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

Zhaoyuan Yin\textsuperscript{1}, Jia Zheng\textsuperscript{2}, Weixin Luo\textsuperscript{3}, Shenhan Qian\textsuperscript{4}, Hanling Zhang\textsuperscript{5,*}, Shenghua Gao\textsuperscript{4,6}

\textsuperscript{1}College of Computer Science and Electronic Engineering, Hunan University\textsuperscript{2}KooLab, Manycore \textsuperscript{3}Meituan Group\textsuperscript{4}ShanghaiTech University \textsuperscript{5}School of Design, Hunan University\textsuperscript{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.

![Network Architecture](image)

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

**Quantitative results.** Figure 2 shows the curves of the $J$&$F$ 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**


*Corresponding author.
Figure 2. The curve of $J\&F$ versus the number of rounds on YouTube-VOS dataset.

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 $J\&F$ on each frame.