Video Object Segmentation-aware Video Frame Interpolation
-Supplementary Material-

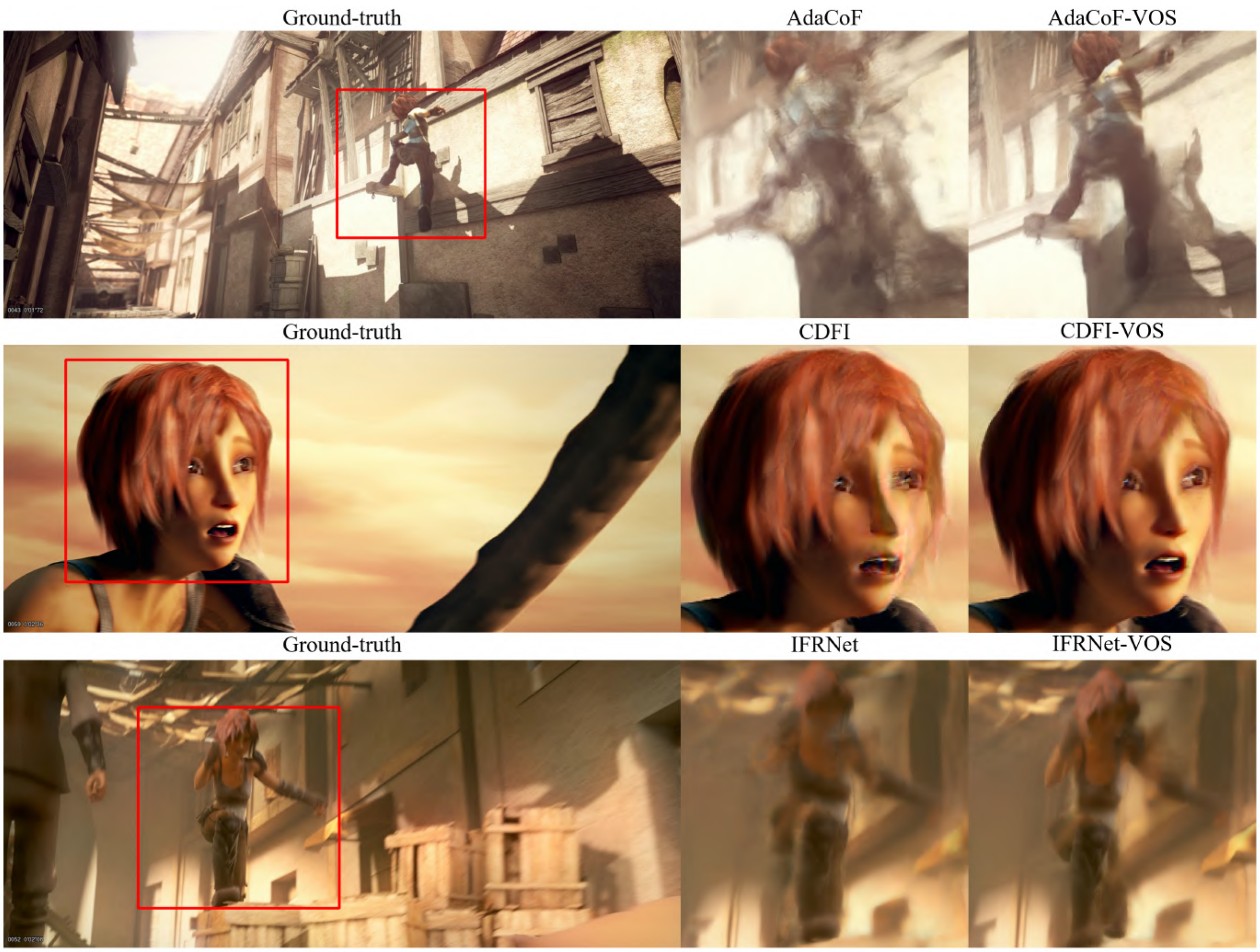

Figure S-1. Examples of 4× VFI results obtained by the baseline models, AdaCoF [5], CDFI [1], and IFRNet [4], trained with and without the proposed VOS-VFI.

A. VFI with a scale factor of 4.

The proposed VOS-VFI can be applied to any VFI model. Consequently, arbitrary-rate VFI is feasible if VOS-VFI is applied to VFI models supporting such a capability. However, among the three representative baseline models chosen in our experiments, AdaCoF [5] and CDFI [1] do not support arbitrary-rate VFI. In addition, IFRNet [4] requires dedicated weight parameters for multi-frame interpolation. Therefore, to evaluate the performance of VOS-VFI on another scale factor in addition to the standard scale factor 2, we performed $2\times$ VFI twice to have $4\times$ VFI results, which is a common strategy [3, 7]. Table S-1 shows the quantitative performance evaluation results on the DAVIS 2016 and 2017 datasets. The results showed a similar tendency as those obtained for the scale factor of 2. Specifically, the

Model	DAVIS 2016					DAVIS 2017				
	PSNR$_\uparrow$	†PSNR$_\uparrow$	$J\&F_\uparrow$	$J_\uparrow$	$F_\uparrow$	PSNR$_\uparrow$	†PSNR$_\uparrow$	$J\&F_\uparrow$	$J_\uparrow$	$F_\uparrow$
AdaCoF [5]	**21.03**	26.72	73.2	73.1	73.2	**22.15**	26.63	69.6	67.3	71.8
AdaCoF-VOS	20.96	**26.81**	**77.8**	**77.6**	**78.0**	22.13	**26.71**	**73.6**	**71.0**	**76.2**
CDFI [1]	21.01	26.77	78.6	77.8	79.4	21.98	26.69	75.2	72.1	78.4
CDFI-VOS	**21.05**	**26.88**	**80.0**	**79.2**	**80.8**	**22.03**	**26.80**	**76.1**	**72.9**	**79.3**
IFRNet [4]	21.32	26.83	78.7	78.6	78.7	22.33	26.76	75.2	72.5	78.0
IFRNet-VOS	**21.36**	**26.94**	**79.3**	**79.1**	**79.5**	**22.37**	**26.82**	**76.0**	**73.1**	**78.9**

Table S-1. Quantitative results of the three baseline models trained with and without VOS-VFI for 4× interpolation. The performance is evaluated in terms of the image quality PSNR and segmentation accuracy ($J\&F$, J, and F) on the DAVIS 2016 and 2017 datasets. † represents PSNR scores on the foreground objects obtained by masking out the background using the ground-truth segmentation maps.

proposed VOS-VFI improved the segmentation accuracy of the baseline models by 4.6%, 1.4%, and 0.6% for AdaCoF, CDFI, and IFRNet for DAVIS 2016, respectively, and 4.0%, 0.9%, and 0.8% for AdaCoF, CDFI, and IFRNet for DAVIS 2017, respectively, in terms of $J\&F$.

Fig. S-1 shows several results obtained for 4× VFI on the HD dataset. As can be seen, the proposed VOS-VFI contributed to the baseline models by improving the image quality of interpolated frames.

B. More results

First, we experimented the method that uses the correspondence-wise loss (CoRR) [2] for achieving better visual quality, which can be adopted to any other VFI models as our approach. However, as shown in Table S-2, we could not obtain performance improvements using CoRR.

Next, we applied VOS-VFI to a more recent transformer-based VFI baseline [8]. VOS-VFI also introduced non-marginal improvements, especially for VOS metrics, on this latest baseline. Our source code can be found in our project page[1]. We expect that the proposed VOS-VFI training framework can be applied to upcoming VFI models to boost performance without increasing their number of parameters and inference time.

Lastly, Table S-3 provides the NIQE/PI/NIMA scores obtained from four datasets for clear performance comparisons. As can be seen, the proposed method improved these perceptual metrics on all datasets. Table S-4 provides the PSNR scores on the four datasets, where the foreground PSNRs were only measured for the DAVIS datasets using the ground-truth segmentation maps. Although the proposed VOS-VFI showed some improvements in foreground object synthesis, we could not obtain consistent performance improvements in terms of the PSNR. Note that the PSNR is not correlated with the human perceptual quality of interpolated frames [6], and VOS-VFI showed effectiveness in the perceptual quality metrics, VOS performance (Table 1), video tracking performance (Table 3), object pose estimation performance (Table 5), and user studies (Section 4.2.1). Check our project page for more results.

[1] https://github.com/junsang7777/VOS-VFI

Model	DAVIS 2016				
	PSNR	†PSNR	$J\&F_\uparrow$	$J_\uparrow$	$F_\uparrow$
AdaCoF	**25.11**	25.62	85.9	84.9	86.8
AdaCoF-VOS	25.03	**25.72**	**87.0**	**85.9**	**88.2**
AdaCoF-CoRR [2]	24.87	25.66	85.1	84.2	86.0
EMA [8]	**27.24**	26.03	88.0	86.7	89.3
EMA-VOS	27.19	**26.06**	**88.8**	**87.6**	**90.1**
EMA [8]-CoRR [2]	27.10	25.80	86.9	85.9	88.0

Table S-2. Evaluation using the additional methods [2, 8]. † represents a foreground PSNR.

Model	DAVIS 2016			DAVIS 2017			Vimeo90K (val)			UCF101 (val)		
	NIQE$_\downarrow$	PI$_\downarrow$	NIMA$_\uparrow$	NIQE$_\downarrow$	PI$_\downarrow$	NIMA$_\uparrow$	NIQE$_\downarrow$	PI$_\downarrow$	NIMA$_\uparrow$	NIQE$_\downarrow$	PI$_\downarrow$	NIMA$_\uparrow$
AdaCoF [5]	3.443	3.338	4.565	3.545	3.402	4.502	5.180	4.104	4.765	7.272	5.695	4.018
AdaCoF-VOS	**3.431**	**3.329**	**4.586**	**3.534**	**3.400**	**4.505**	**5.153**	**4.094**	**4.771**	**7.237**	**5.678**	**4.026**
CDFI [1]	3.081	2.845	4.568	3.267	3.015	4.485	4.933	3.832	4.873	6.878	5.421	3.987
CDFI-VOS	**3.067**	**2.658**	**4.671**	**3.254**	**3.002**	**4.492**	**4.910**	**3.822**	**4.879**	**6.875**	**5.408**	**3.991**
IFRNet [4]	3.534	3.304	4.407	3.668	3.494	4.351	5.062	3.969	4.820	7.191	5.665	**4.023**
IFRNet-VOS	**3.519**	**3.294**	**4.416**	**3.651**	**3.483**	**4.361**	**5.021**	**3.935**	**4.824**	**7.115**	**5.617**	4.020

Table S-3. Evaluation in terms of the three representative perceptual quality metrics.

Model	DAVIS 2016		DAVIS 2017		Vimeo90K (val)	UCF101 (val)
	PSNR$_\uparrow$	†PSNR$_\uparrow$	PSNR$_\uparrow$	†PSNR$_\uparrow$	PSNR$_\uparrow$	PSNR$_\uparrow$
AdaCoF [5]	**25.11**	25.62	**26.23**	26.13	**34.34**	35.16
AdaCoF-VOS	25.03	**25.72**	26.21	**26.22**	34.26	35.16
CDFI [1]	25.68	25.74	26.71	26.24	35.17	35.21
CDFI-VOS	**25.75**	**25.80**	**26.79**	**26.30**	**35.28**	**35.25**
IFRNet [4]	26.70	25.91	27.57	26.44	35.73	35.26
IFRNet-VOS	**26.74**	**25.98**	**27.60**	**26.50**	**35.80**	**35.28**

Table S-4. Evaluation in terms of the PSNR. † represents a foreground PSNR.

References

[1] Tianyu Ding, Luming Liang, Zhihui Zhu, and Ilya Zharkov. CDFI: Compression-driven network design for frame interpolation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8001–8011, 2021.

[2] Daniel Geng, Max Hamilton, and Andrew Owens. Comparing correspondences: Video prediction with correspondence-wise losses. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3365–3376, 2022.

[3] Zhewei Huang, Tianyuan Zhang, Wen Heng, Boxin Shi, and Shuchang Zhou. Real-time intermediate flow estimation for video frame interpolation. In *Proceedings of the European Conference on Computer Vision*, pages 624–642, 2022.

[4] Lingtong Kong, Boyuan Jiang, Donghao Luo, Wenqing Chu, Xiaoming Huang, Ying Tai, Chengjie Wang, and Jie Yang. IFRNet: Intermediate feature refine network for efficient frame interpolation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1969–1978, 2022.

[5] Hyeongmin Lee, Taeoh Kim, Tae-young Chung, Daehyun Pak, Yuseok Ban, and Sangyoun Lee. AdaCoF: Adaptive collaboration of flows for video frame interpolation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5316–5325, 2020.

[6] Hui Men, Hanhe Lin, Vlad Hosu, Daniel Maurer, Andres Bruhn, and Dietmar Saupe. Technical report on visual quality assessment for frame interpolation. *arXiv preprint arXiv:1901.05362*, 2019.

[7] Junheum Park, Keunsoo Ko, Chul Lee, and Chang-Su Kim. BMBC: Bilateral motion estimation with bilateral cost volume for video interpolation. In *Proceedings of the European Conference on Computer Vision*, pages 109–125, 2020.

[8] Guozhen Zhang, Yuhan Zhu, Haonan Wang, Youxin Chen, Gangshan Wu, and Limin Wang. Extracting motion and appearance via inter-frame attention for efficient video frame interpolation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5682–5692, 2023.