ST-MFNet: A Spatio-Temporal Multi-Flow Network for Frame Interpolation Supplementary Material

Duolikun Danier Fan Zhang David Bull
University of Bristol
{duolikun.danier, fan.zhang, dave.bull}@bristol.ac.uk

The Supplementary Material is organized as follows. Section A presents the the discriminator architecture used for training ST-MFNet. Section B provides additional ablation study results. The visualization of the multi-scale multi-flows is given in Section C. Section D presents the additional quantitative evaluation results in terms of LPIPS [28]. Results on multi-frame interpolation are provided in Section E. An additional experiment to validate the model design is described in Section F. The experimental configuration of the user study is described in Section G. Section H provides a link to a supplementary demo video. Finally, Section I summarizes the license information for all data and code assets used in this paper.

A. Discriminator for ST-MFNet

The architecture of the discriminator employed in this work is illustrated in Figure 1; this was originally designed to train ST-GAN [27] for texture synthesis. It contains a temporal and a spatial branch. The former takes the differences between the interpolated output I_t^{out} (where $t = 1.5$) of ST-MFNet and its two adjacent original frames I_1, I_2 as input. The differences here represent the high-frequency temporal information within these three frames. The spatial branch in this network processes the ST-MFNet output I_t^{out} to generate spatial features. Finally, the temporal and spatial features generated in these two branches are concatenated before fed into the final fully connected layers.

B. Additional Ablation Study Results

In the main paper, we presented key ablation study results where the primary contributions in the proposed ST-MFNet are evaluated. Here the effectiveness of the up-sampling scale is further investigated, which has been employed during the multi-flow prediction in the MIFNet branch (see Section 3.1 of the main paper). In addition, we present the quantitative ablation study results for the ST-GAN in terms of a perceptually-oriented metric, the Learned Perceptual Image Patch Similarity (LPIPS) [28].

Up-sampling. To evaluate the contribution of the up-sampling scale during the multi-flow prediction, the version of ST-MFNet (Ours-*w/o US*) with only two multi-flow estimation heads (at $l = 0, 1$ scales) were implemented. It was also trained and evaluated using the same configurations described in the main paper. Its interpolation results are summarized in Table 1 alongside more comprehensive ablation study results for the other four variants of ST-MFNet (described in the main paper). It can be observed that Ours-*w/o US* was outperformed by the full version of ST-MFNet (Ours) on all test datasets. The performance difference can also be demonstrated through visual comparison as shown in Figure 2. All of these confirm the effectiveness of the up-sampling scale in multi-flow estimation.

ST-GAN. In the main paper, due to space limitations, we only evaluated the effectiveness of the adopted ST-GAN using visual examples. Here we additionally present the quantitative ablation study results for the adopted ST-GAN. For this purpose, we evaluate the same variants of ST-MFNet as described in Section 5.1 (the ST-GAN sub-section) of the main paper, that is, the distortion-oriented version (Our-$\mathcal{L}_{lap}$), the version fine-tuned with ST-GAN (Our-$\mathcal{L}_p$), the version fine-tuned with FIGAN [12] and the version fine-tuned with TGAN [22]. Table 2 summarizes the performance of these variants on all four test sets in terms of LPIPS. It can be clearly observed from the table that the ST-GAN adopted in our work provides the best overall LPIPS performance, indicating its effectiveness for enhancing perceptual quality of the interpolated results.

C. Visualization of Motion Fields

To better understand the effectiveness of the multi-scale multi-flow estimation in the MIFNet branch, the predicted multi-flows are visualized here in the same manner as done in [12]. That is, the mean flow maps at scale l, $\bar{G}^l_{t\to n}$ (where $n = 1, 2$), are obtained using Equations (1) and (2), and shown in Figure 3. Note that for the purpose of visualization, the flows at the down- and up-sampled scales are re-scaled to the original resolution using the nearest neighbor

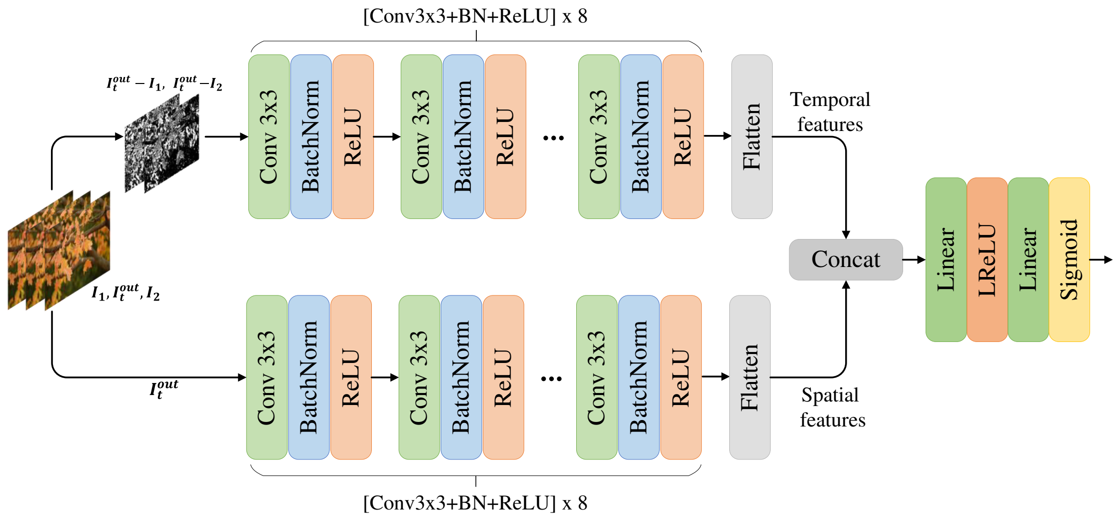


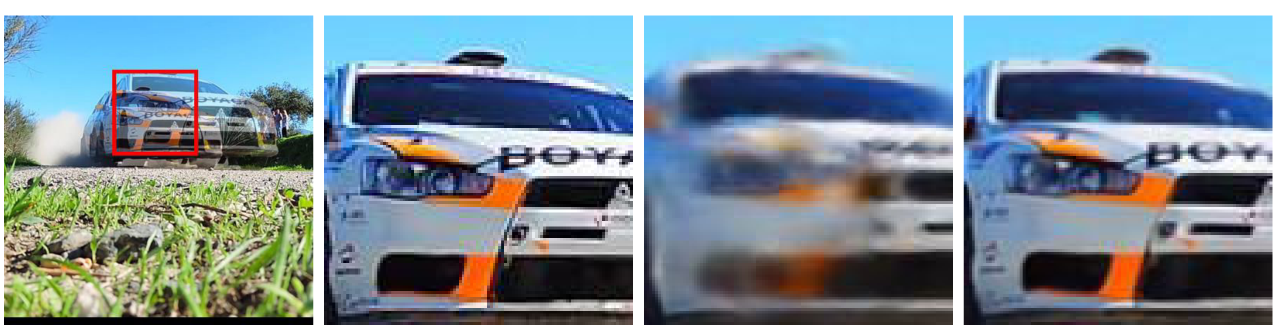
Figure 1. Architecture of the discriminator used for training ST-MFNet.

	UCF101	DAVIS	SNU-FILM				VFITex
			Easy	Medium	Hard	Extreme	
Ours-*w/o BLFNet*	33.218/0.970	27.767/0.881	40.655/0.990	36.890/0.984	31.205/0.947	25.492/0.869	28.498/0.915
Ours-*w/o MIFNet*	33.202/0.969	27.886/0.889	40.331/0.991	36.530/0.982	31.321/0.949	25.620/0.871	28.357/0.911
Ours-*w/o TENet*	32.895/0.970	27.484/0.880	40.275/0.991	35.983/0.980	30.527/0.937	25.374/0.864	28.241/0.910
Ours-*unet*	33.378/0.970	28.096/0.892	40.616/0.991	36.797/0.984	31.383/0.950	25.680/. 872	28.898/0.925
Ours-*w/o US*	33.371/0.970	28.155/0.893	40.248/0.990	36.689/0.983	31.384/0.949	25.636/0.873	28.977/0.925
Ours	33.384/0.970	28.287/0.895	40.775/0.992	37.111/0.985	31.698/0.951	25.810/0.874	29.175/0.929

Table 1. Comprehensive ablation study results on ST-MFNet.

	UCF101	DAVIS	SNU-FILM				VFITex
			Easy	Medium	Hard	Extreme	
Ours-$\mathcal{L}_{lap}$	0.036	0.125	0.019	0.036	0.073	0.148	0.216
TGAN	0.034	0.117	0.019	0.033	0.068	0.142	0.213
FIGAN	0.036	0.119	0.020	0.035	0.070	0.146	0.216
Ours-$\mathcal{L}_p$	0.033	0.116	0.017	0.031	0.065	0.140	0.210

Table 2. Quantitative ablation study results for ST-GAN, in terms of LPIPS.



(a) Overlay (b) GT (c) Ours-*w/o US* (d) Ours-*w/ US*

Figure 2. Qualitative results interpolated by the ST-MFNet with the up-sampled scale removed (Ours-*w/o US*) and the full version of ST-MFNet (Ours-*w/ US*). Here "Overlay" means the overlaid adjacent frames.

(a) Overlay (b) $\bar{G}^{l=0}_{t\rightarrow 1}$ (c) $\bar{G}^{l=0}_{t\rightarrow 2}$ (d) $\bar{G}^{l=1}_{t\rightarrow 1}$ (e) $\bar{G}^{l=1}_{t\rightarrow 2}$ (f) $\bar{G}^{l=-1}_{t\rightarrow 1}$ (g) $\bar{G}^{l=-1}_{t\rightarrow 2}$

Figure 3. Visualization of the multi-scale multi-flows predicted by the network.

filter.

$$\mathbf{g}(x,y,i) = (\boldsymbol{\alpha}(x,y,i), \boldsymbol{\beta}(x,y,i)) \tag{1}$$

$$\bar{G}^{l}_{t\rightarrow n}(x,y) = \sum_{i=1}^{N} \mathbf{w}(x,y,i)\mathbf{g}(x,y,i) \tag{2}$$

It can be observed from Figure 3 that compared to the mean flow map at the original scale ($l = 0$), the flows estimated for the down-sampled scale ($l = 1$) tend to depict the general motion coarsely in different regions. On the other hand, the flow maps at the up-sampled scale ($l = -1$) reflect more detailed motion information.

D. Comprehensive Evaluation Results

In the main paper, we presented our quantitative evaluation results of the proposed ST-MFNet and 14 competing methods in terms of PSNR and SSIM. Here, we additionally evaluate these methods in terms of LPIPS. The full results on the test sets UCF101 [24], DAVIS [21] and VFITex are summarized in Table 3, and the results on SNU-FILM [6] are shown in Table 4.

E. Results of 4× and 8× Interpolation

The performance of the proposed ST-MFNet on multi-frame interpolation task is also evaluated, and compared to three best-performing benchmark algorithms: QVI [25], FLAVR [11] and Softsplat [18]. The algorithms were applied recursively to generate all the intermediate frames. The 11 test sequences at 240 FPS in the GoPro dataset [17] were used as the test set for 4× and 8× interpolation. Table 5 summarizes the results, where it can be seen that ST-MFNet shows the best overall performance.

F. Validation of Model Design

The proposed ST-MFNet combines multi-flow and single-flow based warping methods to enhance the interpolation quality of both complex and large motions. A natural question to ask is whether the performance of the model comes from the specific model design or simply from ensembling effect. To address this question, we create an ensemble model as a baseline, which simply combines AdaCoF [12] and Softsplat [18] through arithmetic averaging. This baseline model was trained under the same configurations as ST-MFNet and compared to the latter quantitatively. The results are summarized in Table 6, where it is noted that although ensembling of AdaCoF and Softsplat does provide some benefit, the gain is marginal. This implies that the main source of the performance gain in ST-MFNet is the model design.

G. User Study

The user study was conducted in a darkened, lab-based environment. The test sequences were played on a SONY PVM-X550 display, with screen size 124.2×71.8cm. The display resolutions were configured to 1920×1080 (spatial) and 60Hz (temporal), and the viewing distance was 2.15 meters (three times the screen height) [9]. The presentation of video sequences was controlled by a Windows PC running Matlab Psychtoolbox [3]. In each trial, a pair of videos to be compared were played twice, then the participant was asked to select the video with better perceived quality through an interface developed using the Psychtoolbox. This user study and the use of human data have undergone an internal ethics review and has been approved by the Institutional Review Board.

H. Video Demo

A video containing interpolation examples generated by ST-MFNet and more visual comparisons is available via this link: https://drive.google.com/file/d/1zpE3rCQNJi4e8ADNWKbJA5wTvPl1KZSj/view?usp=sharing.

	UCF101			DAVIS			VFITex		
	PSNR (↑)	SSIM (↑)	LPIPS (↓)	PSNR (↑)	SSIM (↑)	LPIPS (↓)	PSNR (↑)	SSIM (↑)	LPIPS (↓)
DVF [13]	32.251	0.965	0.036	20.403	0.673	0.274	19.946	0.709	0.389
SuperSloMo [10]	32.547	0.968	0.028	26.523	0.866	0.119	27.914	0.911	0.217
SepConv [19]	32.524	0.968	0.035	26.441	0.853	0.169	27.635	0.907	0.230
DAIN [4]	32.524	0.968	0.030	27.086	0.873	0.117	27.314	0.909	0.212
BMBC [20]	32.729	0.969	0.032	26.835	0.869	0.125	27.337	0.904	0.220
AdaCoF [12]	32.610	0.968	0.033	26.445	0.854	0.158	27.639	0.904	0.222
FeFlow [8]	32.520	0.967	0.036	26.555	0.856	0.169	OOM	OOM	OOM
CDFI [7]	32.653	0.968	0.024	26.471	0.857	0.157	27.576	0.906	0.218
CAIN [6]	32.537	0.968	0.037	26.477	0.857	0.197	28.184	0.911	0.240
Softsplat [18]	32.835	0.969	0.037	27.582	0.881	0.116	28.813	0.924	0.221
EDSC [5]	32.677	0.969	0.033	26.968	0.860	0.142	27.641	0.904	0.222
XVFI [23]	32.224	0.966	0.038	26.565	0.863	0.125	27.759	0.909	0.218
QVI [25]	32.668	0.967	0.036	27.483	0.883	0.181	28.819	0.926	0.210
FLAVR [11]	33.389	0.971	0.035	27.450	0.873	0.190	28.487	0.915	0.233
ST-MFNet (Ours-$\mathcal{L}_{lap}$)	33.384	0.970	0.036	28.287	0.895	0.125	29.175	0.929	0.216
ST-MFNet (Ours-$\mathcal{L}_{p}$)	33.364	0.970	0.033	28.172	0.892	0.116	28.945	0.924	0.210

Table 3. Quantitative comparison results for our model and 14 tested methods on UCF101, DAVIS and VFITex, in terms of PSNR, SSIM and LPIPS. OOM denotes cases where our GPU runs out of memory for the evaluation. For each column, the best result is colored in red and the second best is colored in blue. Underlined scores denote the performance of pre-trained models rather than our re-trained versions.

	SNU-FILM											
	Easy			Medium			Hard			Extreme		
	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS	PSNR	SSIM	LPIPS
DVF [13]	27.528	0.876	0.109	24.091	0.817	0.166	21.556	0.760	0.231	19.709	0.705	0.303
SuperSloMo [10]	36.255	0.984	0.025	33.802	0.973	0.034	29.519	0.930	0.068	24.770	0.855	0.141
SepConv [19]	39.894	0.990	0.022	35.264	0.976	0.043	29.620	0.926	0.094	24.653	0.851	0.183
DAIN [4]	39.280	0.989	0.020	34.993	0.976	0.033	29.752	0.929	0.082	24.819	0.850	0.142
BMBC [20]	39.809	0.990	0.020	35.437	0.978	0.034	29.942	0.933	0.088	24.715	0.856	0.145
AdaCoF [12]	39.912	0.990	0.021	35.269	0.977	0.039	29.723	0.928	0.080	24.656	0.851	0.152
FeFlow [8]	39.591	0.990	0.022	35.014	0.977	0.041	29.466	0.928	0.090	24.607	0.852	0.182
CDFI [7]	39.881	0.990	0.019	35.224	0.977	0.036	29.660	0.929	0.081	24.645	0.854	0.163
CAIN [6]	39.890	0.990	0.021	35.630	0.978	0.037	29.998	0.931	0.097	25.060	0.857	0.203
Softsplat [18]	40.165	0.991	0.021	36.017	0.979	0.036	30.604	0.937	0.066	25.436	0.864	0.119
EDSC [5]	39.792	0.990	0.023	35.283	0.977	0.040	29.815	0.929	0.080	24.872	0.854	0.153
XVFI [23]	38.849	0.989	0.022	34.497	0.975	0.039	29.381	0.929	0.075	24.677	0.855	0.139
QVI [25]	36.648	0.985	0.019	34.637	0.978	0.032	30.614	0.947	0.066	25.426	0.866	0.140
FLAVR [11]	40.135	0.990	0.021	35.988	0.979	0.049	30.541	0.937	0.112	25.188	0.860	0.218
ST-MFNet (Ours-$\mathcal{L}_{lap}$)	40.775	0.992	0.019	37.111	0.985	0.036	31.698	0.951	0.073	25.810	0.874	0.148
ST-MFNet (Ours-$\mathcal{L}_{p}$)	40.542	0.991	0.017	36.964	0.983	0.031	31.580	0.949	0.065	25.764	0.871	0.140

Table 4. Quantitative comparison results for our model and 14 tested methods on SNU-FILM dataset, in terms of PSNR, SSIM and LPIPS. For each column, the best result is colored in red and the second best is colored in blue. Underlined scores denote the performance of pre-trained models rather than our re-trained versions.

I. Attribution of Assets

The data and code assets employed in this work and their corresponding license information are summarized in Table 7 and 8 respectively.

	GoPro-4×			GoPro-8×		
	PSNR (↑)	SSIM (↑)	LPIPS (↓)	PSNR (↑)	SSIM (↑)	LPIPS (↓)
QVI [25]	29.324	0.927	0.049	29.280	0.929	0.048
FLAVR [11]	28.911	0.914	0.110	29.512	0.922	0.101
Softsplat [18]	28.858	0.908	0.072	29.663	0.918	0.067
ST-MFNet (Ours-$\mathcal{L}_{lap}$)	29.892	0.926	0.098	30.568	0.934	0.092

Table 5. Quantitative comparison results for 4× and 8× interpolation on GoPro dataset in terms of PSNR, SSIM and LPIPS. For each column, the best result is colored in red and the second best is colored in blue.

			AdaCoF	Softsplat	AdaCoF+Softsplat	ST-MFNet (ours-$\mathcal{L}_{lap}$)
UCF101		PSNR (↑)	32.488	32.683	32.729	33.384
		SSIM (↑)	0.968	0.969	0.969	0.970
DAVIS		PSNR (↑)	26.445	27.359	27.361	28.287
		SSIM (↑)	0.854	0.878	0.878	0.895
SNU-FILM	Easy	PSNR (↑)	39.912	40.021	40.083	40.775
		SSIM (↑)	0.990	0.991	0.991	0.992
	Medium	PSNR (↑)	35.269	35.833	35.841	37.111
		SSIM (↑)	0.977	0.979	0.979	0.985
	Hard	PSNR (↑)	29.723	30.412	30.449	31.698
		SSIM (↑)	0.928	0.936	0.937	0.951
	Extreme	PSNR (↑)	24.656	25.242	25.258	25.810
		SSIM (↑)	0.851	0.862	0.864	0.874
VFITex		PSNR (↑)	27.639	28.620	28.629	29.175
		SSIM (↑)	0.904	0.922	0.923	0.929

Table 6. Quantitative evaluation results of the proposed ST-MFNet and a simple baseline that combines AdaCoF and Softsplat. For each row, the best result is colored in red and the second best is colored in blue. Note Softsplat here is trained with Charbonnier loss so that AdaCoF, Softsplat and the baseline only differ in model design.

Dataset	**Dataset URL**	**License / Terms of Use**
Vimeo-90k [26]	http://toflow.csail.mit.edu	MIT license.
BVI-DVC [14]	https://fan-aaron-zhang.github.io/BVI-DVC/	All sequences are allowed for academic research.
UCF101 [24]	https://www.crcv.ucf.edu/research/data-sets/ucf101/	No explicit license terms, but compiled and made available for research use by the University of Central Florida.
DAVIS [21]	https://davischallenge.org	BSD license.
SNU-FILM [6]	https://myungsub.github.io/CAIN/	MIT license .
Xiph [16]	https://media.xiph.org/video/derf	Sequences used are available for research use.
Mitch Martinez Free 4K Stock Footage [1]	http://mitchmartinez.com/free-4k-red-epic-stock-footage/	Sequences used are available for research use.
UVG database [15]	http://ultravideo.fi	Non-commercial Creative Commons BY-NC license.
Pexels [2]	https://www.pexels.com/videos/	All sequences are available for research use.

Table 7. License information for the datasets used in this work.

Method	**Source code URL**	**License / Teams of Use**
DVF [13]	https://github.com/liuziwei7/voxel-flow	Non-commercial research and education only.
SuperSloMo [10]	https://github.com/avinashpaliwal/Super-SloMo	MIT license.
SepConv [19]	https://github.com/sniklaus/sepconv-slomo	Academic purposes only.
DAIN [4]	https://github.com/baowenbo/DAIN	MIT license.
BMBC [20]	https://github.com/JunHeum/BMBC	MIT license.
AdaCoF [12]	https://github.com/HyeongminLEE/AdaCoF-pytorch	MIT license.
FeFlow [8]	https://github.com/CM-BF/FeatureFlow	MIT license.
CAIN [6]	https://github.com/myungsub/CAIN	MIT license.
SoftSplat [18]	https://github.com/sniklaus/softmax-splatting	Academic purposes only.
XVFI [23]	https://github.com/JihyongOh/XVFI	Research and education only.
FLAVR [11]	https://github.com/tarun005/FLAVR	Apache-2.0 License.

Table 8. License information for the code assets used in this work.

References

[1] Mitch Martinez free 4k stock footage. *http://mitchmartinez.com/free-4k-red-epic-stock-footage/*. 6

[2] Pexels videos. *https://www.pexels.com/videos/*. 6

[3] Psychtoolbox-3. https://github.com/Psychtoolbox-3/Psychtoolbox-3. 3

[4] Wenbo Bao, Wei-Sheng Lai, Chao Ma, Xiaoyun Zhang, Zhiyong Gao, and Ming-Hsuan Yang. Depth-aware video frame interpolation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3703–3712, 2019. 4, 6

[5] Xianhang Cheng and Zhenzhong Chen. Multiple video frame interpolation via enhanced deformable separable convolution. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 4

[6] Myungsub Choi, Heewon Kim, Bohyung Han, Ning Xu, and Kyoung Mu Lee. Channel attention is all you need for video frame interpolation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 10663–10671, 2020. 3, 4, 6

[7] 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. 4

[8] Shurui Gui, Chaoyue Wang, Qihua Chen, and Dacheng Tao. Featureflow: Robust video interpolation via structure-to-texture generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14004–14013, 2020. 4, 6

[9] Recommendation ITU-R BT. 500-11, methodology for the subjective assessment of the quality of television pictures,". *International Telecommunication Union, Tech. Rep*, 2002. 3

[10] Huaizu Jiang, Deqing Sun, Varun Jampani, Ming-Hsuan Yang, Erik Learned-Miller, and Jan Kautz. Super slomo: High quality estimation of multiple intermediate frames for video interpolation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9000–9008, 2018. 4, 6

[11] Tarun Kalluri, Deepak Pathak, Manmohan Chandraker, and Du Tran. Flavr: Flow-agnostic video representations for fast frame interpolation. *arXiv preprint arXiv:2012.08512*, 2020. 3, 4, 5, 6

[12] 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. 1, 3, 4, 6

[13] Ziwei Liu, Raymond A Yeh, Xiaoou Tang, Yiming Liu, and Aseem Agarwala. Video frame synthesis using deep voxel flow. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4463–4471, 2017. 4, 6

[14] Di Ma, Fan Zhang, and David Bull. BVI-DVC: A training database for deep video compression. *IEEE Transactions on Multimedia*, pages 1–1, 2021. 6

[15] Alexandre Mercat, Marko Viitanen, and Jarno Vanne. UVG dataset: 50/120fps 4k sequences for video codec analysis and development. In *Proceedings of the 11th ACM Multimedia Systems Conference*, pages 297–302, 2020. 6

[16] Chris Montgomery et al. Xiph. org video test media (derf's collection), the xiph open source community, 1994. *Online, https://media. xiph. org/video/derf*, 3. 6

[17] Seungjun Nah, Tae Hyun Kim, and Kyoung Mu Lee. Deep multi-scale convolutional neural network for dynamic scene deblurring. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3883–3891, 2017. 3

[18] Simon Niklaus and Feng Liu. Softmax splatting for video frame interpolation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5437–5446, 2020. 3, 4, 5, 6

[19] Simon Niklaus, Long Mai, and Feng Liu. Video frame interpolation via adaptive separable convolution. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 261–270, 2017. 4, 6

[20] Junheum Park, Keunsoo Ko, Chul Lee, and Chang-Su Kim. BMBC: Bilateral motion estimation with bilateral cost volume for video interpolation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16*, pages 109–125. Springer, 2020. 4, 6

[21] Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus Gross, and Alexander Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 724–732, 2016. 3, 6

[22] Masaki Saito, Eiichi Matsumoto, and Shunta Saito. Temporal generative adversarial nets with singular value clipping. In *Proceedings of the IEEE international conference on computer vision*, pages 2830–2839, 2017. 1

[23] Hyeonjun Sim, Jihyong Oh, and Munchurl Kim. XVFI: extreme video frame interpolation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2021. 4, 6

[24] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. UCF101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 3, 6

[25] Xiangyu Xu, Li Siyao, Wenxiu Sun, Qian Yin, and Ming-Hsuan Yang. Quadratic video interpolation. In *NeurIPS*, 2019. 3, 4, 5

[26] 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. 6

[27] Kun Yang, Dong Liu, Zhibo Chen, Feng Wu, and Weiping Li. Spatiotemporal generative adversarial network-based dynamic texture synthesis for surveillance video coding. *IEEE Transactions on Circuits and Systems for Video Technology*, 2021. 1

[28] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the*

IEEE conference on computer vision and pattern recognition, pages 586–595, 2018. 1