Supplementary Material for Learning to Recommend Frame for Interactive Video Object Segmentation in the Wild

Zhaoyuan Yin[1], Jia Zheng[2], Weixin Luo[3], Shenhan Qian[4], Hanling Zhang[5]*, Shenghua Gao[4,6]
[1]College of Computer Science and Electronic Engineering, Hunan University
[2]KooLab, Manycore [3]Meituan Group
[4]ShanghaiTech University [5]School of Design, Hunan University
[6]Shanghai Engineering Research Center of Intelligent Vision and Imaging
{zyyin, jh_hlzhang}@hnu.edu.cn jiajia@qunhemail.com luoweixin@meituan.com
{qianshh, gaoshh}@shanghaitech.edu.cn

In this supplementary material, we first present the details of the network architecture. Then, we show the quantitative results on YouTube-VOS dataset. Finally, we provide more qualitative results of IPN [3] and MANet [2] on DAVIS dataset [4] and YouTube-VOS dataset [5].

Network architecture. We divide state s^t into a sequence of N pairs of segmentation quality q^t and recommendation history h^t, and process it frame by frame. The state s^t is firstly fed into two fully connected layers with both 128 feature dimensions to obtain feature sequence. Then, we use a Bi-Directional LSTM unit with 128 hidden size to capture the temporal information of the feature sequence. Finally, Q value of each frame is obtained via two fully connected layers with 128 and 1 feature dimensions. The detail of the network architecture is shown in Figure 1.

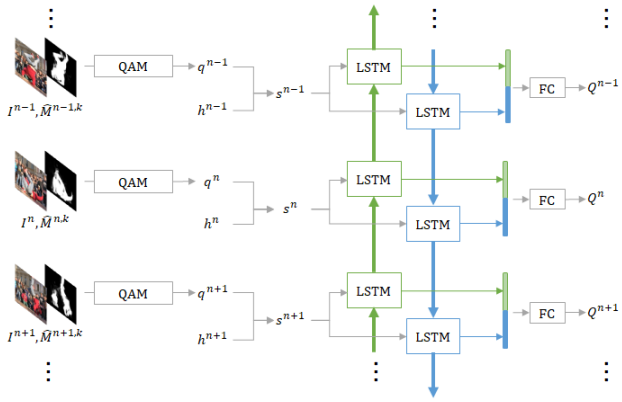


Figure 1. Network architecture. QAM denotes the segmentation quality assessment module.

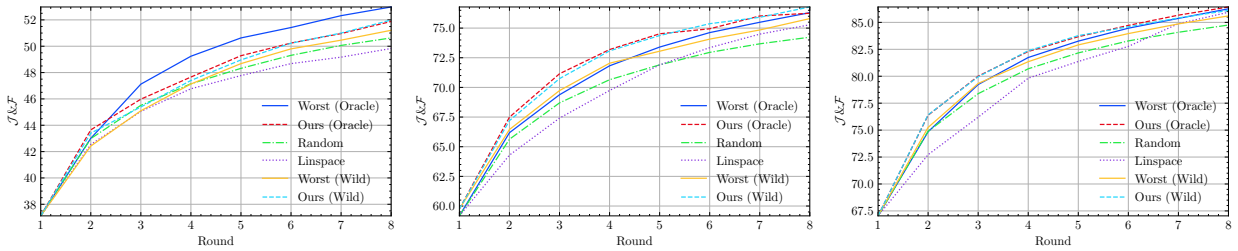
Quantitative results. Figure 2 shows the curves of the $\mathcal{J}\&\mathcal{F}$ versus the number of rounds on YouTube-VOS dataset.

Qualitative results. We show more qualitative results for the IPN [3] and MANet [2] in Figure 3a and 3b. Similar to the case of ATNet [1], the interactive video object segmentation algorithms combined with our agent can produce more accurate segmentation masks.

References

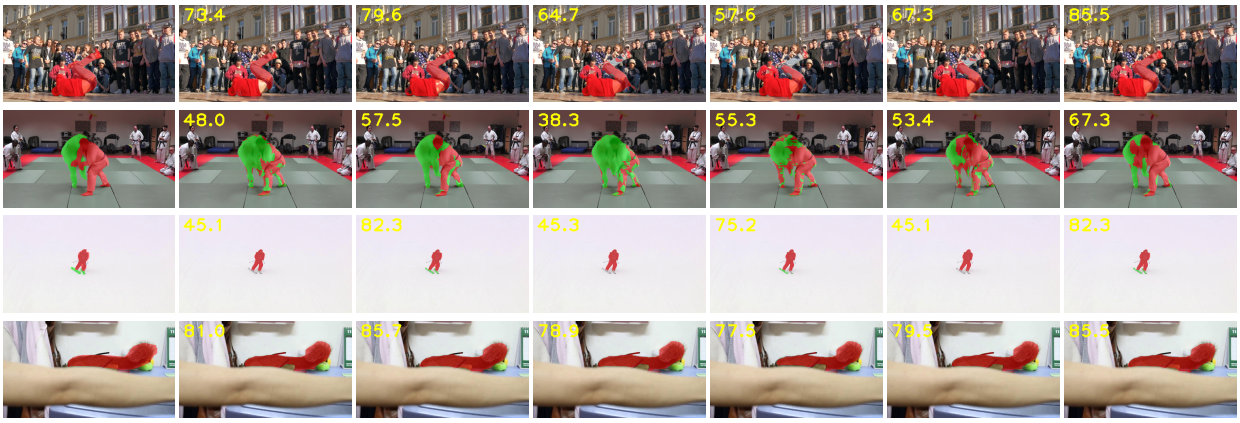
[1] Yuk Heo, Yeong Jun Koh, and Chang-Su Kim. Interactive video object segmentation using global and local transfer modules. In *ECCV*, pages 297–313, 2020. 1, 2
[2] Jiaxu Miao, Yunchao Wei, and Yi Yang. Memory aggregation networks for efficient interactive video object segmentation. In *CVPR*, pages 10366–10375, 2020. 1, 2
[3] Seoung Wug Oh, Joon-Young Lee, Ning Xu, and Seon Joo Kim. Fast user-guided video object segmentation by interaction-and-propagation networks. In *CVPR*, pages 5247–5256, 2019. 1, 2
[4] Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alexander Sorkine-Hornung, and Luc Van Gool. The 2017 davis challenge on video object segmentation. *CoRR*, abs/1704.00675, 2017. 1
[5] Ning Xu, Linjie Yang, Yuchen Fan, Jianchao Yang, Dingcheng Yue, Yuchen Liang, Brian Price, Scott Cohen, and Thomas Huang. Youtube-vos: Sequence-to-sequence video object segmentation. In *ECCV*, pages 585–601, 2018. 1

*Corresponding author.

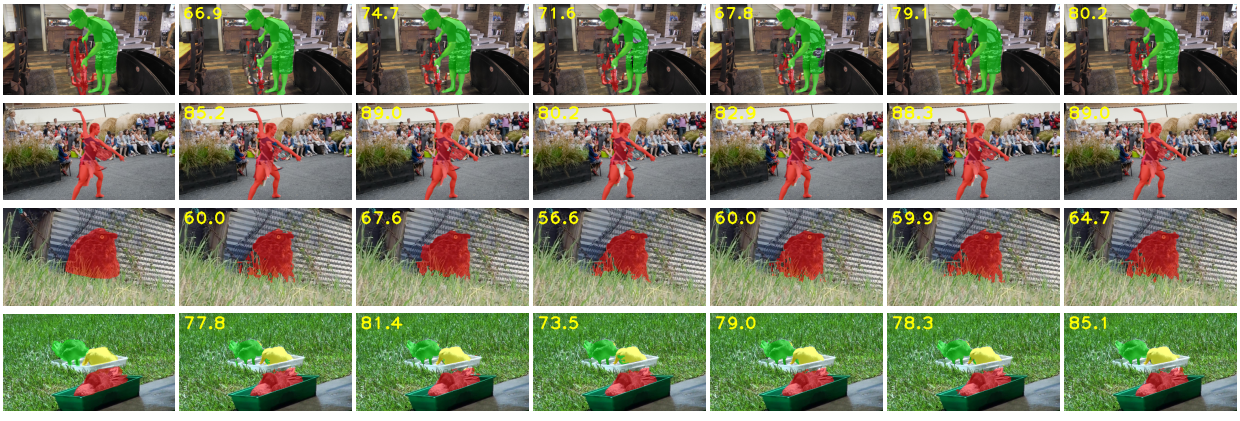


(a) IPN [3] (b) MANet [2] (c) ATNet [1]

Figure 2. The curve of $\mathcal{J}\&\mathcal{F}$ versus the number of rounds on YouTube-VOS dataset.



(a) IPN [3].



(b) MANet [2].

Figure 3. Qualitative comparison on DAVIS (first two rows) and YouTube-VOS dataset (the last two rows). All result masks are sampled after 8 rounds. The ground truth is available ("Oracle") in the second and third columns, while the ground truth is unknown ("Wild") in the last four columns. We show the segmentation quality $\mathcal{J}\&\mathcal{F}$ on each frame.