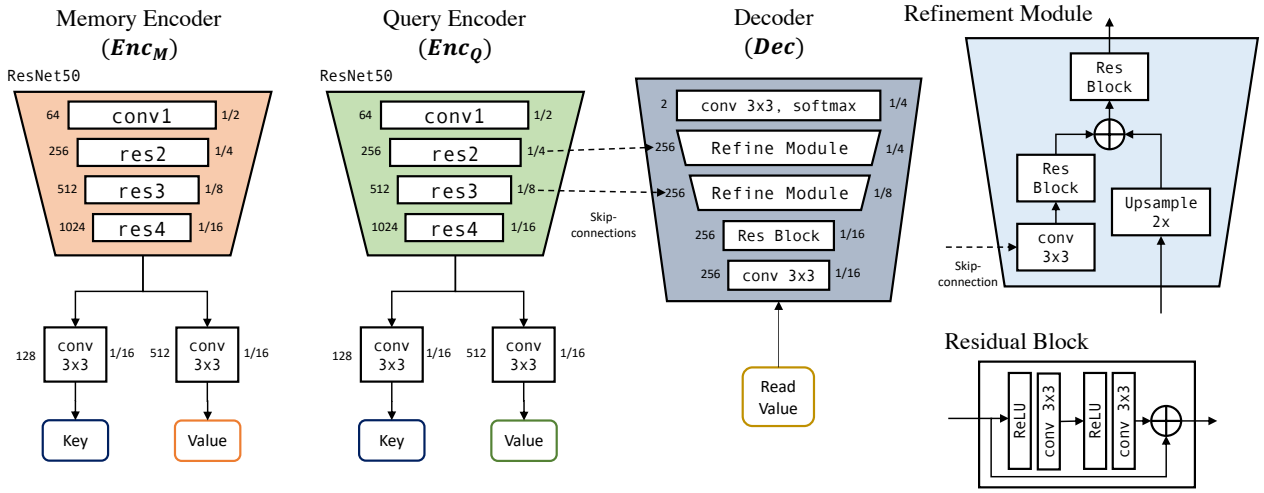

Figure 1: A detailed illustration of three sub-networks: the memory encoder, the query encoder, and the decoder. $\bigoplus$ indicates element-wise addition. The output channel dimension and the relative spatial scale of each layer (block) is shown on the left and the right, respectively.

	OL	$\mathcal{J}$ Mean	$\mathcal{F}$ Mean
OSMN [28]		37.7	44.9
FAVOS [4]		42.9	44.2
OSVOS [2]	✓	47.0	54.8
OnAVOS [26]	✓	49.9	55.7
OSVOS^S [2]	✓	52.9	62.1
RGMP [19]		51.3	54.4
FEELVOS [25]		55.1	60.4
Lucid [14]	✓	63.4	69.9
CINN [1]	✓	64.5	70.5
DyeNet [15]	✓	65.8	70.5
PReMVOS [16]	✓	67.5	**75.7**
Ours		**69.3**	75.2

Table 1: The quantitative evaluation on DAVIS-2017 test-dev set. OL indicates online learning.

[4] Jingchun Cheng, Yi-Hsuan Tsai, Wei-Chih Hung, Shengjin Wang, and Ming-Hsuan Yang. Fast and accurate online video object segmentation via tracking parts. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2, 3

[5] Jingchun Cheng, Yi-Hsuan Tsai, Shengjin Wang, and Ming-Hsuan Yang. Segflow: Joint learning for video object segmentation and optical flow. *IEEE International Conference on Computer Vision (ICCV)*, 2017. 3

[6] Hai Ci, Chunyu Wang, and Yizhou Wang. Video object segmentation by learning location-sensitive embeddings. In *European Conference on Computer Vision (ECCV)*, 2018. 3

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 1

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *European Conference on Computer Vision (ECCV)*, pages 630–645. Springer, 2016. 1

[9] Yuan-Ting Hu, Jia-Bin Huang, and Alexander Schwing. Maskrnn: Instance level video object segmentation. In *Advances in Neural Information Processing Systems*, 2017. 3

[10] Yuan-Ting Hu, Jia-Bin Huang, and Alexander Schwing. Videomatch: Matching based video object segmentation. In *European Conference on Computer Vision (ECCV)*, 2018. 3

[11] B. Leibe J. Luiten, P. Voigtlaender. Premvos: Proposal-generation, refinement and merging for the davis challenge on video object segmentation 2018. *The 2018 DAVIS Challenge on Video Object Segmentation - CVPR Workshops*, 2018. 3

[12] Varun Jampani, Raghudeep Gadde, and Peter V Gehler. Video propagation networks. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 3

[13] Joakim Johnander, Martin Danelljan, Emil Brissman, Fahad Shahbaz Khan, and Michael Felsberg. A generative appearance model for end-to-end video object segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 3

[14] Anna Khoreva, Rodrigo Benenson, Eddy Ilg, Thomas Brox, and Bernt Schiele. Lucid data dreaming for video object segmentation. *International Journal of Computer Vision*, pages 1–23. 2

[15] Xiaoxiao Li and Chen Change Loy. Video object segmentation with joint re-identification and attention-aware mask propagation. In *European Conference on Computer Vision (ECCV)*, 2018. 2, 3

	OL	$\mathcal{J}$ Mean	$\mathcal{F}$ Mean	Time
MaskRNN† [9]		56.3	-	-
BVS [18]		60.0	58.8	0.37s
SFL† [5]		67.4	66.7	-
OFL [24]		68.0	63.4	120s
MSK† [20]		69.9	-	-
PLM [29]	✓	70.0	62.0	0.3s
VPN [12]		70.2	65.5	0.63s
OSMN [28]		74.0	72.9	0.14s
SFL [5]	✓	74.8	74.5	7.9s
PML [3]		75.5	79.3	0.27s
S2S **(+YV)** [27]	✓	79.1	-	9s
MSK [20]	✓	79.7	75.4	12s
OSVOS [2]	✓	79.8	80.6	9s
MaskRNN [9]	✓	80.7	80.9	-
VidMatch [10]		81.0	-	0.32s
FEELVOS **(+YV)** [25]		81.1	82.2	0.45s
RGMP [19]		81.5	82.0	0.13s
A-GAME **(+YV)** [13]		82.0	-	0.07s
FAVOS [4]		82.4	79.5	1.8s
LSE [6]	✓	82.9	80.3	-
CINN [1]	✓	83.4	85.0	>30s
PReMVOS [11]	✓	84.9	88.6	>30s
OSVOS^S [17]	✓	85.6	86.4	4.5s
OnAVOS [26]	✓	86.1	84.9	13s
DyeNet [15]	✓	86.2	-	2.32s
Ours		84.8	88.1	0.16s
Ours **(+YV)**		**88.7**	**89.9**	0.16s

Table 2: Quantitative evaluation on DAVIS-2016 validation set. OL indicates online learning. Time shows the approximated runtime (seconds per frame). † indicates a variant without the use of online learning. **(+YV)** indicates the use of Youtube-VOS for training.

[16] Jonathon Luiten, Paul Voigtlaender, and Bastian Leibe. Premvos: Proposal-generation, refinement and merging for video object segmentation. In *Asian Conference on Computer Vision*, 2018. 1, 2

[17] Kevis-Kokitsi Maninis, Sergi Caelles, Yuhua Chen, Jordi Pont-Tuset, Laura Leal-Taixé, Daniel Cremers, and Luc Van Gool. Video object segmentation without temporal information. *arXiv preprint arXiv:1709.06031*, 2017. 3

[18] Nicolas Märki, Federico Perazzi, Oliver Wang, and Alexander Sorkine-Hornung. Bilateral space video segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 743–751, 2016. 3

[19] Seoung Wug Oh, Joon-Young Lee, Kalyan Sunkavalli, and Seon Joo Kim. Fast video object segmentation by reference-guided mask propagation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 1, 2, 3

[20] Federico Perazzi, Anna Khoreva, Rodrigo Benenson, Bernt Schiele, and Alexander Sorkine-Hornung. Learning video object segmentation from static images. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 1, 3

[21] Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus Gross, and Alexander Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 1

[22] Pedro O Pinheiro, Tsung-Yi Lin, Ronan Collobert, and Piotr Dollár. Learning to refine object segments. In *European Conference on Computer Vision (ECCV)*, pages 75–91. Springer, 2016. 1

[23] Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alexander Sorkine-Hornung, and Luc Van Gool. The 2017 davis challenge on video object segmentation. *arXiv:1704.00675*, 2017. 1

[24] Yi-Hsuan Tsai, Ming-Hsuan Yang, and Michael J Black. Video segmentation via object flow. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3899–3908, 2016. 3

[25] Paul Voigtlaender, Yuning Chai, Florian Schroff, Hartwig Adam, Bastian Leibe, and Liang-Chieh Chen. Feelvos: Fast end-to-end embedding learning for video object segmentation. In *CVPR*, 2019. 2, 3

[26] Paul Voigtlaender and Bastian Leibe. Online adaptation of convolutional neural networks for video object segmentation. In *British Machine Vision Conference*, 2017. 2, 3

[27] Ning Xu, Linjie Yang, Dingcheng Yue, Jianchao Yang, Brian Price, Jimei Yang, Scott Cohen, Yuchen Fan, Yuchen Liang, and Thomas Huang. Youtube-vos: Sequence-to-sequence video object segmentation. In *European Conference on Computer Vision (ECCV)*, 2018. 3

[28] Linjie Yang, Yanran Wang, Xuehan Xiong, Jianchao Yang, and Aggelos K Katsaggelos. Efficient video object segmentation via network modulation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2, 3

[29] Jae Shin Yoon, Francois Rameau, Junsik Kim, Seokju Lee, Seunghak Shin, and In So Kweon. Pixel-level matching for video object segmentation using convolutional neural networks. In *IEEE International Conference on Computer Vision (ICCV)*, 2017. 3