Across Scales & Across Dimensions: Temporal Super-Resolution using Deep Internal Learning

Supplementary Material

Paper ID 132

PLEASE VIEW THE ATTACHED VIDEO

1 TSR Benchmark

We curated a challenging dataset of 25 LTR (low temporal resolution) videos of very complex fast dynamic scenes, as if recorded by a 'slow' video camera (30 fps with full inter-frame exposure time). These videos contain complex scene dynamics, highly non-rigid motions, severe motion-blurs and/or severe motion-aliasing. Fig. 1 (in this document) shows a sample frame from each video, and Table 1 (in this document) summarises the lengths and frame sizes of the HTR (high temporal resolution) ground-truth videos.

The dataset was generated from 25 very different real videos recorded with high speed (mostly 240 fps) consumer cameras. Some of these videos were downloaded from the web, and some taken from the Need-for-Speed [4] benchmark. The LTR videos were generated from these HTR videos by *blurring and sub-sampling them in time* by a factor of 8 (by averaging every 8 frames). This is equivalent to a 'slower' video camera recording the same dynamic scene, with full inter-frame exposure-time, at $1/8$ framerate (30 fps). These LTR videos were fed as inputs to the different algorithms.

We further split our dataset into two groups: (i) 13 extremely challenging videos, not only with severe motion blur, but also with severe motion aliasing and/or complex highly non-rigid motions (e.g., splashing water, flickering fire, etc.); (ii) 12 less challenging videos, still with sever motion blur, but mostly rigid motions. This was done in order to highlight the type of videos that benefit the most from Internal-Learning.

Table 2 (in this document) lists the average per-frame PSNR, structural similarity (SSIM), and a perceptual measure (LPIPS [5]), computed for each method on each video. To avoid boundary effects we did not include the first and last 30 frames of each sequence. We also disregarded a 20-pixel boundary around each frame when computing per-frame PSNR. This wide masking of the boundaries was done to accommodate large margin that some of the other algorithms require.

The results in Table 2 (of this document) indicate that sophisticated frame-interpolation methods (DAIN [1],NVIDIA SloMo [2]) are not adequate for the task of Temporal Super Resolution (TSR), and are significantly inferior (-1 dB) on LTR videos compared to dedicated TSR methods (Ours and Flawless [3]). Flawless and Ours provide comparable quantitative results on the dataset, even though Flawless is a pre-trained supervised

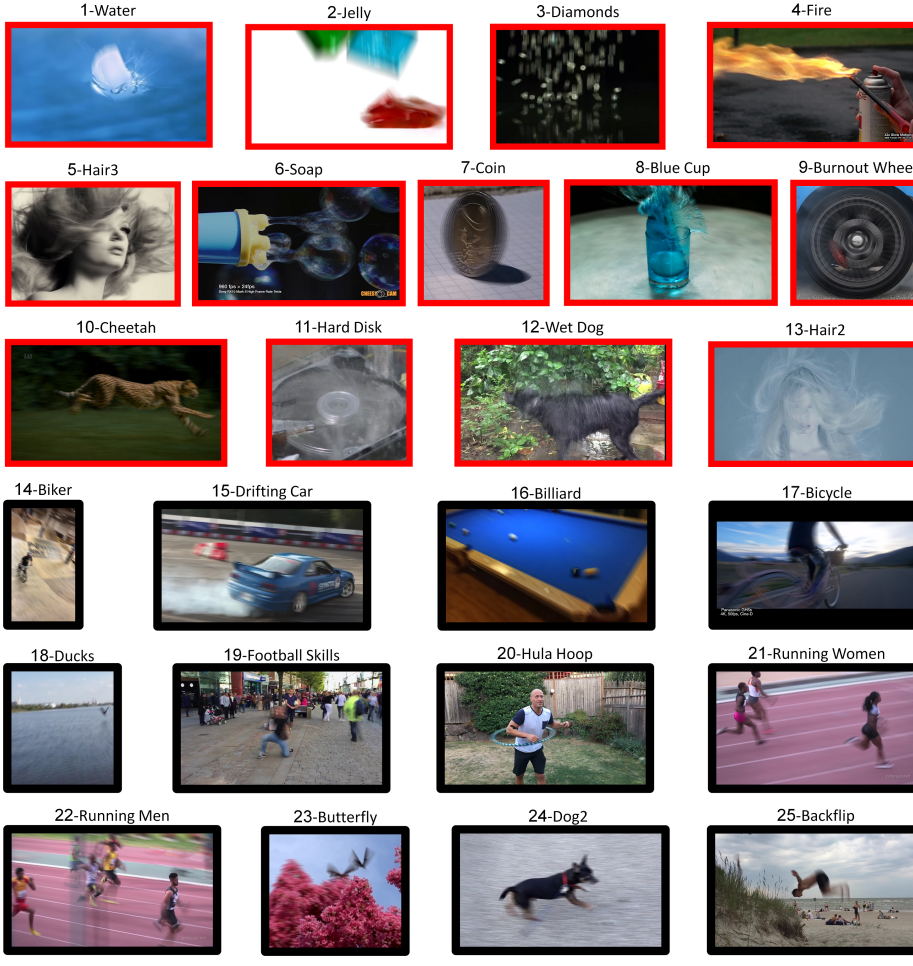

Fig. 1: **Representative frames of the TSR benchmark:** One frame from each LTR video in the benchmark. The video numbering is consistent with that used in all the tables of this document. Red frames mark the subset of more challenging videos.

#	Video Name	# Frames	Resolution	#	Video Name	# Frames	Resolution
1	Water	280	640 x 360	14	Biker	720	1280 x 720
2	Jelly	480	640 x 360	15	Drifting Car	400	1280 x 720
3	Diamonds	360	448 x 360	16	Billiard	480	1280 x 720
4	Fire	440	1280 x 720	17	Bicycle	240	1280 x 720
5	Hair3	360	704 x 512	18	Ducks	840	512 x 568
6	Soap	480	1280 x 720	19	Football Skills	440	1280 x 720
7	Coin	200	512 x 432	20	Hula Hoop	320	1280 x 720
8	Blue Cup	240	960 x 536	21	Running Women	440	1280 x 720
9	Burnout Wheel	440	544x560	22	Running Men	440	1280 x 720
10	Cheetah	240	1280 x 704	23	Butterfly	440	560 x 480
11	Hard Disk	360	704 x 720	24	Dog2	400	1280 x 720
12	Wet Dog	400	1280 x 720	25	Backflip	400	1280 x 720
13	Hair2	400	1280 x 720				

Table 1: **Details of the TSR Benchmark:** Lengths and frame sizes of the HTR videos.

method, whereas Ours is unsupervised and requires no prior training examples. Moreover, on the subset of extremely challenging videos (with highly complex non-rigid motions), our Zero-Shot TSR outperforms the state-of-the-art externally trained Flawless [3]. Since rigid motions are easier to model and capture in an external training set, Flawless provided high-quality results (better than ours) on the videos which are dominated by rigid motions. However, even in those videos, when focusing on the areas with non-rigid motions, our method visually outperforms the externally trained Flawless. While these non-rigid areas are smaller in those videos (hence have negligible effect on PSNR), they often tend to be the salient and more interesting regions in the frame. Such examples can be found in the Supplementary-Video (e.g., the billiard-ball in Video-16, the hula-hoop in Video-20, the person performing a back-flip in Video-25, etc.), as well as in *Fig.1 of the paper*.

2 Ablation Study

Table 3 (in this document) details the ablation study designed to examine the power of cross-dimension augmentations for all videos in the dataset. It compares the performance of our network when: (i) Training only on examples from *same-dimension* ('Within'); (ii) Training only on examples *across-dimensions* ('Across'); (iii) Training each video on its best configuration – 'within', 'across', or on both.

Since our atomic TSRx2 network is trained only on a coarse spatial scale of the video, we performed the ablation study at that scale (hence the differences between the numeric values in Tables 2 and 3). Table 2 indicates that, on the average, the cross-dimension augmentations are more informative than the within (same-dimension) augmentations. However, since different videos have different preferences, training each video with its best within and/or across configuration provides a small additional overall improvement. For more details see paper.

#		Ours			Flawless			DAIN			Nvidia SloMo			Linear Interp.		
		PSNR [dB]	SSIM	LPIPS	PSNR [dB]	SSIM	LPIPS	PSNR [dB]	SSIM	LPIPS	PSNR [dB]	SSIM	LPIPS	PSNR [dB]	SSIM	LPIPS
1	Water	**31.31**	**0.965**	0.098	30.90	0.962	**0.093**	30.45	0.959	0.118	30.46	0.958	0.117	29.92	0.950	0.132
2	Jelly	**26.94**	**0.937**	0.083	25.87	0.925	**0.082**	24.81	0.917	0.107	24.66	0.913	0.108	24.50	0.910	0.116
3	Diamonds	**22.70**	**0.819**	0.262	22.42	**0.819**	**0.218**	20.59	0.755	0.349	20.87	0.764	0.321	20.82	0.749	0.348
4	Fire	**29.41**	**0.974**	0.092	28.53	0.971	0.105	28.66	0.972	**0.090**	28.52	0.970	0.093	28.54	0.973	0.094
5	Hair3	**24.59**	**0.877**	**0.235**	24.00	0.871	0.252	23.77	0.863	0.280	23.50	0.853	0.266	23.53	0.861	0.276
6	Soap	**27.20**	**0.932**	0.145	26.43	0.927	**0.144**	26.39	0.922	0.162	26.10	0.920	0.160	26.43	0.924	0.164
7	Coin	**24.43**	**0.897**	0.242	24.19	**0.897**	**0.208**	22.73	0.879	0.233	22.67	0.871	0.231	22.69	0.881	0.227
8	Blue Cup	**27.19**	**0.922**	**0.183**	26.95	0.919	0.199	26.48	0.913	0.210	26.39	0.911	0.206	26.19	0.910	0.227
9	Burnout Wheel	**24.71**	**0.885**	0.199	23.35	0.865	0.199	23.38	0.867	0.197	23.39	0.860	0.204	23.78	0.876	**0.189**
10	Cheetah	28.88	**0.889**	**0.375**	**29.17**	0.887	0.392	28.40	0.881	0.389	28.41	0.880	0.405	27.98	0.866	0.430
11	Hard Disk	**28.12**	**0.959**	**0.230**	27.59	0.955	0.254	27.75	0.956	0.252	27.72	0.954	0.250	27.89	0.958	0.251
12	Wet Dog	**29.67**	**0.947**	**0.121**	29.35	0.945	0.130	29.09	0.943	0.145	28.93	0.939	0.146	29.17	0.942	0.145
13	Hair2	39.44	0.987	**0.126**	39.73	**0.988**	0.167	**39.75**	0.987	0.166	39.73	0.987	0.157	39.48	0.987	0.159
	Avg. Challenging	**28.05**	**0.922**	**0.184**	27.58	0.918	0.188	27.10	0.909	0.208	27.03	0.906	0.205	26.99	0.907	0.212
14	Biker	27.62	0.906	0.291	**29.05**	**0.929**	**0.208**	26.81	0.900	0.308	26.83	0.900	0.310	25.44	0.883	0.354
15	Drifting Car	29.44	0.942	0.216	**29.67**	**0.950**	**0.142**	28.25	0.935	0.217	28.15	0.933	0.220	26.83	0.918	0.265
16	Billiard	35.79	0.978	**0.062**	**36.21**	**0.982**	0.064	34.51	0.976	0.081	34.50	0.976	0.084	32.81	0.971	0.103
17	Bicycle	**26.84**	0.917	0.204	26.69	**0.921**	**0.186**	25.64	0.910	0.221	25.41	0.907	0.224	24.75	0.905	0.233
18	Ducks	27.24	0.828	0.324	**27.66**	**0.842**	**0.306**	27.02	0.823	0.356	27.16	0.826	0.366	26.76	0.819	0.386
19	Football Skills	31.31	0.967	0.058	**31.68**	**0.972**	**0.047**	30.19	0.963	0.065	29.82	0.960	0.071	29.34	0.958	0.075
20	Hula Hoop	28.88	0.915	0.201	**29.83**	**0.952**	**0.092**	27.74	0.909	0.204	27.69	0.907	0.203	26.39	0.864	0.280
21	Running Women	26.49	0.907	0.229	**26.87**	**0.927**	**0.189**	25.42	0.900	0.248	25.63	0.900	0.252	24.81	0.890	0.271
22	Running Men	25.91	0.921	0.194	**26.11**	**0.931**	**0.164**	24.81	0.908	0.220	24.96	0.908	0.224	24.39	0.900	0.243
23	Butterfly	**27.85**	0.939	0.141	27.70	**0.951**	**0.080**	26.38	0.929	0.156	26.44	0.927	0.156	25.63	0.904	0.196
24	Dog2	22.27	0.641	0.474	**22.78**	**0.676**	**0.382**	22.13	0.632	0.509	22.08	0.624	0.495	21.82	0.619	0.513
25	Backflip	32.55	0.970	0.064	**32.82**	**0.981**	**0.047**	31.07	0.973	0.070	30.70	0.970	0.077	29.76	0.965	0.090
	Avg. Full Set	**28.27**	**0.910**	0.190	28.22	**0.910**	**0.170**	27.29	0.900	0.210	27.23	0.900	0.210	26.79	0.890	0.230

Table 2: **Comparing temporal upsampling×8 results on our TSR video dataset.** *When applied to LTR videos with severe motion blur and motion aliasing, frame interpolation methods (e.g., Nvidia SlowMo [2] and DAIN [1]) score significantly lower. However, even methods trained to overcome such challenges, but were trained on external datasets (Flawless [3]), struggle on videos that do not represent the typical motions and dynamic behaviors they were trained on. Videos 1-13 are such challenging examples.*

#		Only Within			Only Across			Best Config		
		PSNR [dB]	SSIM	LPIPS	PSNR [dB]	SSIM	LPIPS	PSNR [dB]	SSIM	LPIPS
1	Water	34.39	0.980	0.012	**35.52**	0.987	0.008	**35.52**	**0.987**	**0.008**
2	Jelly	29.32	0.953	0.016	**29.92**	**0.955**	0.017	**29.92**	**0.955**	**0.015**
3	Diamonds	27.81	0.930	**0.036**	28.06	0.932	0.037	**28.38**	**0.933**	**0.036**
4	Fire	32.44	0.976	0.037	**32.50**	**0.976**	**0.036**	**32.50**	**0.976**	**0.036**
5	Hair3	28.67	0.927	0.052	**28.87**	**0.929**	**0.050**	**28.87**	**0.929**	**0.050**
6	Soap	31.35	0.953	0.049	**31.51**	**0.954**	**0.049**	**31.51**	**0.954**	**0.049**
7	Coin	26.64	0.900	0.103	**28.13**	**0.923**	**0.083**	**28.13**	**0.923**	**0.083**
8	Blue Cup	32.60	0.966	0.032	**32.85**	**0.968**	**0.031**	**32.85**	**0.968**	**0.031**
9	Burnout Wheel	31.31	0.948	0.031	**31.54**	**0.950**	**0.031**	**31.54**	**0.950**	**0.031**
10	Cheetah	35.68	**0.971**	**0.036**	**35.72**	0.970	0.038	**35.72**	**0.971**	**0.036**
11	Hard Disk	32.56	0.960	0.057	**32.62**	0.960	0.056	**32.62**	**0.960**	**0.055**
12	Wet Dog	34.20	0.972	**0.024**	34.12	0.972	0.024	**34.26**	**0.972**	**0.024**
13	Hair2	43.92	0.992	0.022	**43.96**	0.992	0.022	**43.96**	**0.992**	**0.021**
	Avg. Challenging	32.38	0.956	0.039	32.72	0.959	0.037	**32.75**	**0.959**	**0.037**
14	Biker	34.24	0.970	0.038	36.37	0.982	0.019	**36.64**	**0.984**	0.018
15	Drifting Car	34.92	0.979	0.030	35.15	0.980	0.028	**35.24**	**0.980**	0.028
16	Billiard	44.13	0.997	0.002	**44.53**	**0.997**	0.002	**44.53**	**0.997**	0.002
17	Bicycle	31.12	0.952	0.053	**31.20**	0.953	0.052	**31.20**	**0.953**	0.052
18	Ducks	35.74	0.956	0.068	35.83	0.957	0.068	**35.84**	**0.957**	0.064
19	Football Skills	**38.87**	**0.994**	0.006	38.58	0.993	0.006	**38.87**	**0.994**	0.006
20	Hula Hoop	**39.92**	**0.994**	0.004	39.44	0.993	0.005	**39.92**	**0.994**	0.004
21	Running Women	33.92	0.978	0.011	34.20	0.979	0.010	**34.24**	**0.980**	0.010
22	Running Men	31.13	0.962	0.034	**31.31**	**0.963**	0.032	**31.31**	**0.963**	0.032
23	Butterfly	**32.39**	**0.983**	0.017	32.06	0.980	0.018	**32.39**	**0.983**	0.017
24	Dog2	28.93	0.864	0.108	28.98	0.866	0.106	**29.02**	**0.866**	0.106
25	Backflip	42.82	0.998	0.002	43.14	0.998	0.002	**43.33**	**0.998**	0.002
	Avg. Full Set	33.96	0.962	0.035	34.25	0.964	**0.033**	**34.33**	**0.965**	**0.033**

Table 3: **Ablation study: 'Within' vs. 'Across' examples:** *Results of our atomic TSRx2 network, when trained on examples extracted from: (i) the same-dimension only ('Within'); (ii) across-dimensions only ('Across'); (iii) best configuration for each video – 'within', 'across', or both. Videos 1-13 are the Challenging dataset. The ablation results indicate that on average, the cross-dimension augmentations are more informative than the within (same-dimension) augmentations, leading to an overall improvement in PSNR, SSIM and LPIPS. However, since different videos have different preferences, training each video with its best 'within' and/or 'across' configuration can provide a small additional overall improvement.*

References

1. Bao, W., Lai, W.S., Ma, C., Zhang, X., Gao, Z., Yang, M.H.: Depth-aware video frame interpolation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 3703–3712 (2019) 1, 4
2. Jiang, H., Sun, D., Jampani, V., Yang, M.H., Learned-Miller, E., Kautz, J.: Super slomo: High quality estimation of multiple intermediate frames for video interpolation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018) 1, 4
3. Jin, M., Hu, Z., Favaro, P.: Learning to extract flawless slow motion from blurry videos. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2019) 1, 3, 4
4. Kiani Galoogahi, H., Fagg, A., Huang, C., Ramanan, D., Lucey, S.: Need for speed: A benchmark for higher frame rate object tracking. In: Proceedings of the IEEE International Conference on Computer Vision (CVPR) (2017) 1
5. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018) 1