Revisiting Adaptive Convolutions for Video Frame Interpolation
— Supplementary Material —

Simon Niklaus
Adobe Research

Long Mai
Adobe Research

Oliver Wang
Adobe Research

1. Middlebury Benchmark

A screenshot of the interpolation error category from the Middlebury benchmark for optical flow [1] is shown Figure 1, where our SepConv++ currently ranks second among all published methods. Please note that DCM is an unpublished paper that, as per the title, distills a cheating model.

2. Structural Similarity

We focus on PSNR in our main paper since SSIM [4] is subject to unexpected and unintuitive results [3]. However, we provide relevant results with SSIM instead of PSNR in Tables 1–3. These results support our claims and are generally aligned with PSNR in terms of relative improvements.

References

[1] Simon Baker, Daniel Scharstein, J. P. Lewis, Stefan Roth, Michael J. Black, and Richard Szeliski. A Database and Evaluation Methodology for Optical Flow. *International Journal of Computer Vision*, 92(1):1–31, 2011. 1, 2

[2] Ziwei Liu, Raymond A. Yeh, Xiaoou Tang, Yiming Liu, and Aseem Agarwala. Video Frame Synthesis Using Deep Voxel Flow. In *IEEE International Conference on Computer Vision*, 2017. 2

[3] Jim Nilsson and Tomas Akenine-Möller. Understanding SSIM. *arXiv/2006.13846*, 2020. 1

[4] Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image Quality Assessment: From Error Visibility to Structural Similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. 1

[5] Tianfan Xue, Baian Chen, Jiajun Wu, Donglai Wei, and William T. Freeman. Video Enhancement With Task-Oriented Flow. *International Journal of Computer Vision*, 127(8):1106–1125, 2019. 2

Average interpolation error	avg. rank	Mequon (Hidden texture) im0 GT im1 all	Mequon disc	Mequon untext	Schefflera (Hidden texture) im0 GT im1 all	Schefflera disc	Schefflera untext	Urban (Synthetic) im0 GT im1 all	Urban disc	Urban untext	Teddy (Stereo) im0 GT im1 all	Teddy disc	Teddy untext	Backyard (High-speed camera) im0 GT im1 all	Backyard disc	Backyard untext	Basketball (High-speed camera) im0 GT im1 all	Basketball disc	Basketball untext	Dumptruck (High-speed camera) im0 GT im1 all	Dumptruck disc	Dumptruck untext	Evergreen (High-speed camera) im0 GT im1 all	Evergreen disc	Evergreen untext
SoftSplat [169]	3.4	**2.06** 1	**3.06** 1	1.14 6	**2.80** 1	3.91 3	**1.24** 1	1.99 2	2.73 2	1.21 4	**3.84** 1	**4.64** 1	**2.69** 1	8.10 13	10.0 13	**2.96** 1	4.10 4	7.53 4	1.98 3	5.49 4	12.1 4	1.39 4	5.40 2	8.33 2	1.50 4
DCM [185]	8.5	2.16 3	3.30 2	1.16 8	2.82 2	**3.83** 1	1.27 9	**1.89** 1	**2.58** 1	**1.10** 1	4.09 3	4.98 3	2.81 3	8.59 16	10.7 16	3.09 13	4.87 17	9.18 17	2.00 6	6.17 14	13.5 15	1.44 12	6.14 9	9.45 9	1.65 22
SepConv++ [186]	10.7	2.39 18	4.17 20	1.20 21	2.98 5	4.21 6	1.28 16	3.34 20	3.23 5	2.20 83	4.49 10	5.81 13	2.87 5	7.64 4	9.42 4	2.97 2	3.77 2	6.80 2	**1.96** 1	5.26 3	11.6 3	**1.36** 1	5.71 6	8.86 6	**1.45** 1
FGME [158]	10.8	2.08 2	3.34 4	**0.98** 1	3.32 18	4.43 10	1.63 109	2.46 4	3.28 6	1.41 14	4.08 2	4.85 2	3.05 15	**7.36** 1	**9.08** 1	3.03 6	4.17 6	7.62 6	2.06 17	4.95 2	10.7 2	1.44 12	5.45 3	8.41 3	1.57 13
BMBC [172]	11.9	2.30 11	3.40 6	1.20 21	3.07 6	4.25 7	1.41 56	3.17 16	4.19 26	1.66 35	4.24 5	5.28 5	2.85 4	7.79 5	9.62 5	3.14 19	4.08 3	7.47 3	2.02 8	5.63 6	12.4 6	1.40 6	5.55 4	8.58 4	1.61 19
EAFI [171]	12.0	2.22 5	3.69 9	1.15 7	3.16 9	4.44 11	1.50 83	2.12 3	2.96 3	1.15 2	4.28 7	5.30 6	2.92 6	8.71 19	10.8 17	3.03 6	4.97 19	9.45 19	1.99 4	5.94 12	13.3 13	1.38 3	5.84 7	9.03 7	1.55 11
STAR-Net [164]	14.0	2.18 4	3.37 5	1.21 39	3.46 27	4.88 26	1.47 75	3.04 14	3.53 11	1.58 28	4.41 9	5.44 9	2.76 2	7.51 2	9.27 2	2.98 4	4.65 9	8.72 9	1.99 4	6.21 15	13.4 14	1.41 7	6.17 10	9.45 9	1.49 3
EDSC [174]	15.2	2.32 15	3.90 14	1.16 8	3.10 7	4.38 9	1.51 85	2.98 12	3.54 12	1.36 12	4.49 10	5.74 11	3.16 28	8.05 12	9.96 12	3.08 11	4.89 18	9.28 18	2.02 8	5.55 5	12.3 5	1.41 7	6.42 17	9.99 18	1.55 11
AdaCoF [165]	19.7	2.41 20	4.10 19	1.26 129	3.10 7	4.32 8	1.43 62	3.48 24	3.31 7	1.78 51	4.84 18	5.94 19	2.93 7	8.68 18	10.8 17	3.14 19	4.13 5	7.59 5	1.97 2	5.77 10	12.9 11	1.37 2	5.60 5	8.67 5	1.48 2
DSepConv [162]	23.3	2.47 21	4.39 26	1.21 39	3.32 18	4.60 18	1.72 128	3.28 17	3.66 13	1.50 21	5.11 25	6.36 23	3.23 62	7.85 6	9.69 6	3.11 16	4.68 11	8.78 11	2.04 15	5.65 7	12.5 7	1.44 12	6.54 21	10.2 21	1.58 16

Figure 1: Screenshot of our IE-ranking in the Middlebury benchmark (taken on the 25th of September, currently private).

		Middlebury Baker *et al.* [1]		Vimeo-90k Xue *et al.* [5]		UCF101 - DVF Liu *et al.* [2]		Xiph - 1K (4K resized to 1K)		Xiph - 2K (4K resized to 2K)	
	training dataset	SSIM ↑	relative improvement	SSIM ↑	relative improvement	SSIM ↑	relative improvement	SSIM ↑	relative improvement	SSIM ↑	relative improvement
original SepConv	prorietary	0.959	–	0.956	–	0.947	–	0.959	–	0.929	–
reimplementation	Vimeo-90k	0.958	–	0.953	–	0.949	–	0.959	–	0.925	–
+ delayed padding	— ,, —	0.959	+ 0.001	0.956	+ 0.003	0.950	+ 0.001	0.960	+ 0.001	0.926	+ 0.001
+ input normalization	— ,, —	0.962	+ 0.003	0.957	+ 0.001	0.950	+ 0.000	0.955	- 0.005	0.925	- 0.001
+ improved network	— ,, —	0.963	+ 0.001	0.957	+ 0.000	0.950	+ 0.000	0.962	+ 0.007	0.930	+ 0.005
+ normalized kernels	— ,, —	0.967	+ 0.004	0.960	+ 0.003	0.950	+ 0.000	0.963	+ 0.001	0.932	+ 0.002
+ contextual training	— ,, —	0.968	+ 0.001	0.961	+ 0.001	0.950	+ 0.000	0.964	+ 0.001	0.933	+ 0.001
+ self-ensembling	— ,, —	0.969	+ 0.001	0.962	+ 0.001	0.951	+ 0.001	0.964	+ 0.000	0.934	+ 0.001

Table 1: Ablation experiments to quantitatively analyze the effects of our proposed techniques. In short, they each positively affect the interpolation quality across different dataset as long as the inter-frame motion does not exceed the kernel size.

		Middlebury Baker *et al.* [1]		Vimeo-90k Xue *et al.* [5]		UCF101 - DVF Liu *et al.* [2]		Xiph - 1K (4K resized to 1K)		Xiph - 2K (4K resized to 2K)	
	training dataset	SSIM ↑	relative improvement	SSIM ↑	relative improvement	SSIM ↑	relative improvement	SSIM ↑	relative improvement	SSIM ↑	relative improvement
SepConv - $\mathcal{L}_1$	prorietary	0.959	–	0.956	–	0.947	–	0.959	–	0.929	–
Ours - $\mathcal{L}_{Ctx}$	Vimeo-90k	0.968	+ 0.009	0.961	+ 0.005	0.950	+ 0.003	0.964	+ 0.005	0.933	+ 0.004
Ours - $\mathcal{L}_{Ctx}$ - 8×	— ,, —	0.969	+ 0.001	0.962	+ 0.001	0.951	+ 0.001	0.964	+ 0.000	0.934	+ 0.001

Table 2: Quantitative comparison with SepConv. We list two separate results of our proposed approach, one without and one with self-ensembling. The self-ensembling is denoted by 8× as it represents a combination of eight independent estimates.

		Middlebury Baker *et al.* [1]		Vimeo-90k Xue *et al.* [5]		UCF101 - DVF Liu *et al.* [2]		Xiph - 1K (4K resized to 1K)		Xiph - 2K (4K resized to 2K)	
	venue	SSIM ↑	absolute rank	SSIM ↑	absolute rank	SSIM ↑	absolute rank	SSIM ↑	absolute rank	SSIM ↑	absolute rank
SepConv - $\mathcal{L}_1$	ICCV 2017	0.959	8th of 10	0.956	9th of 10	0.947	10th of 10	0.959	10th of 10	0.929	7th of 10
CtxSyn - $\mathcal{L}_{Lap}$	CVPR 2018	0.964	7th of 10	0.961	5th of 10	0.949	9th of 10	0.963	5th of 10	0.936	3rd of 10
DAIN	CVPR 2019	0.965	5th of 10	0.964	2nd of 10	0.950	3rd of 10	0.965	2nd of 10	0.939	2nd of 10
CAIN	AAAI 2020	0.951	10th of 10	0.959	8th of 10	0.950	3rd of 10	0.961	8th of 10	0.936	3rd of 10
EDSC - $\mathcal{L}_C$	arXiv 2020	0.967	4th of 10	0.961	5th of 10	0.950	3rd of 10	0.963	5th of 10	OOM	OOM
AdaCoF	CVPR 2020	0.959	8th of 10	0.956	9th of 10	0.950	3rd of 10	0.960	9th of 10	0.927	8th of 10
SoftSplat - $\mathcal{L}_{Lap}$	CVPR 2020	0.971	1st of 10	0.970	1st of 10	0.952	1st of 10	0.969	1st of 10	0.944	1st of 10
BMBC	ECCV 2020	0.965	5th of 10	0.964	2nd of 10	0.950	3rd of 10	0.963	5th of 10	OOM	OOM
Ours - $\mathcal{L}_{Ctx}$	N/A	0.968	3rd of 10	0.961	5th of 10	0.950	3rd of 10	0.964	3rd of 10	0.933	6th of 10
Ours - $\mathcal{L}_{Ctx}$ - 8×	— ,, —	0.969	2nd of 10	0.962	4th of 10	0.951	2nd of 10	0.964	3rd of 10	0.934	5th of 10

Table 3: Quantitative comparison with recent approaches for video frame interpolation. In addition to highlighting the best result by underlining it, we emphasize the second-best result via a dotted underline. Note that some methods were unable to run on 2K footage due to exceeding the 16 gigabytes of memory available on our graphics card (denoted as "OOM").