Modular Interactive Video Object Segmentation: Interaction-to-Mask, Propagation and Difference-Aware Fusion Supplementary Material

1. Notations

This is a quick index of the notations in the order of appearance.

General notations	
r	The current interaction round index.
t^r	The index of the current user-interacting frame.
M^r	The mask results of all the frames of the current round.
M^r_j	The mask result of the j-th frame of the current round.
M^{r-1}	The mask results of all the frames of the previous round.
Notations related to interaction	
$M^{r-1}_{t_r}$	The previous mask result of the current user-interacting frame.
Notations related to propagation	
H, W	The spatial dimensions of the features after the encoder with stride 16.
C^k, C^v	The channel dimensions of the "key" and "value" features respectively.
T	The number of frames in the memory bank.
$\mathbf{k}^M, \mathbf{v}^M$	The extracted "key" and "value" features from the memory encoder.
$\mathbf{k}^Q, \mathbf{v}^Q$	The extracted "key" and "value" features from the query encoder.
$\mathbf{F}_{ij}$	The dot product between the query feature at position i and the memory feature at position j.
$\mathbf{W}_{ij}$	The normalized affinity between the query feature at position i and the memory feature at position j.
k	A hyperparameter. Top-k filtering is applied along the memory dimension.
$\text{Top}^k_j(\mathbf{F})$	Indices of affinities that are top-k in the j-th column (i.e., memory) of $\mathbf{F}$.
$\mathbf{m}_j$	The aggregated memory feature for the query position j.
Notations related to fusion	
t_i	The target frame index to be fused.
t^c	The closest previously interacted frame index in the direction of propagation.
$M^{r'}$	The current mask results after propagation and before fusion.
$\mathcal{D}^+, \mathcal{D}^-$	The positive and negative mask differences respectively.
$\mathcal{A}^+, \mathcal{A}^-$	The positive and negative mask differences aligned with the target frame t_i.
n_r, n_c	The normalized temporal distance between the target frame t_i and the current/previously interacted frames t^r and t^c respectively.

2. Comparisons with KMN

KMN [3] presents two major ideas: 1) hide-and-seek training augmentation and 2) kernelized memory reading. We focus only on 2) here. KMN assumes that each *memory* position should only attend to a local window (specified by a Gaussian

distribution with a fixed σ) in the *query*. Thus, filtering is performed on the *query* for every *memory* position. In contrast, our top-k filtering does not assume any spatial prior and performs on the *memory* for every *query* position. Ours is discrete, leading to a efficient algorithm while KMN slows the algorithm down. Note that the two methods are not at odds with each other – we can use both at the same time to obtain higher accuracy in the `test-dev` set (see our project page) without retraining.

References

[1] Sergi Caelles, Jordi Pont-Tuset, Federico Perazzi, Alberto Montes, Kevis-Kokitsi Maninis, and Luc Van Gool. The 2019 davis challenge on vos: Unsupervised multi-object segmentation. In *arXiv:1905.00737*, 2019.

[2] Seoung Wug Oh, Joon-Young Lee, Ning Xu, and Seon Joo Kim. Video object segmentation using space-time memory networks. In *ICCV*, 2019.

[3] Hongje Seong, Junhyuk Hyun, and Euntai Kim. Kernelized memory network for video object segmentation. In *ECCV*, 2020. 1

[4] Seoung Wug Oh, Joon-Young Lee, Kalyan Sunkavalli, and Seon Joo Kim. Fast video object segmentation by reference-guided mask propagation. In *CVPR*, 2018.

[5] 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.

[6] Jonathon Luiten, Paul Voigtlaender, and Bastian Leibe. Premvos: Proposal-generation, refinement and merging for video object segmentation. In *ACCV*, 2018.

[7] Zongxin Yang, Yunchao Wei, and Yi Yang. Collaborative video object segmentation by foreground-background integration. In *ECCV*, 2020.

[8] Xiankai Lu, Wenguan Wang, Danelljan Martin, Tianfei Zhou, Jianbing Shen, and Van Gool Luc. Video object segmentation with episodic graph memory networks. In *ECCV*, 2020.

[9] Agrim Gupta, Piotr Dollar, and Ross Girshick. LVIS: A dataset for large vocabulary instance segmentation. In *CVPR*, 2019.

[10] Yi Zeng, Pingping Zhang, Jianming Zhang, Zhe Lin, and Huchuan Lu. Towards high-resolution salient object detection. In *ICCV*, 2019.

[11] Jianping Shi, Qiong Yan, Li Xu, and Jiaya Jia. Hierarchical image saliency detection on extended cssd. In *TPAMI*, 2015.

[12] Xiang Li, Tianhan Wei, Yau Pun Chen, Yu-Wing Tai, and Chi-Keung Tang. Fss-1000: A 1000-class dataset for few-shot segmentation. In *CVPR*, 2020.

[13] Ho Kei Cheng, Jihoon Chung, Yu-Wing Tai, and Chi-Keung Tang. Cascadepsp: Toward class-agnostic and very high-resolution segmentation via global and local refinement. In *CVPR*, 2020.

[14] Lijun Wang, Huchuan Lu, Yifan Wang, Mengyang Feng, Dong Wang, Baocai Yin, and Xiang Ruan. Learning to detect salient objects with image-level supervision. In *CVPR*, 2017.

[15] Ning Xu, Linjie Yang, Yuchen Fan, Dingcheng Yue, Yuchen Liang, Jianchao Yang, and Thomas Huang. Youtube-vos: A large-scale video object segmentation benchmark. In *ECCV*, 2018.