Look Before You Match: Instance Understanding Matters in Video Object Segmentation

Supplementary Material

A. Additional Experimental Results

A.1. Multi-scale Inference

Multi-scale evaluation is a commonly used trick in segmentation tasks [1, 2, 4] to boost the performance by merging the results of inputs under different data augmentations. Here we follow XMem [2] to apply image scaling and vertical mirroring and simply average the output probabilities to obtain the final masks.

Method	MS	D16 val			D17 val		
		$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
CFBI [13]	✓	90.7	89.6	91.7	83.3	80.5	86.0
XMem [2]	✓	92.7	92.0	93.5	88.2	85.4	91.0
Ours	✗	92.6	91.5	93.7	87.1	83.7	90.5
Ours	✓	92.9	92.2	93.6	88.6	85.8	91.4
Ours*	✗	92.8	91.8	93.8	88.2	84.5	91.9
Ours*	✓	93.4	92.5	94.2	89.8	86.7	93.0

Table 1. Results on DAVIS 2017 validation and YouTube-VOS validation split with different training data. D: DAVIS 2017, Y: YouTube 2019, S: static images, B: BL30K. ‡ denotes pretraining on the combined DAVIS and YouTube-VOS data.

The results in Table 1 imply that multi-scale inference improves the performance of ISVOS by 0.3% and 1.7% in terms of $\mathcal{J}\&\mathcal{F}$ on DAVIS 2016 / 2017 validation split, and ISVOS still outperforms existing methods.

A.2. Results with different training data

In the main experiment, we follow previous methods [2–4, 8] to first pretrain our model on static images (and BL30K optionally) for fair comparisons. To study the effects of pretraining on the final segmentation results, we additionally conduct experiments to train ISVOS on DAVIS 2017 [9] only, YouTube-VOS 2019 [12] only, and a mix of both. The comparison with existing models are shown in Table 2. We can see that ISVOS achieves competitive results even without incorporating static images and BL30K for pretraining, outperforming all the baseline models by a large margin. When gradually increasing the scale of the training data, the performance of our method can be further boosted.

Method	DAVIS17 val			YT2018 val				
	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{G}$	$\mathcal{J}_s$	$\mathcal{F}_s$	$\mathcal{J}_u$	$\mathcal{F}_u$
SST‡ [5]	82.5	79.9	85.1	81.7	81.2	-	76.0	-
CFBI+‡ [15]	82.9	80.1	85.7	82.0	81.2	86.0	76.2	84.6
JOINT‡ [7]	83.5	80.8	86.2	83.1	81.5	85.9	78.7	86.5
XMem‡	84.5	-	-	84.3	-	-	-	-
Ours‡	85.2	82.1	88.3	84.7	84.5	89.1	78.2	87.0
D only	77.5	75.6	79.4	-	-	-	-	-
Y only	-	-	-	84.9	84.0	88.8	78.8	88.0
S + D + Y	87.1	83.7	90.5	86.3	85.5	90.2	80.5	88.8
S + D + B + Y	88.2	84.5	91.9	86.7	86.1	90.8	81.0	89.0

Table 2. Results on DAVIS 2017 validation and YouTube-VOS validation split with different training data. D: DAVIS 2017, Y: YouTube 2019, S: static images, B: BL30K. ‡ denotes pretraining on the combined DAVIS and YouTube-VOS data (*i.e.*, D + Y).

A.3. Results on Long video datasets

In order to further evaluate the long-term performance of ISVOS, we additionally test our method on the Long-time Video dataset [6], which contains three videos with more than 7,000 frames in total for validation. Considering the video duration is longer and the target object(s) will undergo distinct appearance deformation or scale variations, we set the maximum memory size to 64 during inference. The comparison results are shown in Table 3.

Method	Long-time Video		
	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
RMNet [11]	59.8	59.7	60.0
JOINT [7]	67.1	64.5	69.6
STM [8]	80.6	79.9	81.3
HMMN [10]	81.5	79.9	83.0
STCN [4]	87.3	85.4	89.2
AOT [14]	84.3	83.2	85.4
AFB-URR [6]	83.7	82.9	84.5
XMem [2]	89.8	88.0	91.6
Ours	**90.0**	**88.3**	**91.7**

Table 3. Results on the Long-time Video dataset [6].

We can observe that ISVOS again achieves the best segmentation results measured in different metrics. It is worth mentioning that ISVOS beats the methods specifically designed for long videos, *e.g.*, AFB-URR [6] and XMem [2]. We believe the performance gain is resulted from taking advantage of the instance information in query frame to facilitate the semantic matching.

B. More Visualizations

We show the predicted segmentation masks of ISVOS on DAVIS 2017 val, YouTube-VOS 2018 val, and Long-time Video dataset in Figure 1, Figure 2, Figure 3, respectively. For the short video datasets, *i.e.*, DAVIS and YouTube-VOS, the time interval is 5, while for the long video dataset, *i.e.*, Long-time Video dataset, the time interval is 1 since it is sparsely annotated. We can see that our method could generate accurate masks even for the objects with remarkable appearance variations.

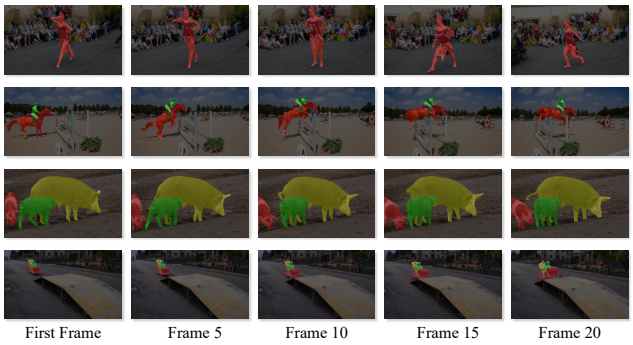

Figure 1. Segmentation results on DAVIS 2017 val split.

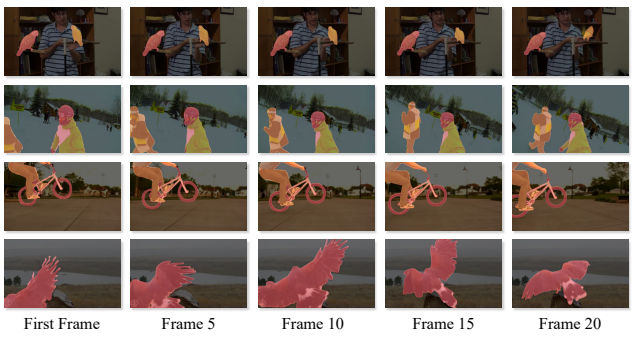

Figure 2. Segmentation results on YouTube-VOS 2018 val split.

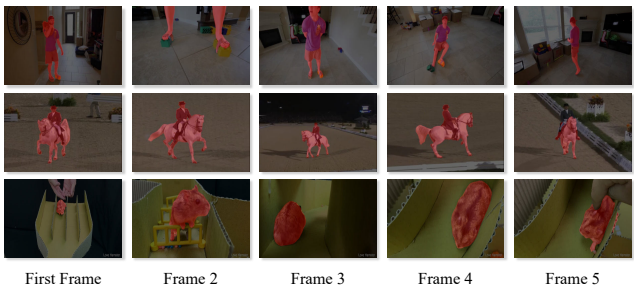

Figure 3. Segmentation results on Long-time Video dataset.

References

[1] Siddhartha Chandra and Iasonas Kokkinos. Fast, exact and multi-scale inference for semantic image segmentation with deep gaussian crfs. In *ECCV*, 2016. 1

[2] Ho Kei Cheng and Alexander G. Schwing. XMem: Long-term video object segmentation with an atkinson-shiffrin memory model. In *ECCV*, 2022. 1, 2

[3] Ho Kei Cheng, Yu-Wing Tai, and Chi-Keung Tang. Modular interactive video object segmentation: Interaction-to-mask, propagation and difference-aware fusion. In *CVPR*, 2021. 1

[4] Ho Kei Cheng, Yu-Wing Tai, and Chi-Keung Tang. Rethinking space-time networks with improved memory coverage for efficient video object segmentation. In *NeurIPS*, 2021. 1

[5] Brendan Duke, Abdalla Ahmed, Christian Wolf, Parham Aarabi, and Graham W Taylor. Sstvos: Sparse spatiotemporal transformers for video object segmentation. In *CVPR*, 2021. 1

[6] Yongqing Liang, Xin Li, Navid Jafari, and Jim Chen. Video object segmentation with adaptive feature bank and uncertain-region refinement. In *NeurIPS*, 2020. 1, 2

[7] Yunyao Mao, Ning Wang, Wengang Zhou, and Houqiang Li. Joint inductive and transductive learning for video object segmentation. In *ICCV*, 2021. 1

[8] Seoung Wug Oh, Joon-Young Lee, Ning Xu, and Seon Joo Kim. Video object segmentation using space-time memory networks. In *ICCV*, 2019. 1

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

[10] Jianping Shi, Qiong Yan, Li Xu, and Jiaya Jia. Hierarchical image saliency detection on extended cssd. *TPAMI*, 2015. 1

[11] Haozhe Xie, Hongxun Yao, Shangchen Zhou, Shengping Zhang, and Wenxiu Sun. Efficient regional memory network for video object segmentation. In *CVPR*, 2021. 1

[12] 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. 1

[13] Zongxin Yang, Yunchao Wei, and Yi Yang. Collaborative video object segmentation by foreground-background integration. In *ECCV*, 2020. 1

[14] Zongxin Yang, Yunchao Wei, and Yi Yang. Associating objects with transformers for video object segmentation. In *NeurIPS*, 2021. 1

[15] Zongxin Yang, Yunchao Wei, and Yi Yang. Collaborative video object segmentation by multi-scale foreground-background integration. *TPAMI*, 2021. 1