Appendix

7. Implementation Details

7.1. Architecture of diff encoder

The strides in diff encoder depends on the resolution of input frame and the size of output diff embeddings, shown as Tab. 7.

size of diff embeddings	resolution	strides
$2 \times 40 \times 80$	640×1280	(2,2,2,2)
	480×960	(3,2,2)
	960×1920	(4,3,2)
$2 \times 10 \times 20$	960×1920	(4,4,3,2)

Table 7. Architecture of diff encoder.

7.2. Architecture of decoder

The architecture of decoders with CCU in different size is given in Tab. 8. C_{init} is the initial channel width of embeddings before feeding to decoder stages. Once feeding the $w_0 \times h_0 \times C_{init}$ embeddings into following stage, for example Stage 1 with stride $s = 5$, the size of output feature maps is $5w_0 \times 5h_0 \times C_1$, where $C_1 = \lfloor C_{init}/r \rfloor$ is the output channel width of Stage 1, $r = 1.2$ is the reduction rate for each stage and $\lfloor x \rfloor$ is the round down operator. K_c is the kernel size in CCU. The minimal and maximal kernel size in different decoder stages are 1 and 5, follow the setting in [3].

resolution	size	C_0	C_{init}	C_1	C_5	strides	K_c
640×1280	0.35	16	32	26	11	(5,4,4,2,2)	3
	0.75	16	48	40	18	(5,4,4,2,2)	3
	1.5	16	68	56	25	(5,4,4,2,2)	3
	3	16	95	79	37	(5,4,4,2,2)	3
480×960	3	16	110	91	42	(5,4,3,2,2)	1
960×1920	1.58	16	68	56	25	(5,4,4,3,2)	1
	3	16	92	76	35	(5,4,4,3,2)	1

Table 8. Architecture of decoder and CCU.

7.3. Experimental Details

In video compression, the network structure would be adjusted for different sizes and bpp, 1.58M with diff embedding in $2 \times 10 \times 20$ for 0.0146 bpp, 3M with diff embedding in $2 \times 10 \times 20$ for 0.0257 bpp and 3M with diff embedding in $2 \times 40 \times 80$ for 0.0517 bpp.

8. Additional quantitative results

8.1. Comparison for video interpolation on DAVIS Dynamic

Interpolation results between different methods on DAVIS Dynamic are shown in Tab. 12. We only compare DNeRV with hybrid-based implicit methods [3] because H-NeRV is the current best implicit method for video representation.

8.2. The effects of different compression techniques

Ablations for various compression technique on UVG is given in Tab. 11. In future work, more advanced model compression methods would be used on the NeRV methods owing to the fewer redundance in the weights.

8.3. The effects of different compression techniques

For the evaluation of video compression, the results of VMAF [32] are demonstrated in Tab. 9.

Bpp	Beauty	Bospho	Honey	Jockey	Ready	Shake	Yacht	avg.
0.015	77.74	71.43	93.71	68.02	53.55	80.74	57.55	71.82
0.025	83.78	78.18	93.16	75.38	60.97	82.53	63.45	76.78
0.05	85.15	77.45	94.22	84.02	67.47	86.13	60.09	79.22

Table 9. Number of VMAF on 960×1920 UVG in different Bpp.

8.4. Ablation results for optimizer

Results for optimizer ablations on Bunny with 0.35M size and 300 epochs is given in Tab. 10. Adan [50] is much more effective than Adam for larger learning rate.

9. Additional qualitative results

9.1. Visualization of video interpolation on UVG

Additional interpolation comparison on UVG is given in Fig. 8 and Fig. 9.

"Jockey" and "ReadySetGo" are two typical videos with large motion and dynamic scenes from UVG. In Fig. 8 and Fig. 9, we could find that the interpolations generated by DNeRV are obviously better than HNeRV. Some subtle spatial structures in interpolations of DNeRV, such as numbers on the screen or flagpole in the distance, remain nearly constant between adjacent frames.

9.2. Visualization of video interpolation on DAVIS Dynamic

Additional interpolation comparison on DAVIS Dynamic is given in Fig. 10, Fig. 11, Fig. 12 and Fig. 13.

DAVIS Dynamic is more difficult than UVG by reason of more dynamic scene changing and fewer frames. Although DNeRV outperforms HNeRV achieving the best results of implicit methods, but there is still much room for improvement. Once increasing the parameter quantity and utilizing task-specific modification, DNeRV could be competitive with state-of-the-art deep interpolation methods.

learning rate	optimizer	50	100	150	200	250	300
5e-4	Adam	24.97/0.769	27.86/0.873	28.99/0.905	30.10/0.920	30.66/0.926	30.80/0.927
	Adan	24.17/0.734	26.42/0.823	27.65/0.862	28.41/0.881	28.80/0.890	28.91/0.893
1e-3	Adam	26.36/0.829	24.67/0.776	27.06/0.841	27.71/0.863	28.46/0.876	28.56/0.878
	Adan	25.53/0.789	28.23/0.879	29.31/0.905	30.03/0.917	30.40/0.922	30.50/0.924
3e-3	Adam	18.39/0.519	18.81/0.548	19.31/0.584	19.18/0.583	19.32/0.591	19.36/0.594
	Adan	27.59/0.865	29.76/0.918	30.59/0.933	31.35/0.941	31.78/0.944	31.89/0.946

Table 10. Optimizer ablations on Bunny in PSNR/SSIM.

UVG	Beauty	Bospho	Honey	Jockey	Ready	Shake	Yacht
N/A	40.00/0.972	36.67/0.965	41.92/0.993	35.75/0.947	28.68/0.917	36.53/0.962	31.10/0.924
8-bit Quant	39.97/0.972	36.64/0.965	41.20/0.993	35.73/0.947	28.66/0.916	36.35/0.961	31.00/0.923
8-bit Quant + Pruning (10%)	39.38/0.971	36.41/0.964	39.95/0.991	35.50/0.946	28.55/0.915	35.42/0.959	30.78/0.921
8-bit Quant + Pruning (20%)	33.72/0.961	34.56/0.957	34.47/0.978	32.25/0.938	27.63/0.905	28.66/0.943	28.84/0.908

Table 11. Compression ablations on UVG in PSNR/SSIM.

9.3. Visualization of video inpainting on DAVIS Dynamic

Additional inpainting comparison on DAVIS Dynamic is shown in Fig. 14, Fig. 15 and Fig. 16.

Due to diff stream and CCU, DNeRV could model different regions of the frame more robustly, reduce the influence of masked regions. Besides, one limitation of DNeRV is that it couldn't model the detail texture well, and we will improve it in the future work.

9.4. Visualization of optical flow and difference stream

We conducted additional experiments on Bunny, following the same setting as Tab. 1a. The PSNR results are 29.13, 29.25, 28.84, and 28.70 in dB for the model sizes of 0.35M, 0.75M, 1.5M, and 3M. The optical flow is computed using Gunner Farneback algorithm by opencv-python 4.5.3 and numpy 1.19.5.

The visualization comparison between optical flow and diff stream is shown in Fig. 7. It can be clearly observed that, although optical flow contains motion information, it loses huge other information in pixel domain. Saliency motion information in optical flow may be key in action recognition or motion prediction, but it cannot bring much help for pixel-level reconstruction tasks. For example, the fluctuation of grass or the change of skin brightness with the light may not help to recognize the rabbit's movements, but they are essential for reconstruction. Diff stream records all these information in unbiased way.

Videos	DNeRV		HNeRV	
	test	train	test	train
Blackswan	23.89/0.712	28.98/0.874	21.67/0.589	28.76/0.865
Bmx-bumps	22.34/0.696	25.96/0.784	19.24/0.549	30.32/0.883
Camel	21.31/0.656	23.79/0.761	20.69/0.586	26.28/0.855
Breakdance	22.28/0.858	27.26/0.937	20.40/0.841	29.53/0.958
Car-round	20.42/0.725	28.91/0.931	16.92/0.560	28.23/0.919
Bmx-trees	21.68/0.644	28.88/0.867	18.39/0.453	28.99/0.872
Car-shadow	22.47/0.734	29.41/0.913	19.35/0.622	28.64/0.897
Cows	20.89/0.629	25.24/0.837	20.45/0.590	24.71/0.815
Dance-twirl	20.95/0.656	29.19/0.872	18.38/0.517	28.70/0.857
Dog	24.91/0.683	29.55/0.857	21.99/0.457	29.85/0.868
Car-turn	24.29/0.737	28.21/0.838	22.34/0.654	27.80/0.828
Dog-agility	20.57/0.730	27.14/0.852	17.14/0.609	26.21/0.818
Drift-straight	19.11/0.645	29.75/0.921	15.62/0.354	29.72/0.916
Drift-turn	21.22/0.649	29.45/0.849	18.44/0.501	28.43/0.815
Goat	20.46/0.554	28.63/0.908	18.22/0.327	27.69/0.891
Libby	24.24/0.688	32.22/0.906	20.00/0.472	30.75/0.871
Mallard-fly	21.81/0.610	28.25/0.809	19.23/0.397	27.26/0.788
Mallard-water	21.24/0.687	27.55/0.882	17.60/0.429	29.23 0.911
Parkour	22.13/0.680	27.32/0.879	18.82/0.488	26.77/0.863
Rollerblade	24.91/0.850	30.52/0.915	21.56/0.782	29.92/0.907
Scooter-black	17.15/0.633	27.26/0.926	14.37/0.416	26.33/0.901
Stroller	23.32/0.718	32.36/0.923	20.47/0.559	31.68/0.905
Average	**21.89/0.690**	28.45/0.875	19.15/0.534	28.44/0.873

Table 12. Interpolation results on DAVIS Dynamic.

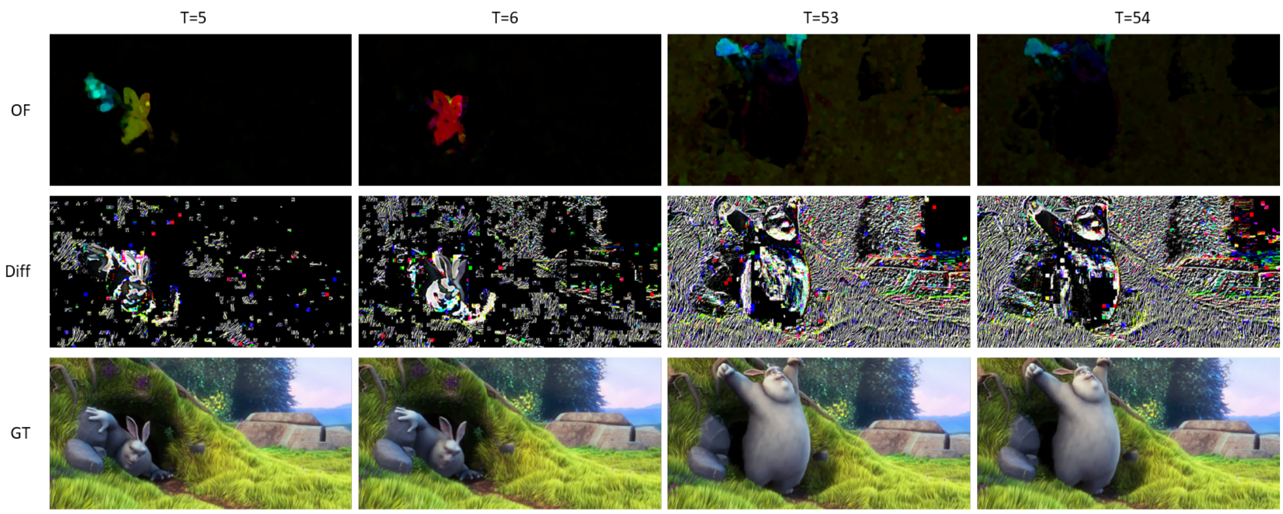

Figure 7. Comparison between optical flow and difference stream.

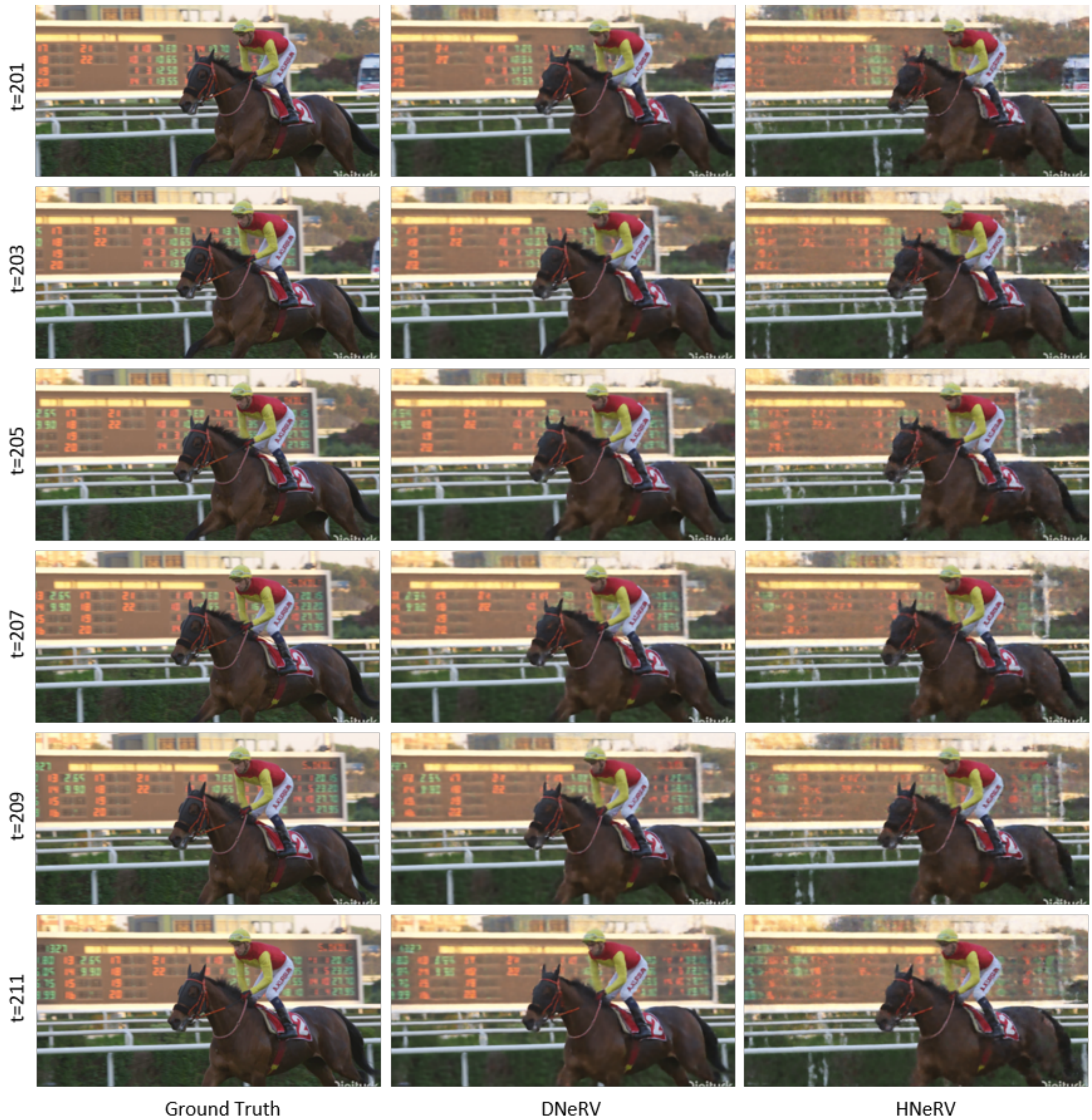

Figure 8. Additional examples for video interpolation on Jockey.

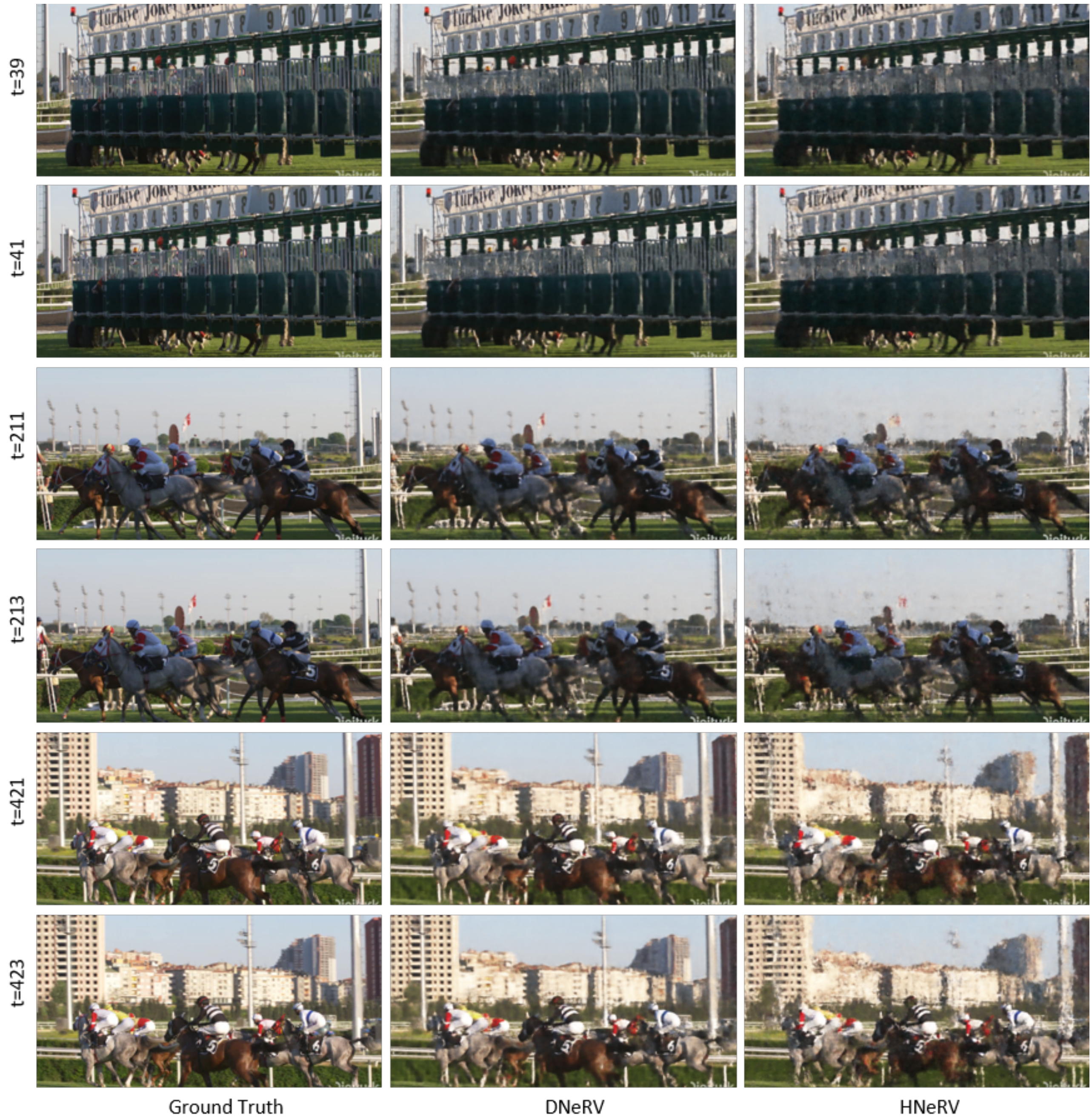

Figure 9. Additional examples for video interpolation on ReadySetGo.

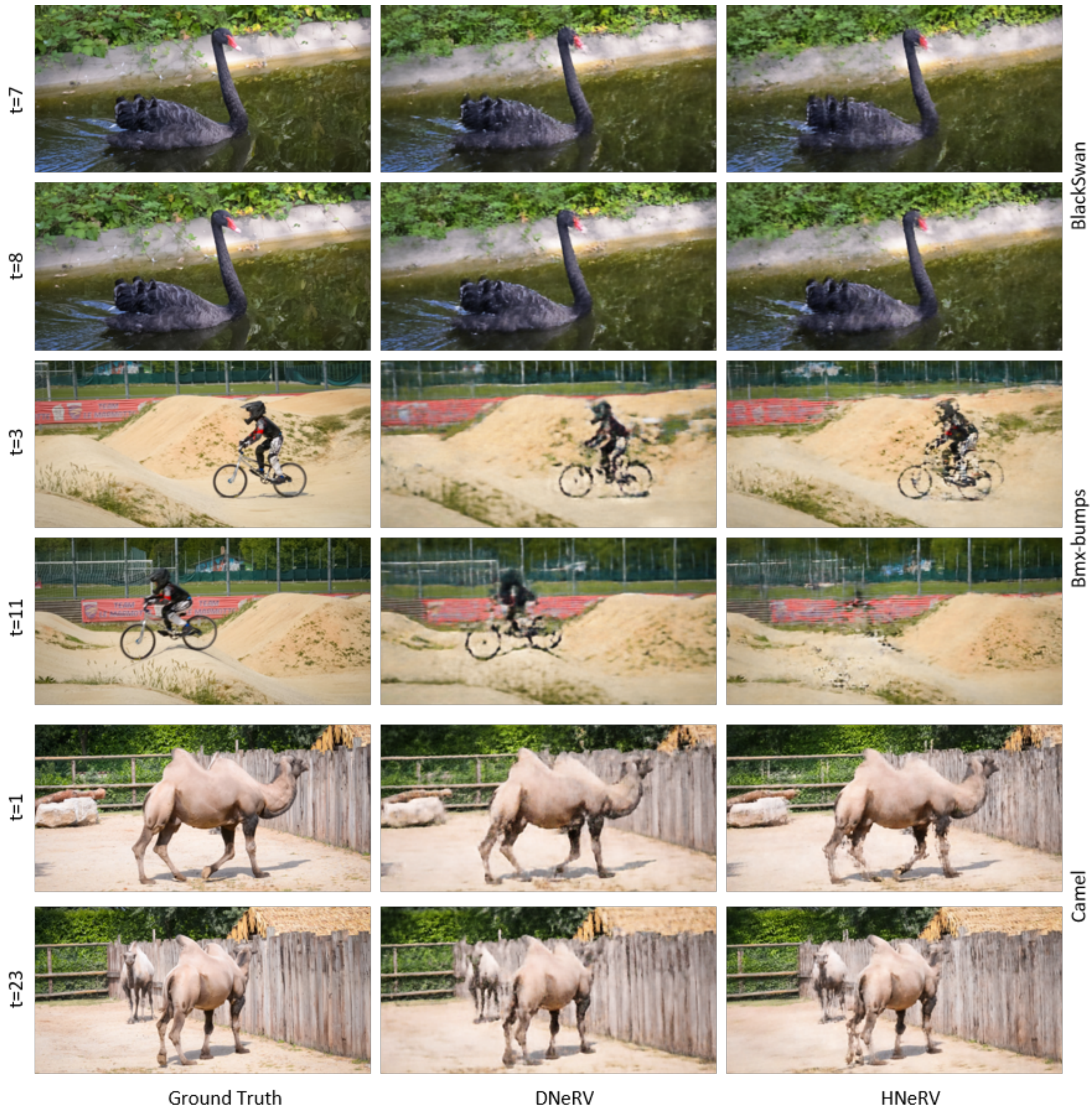

Figure 10. Additional examples for video interpolation on Blackswan, Bmx-bumps and Camel.

Figure 11. Additional examples for video interpolation on Breakdance, Car-roundabout and Car-shadow.

Figure 12. Additional examples for video interpolation on Dance-twril, Drift-straight and Drift-turn.

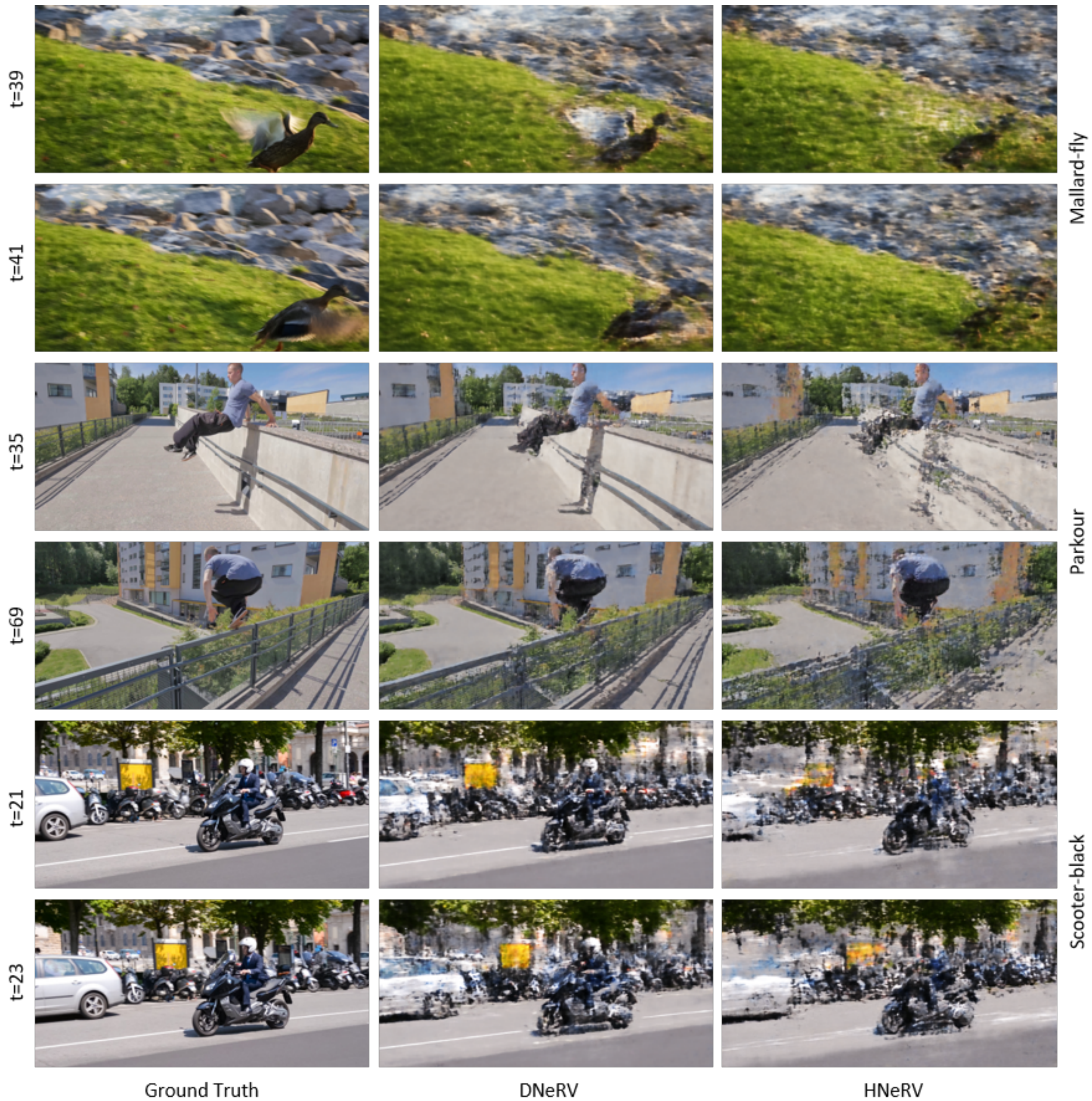

Figure 13. Additional examples for video interpolation on Mallard-fly, Parkour and Scooter-black.

Figure 14. Additional examples for video interpolation on Blackswan, Bmx-bumps and Bmx-trees.

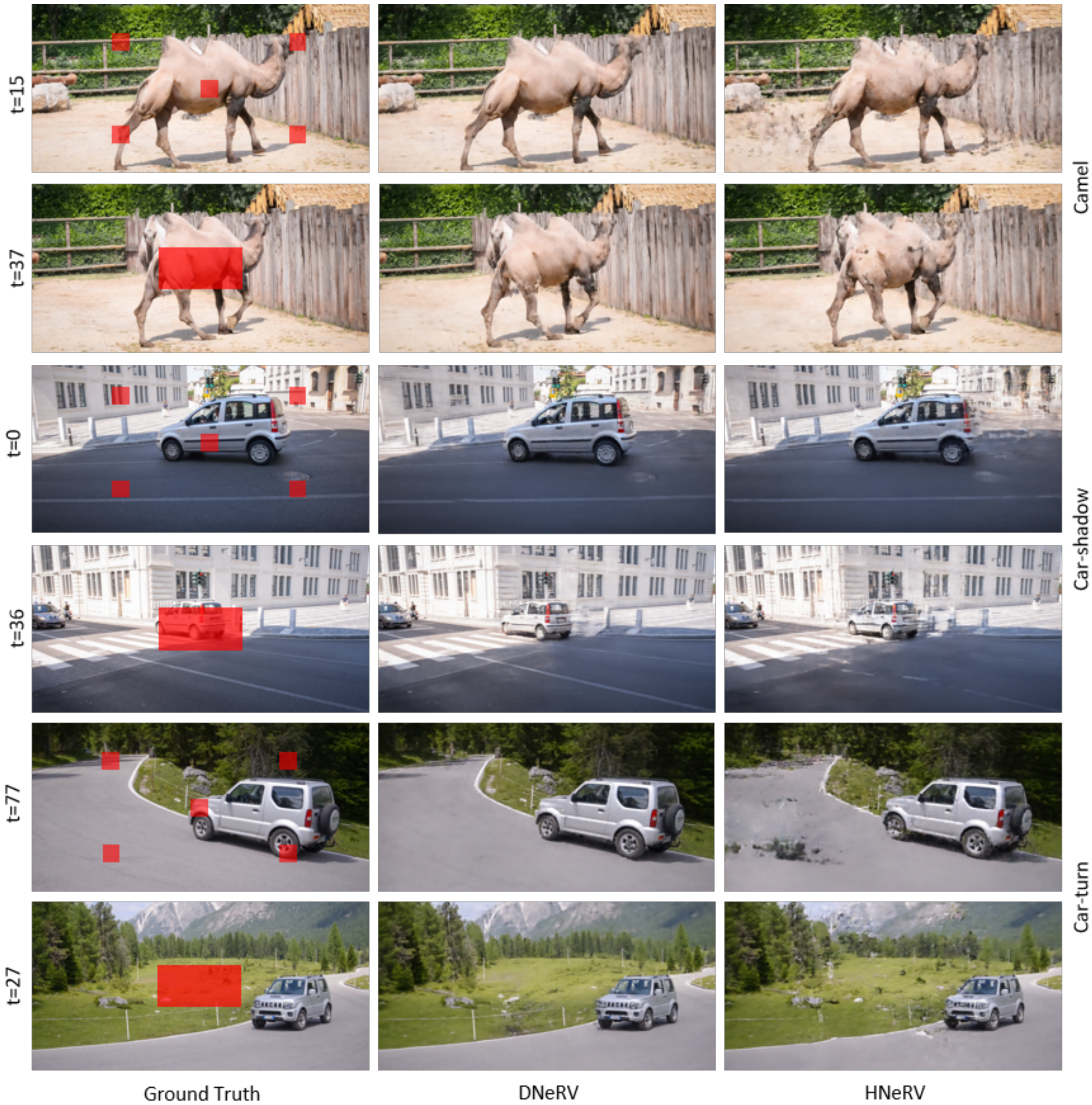

Figure 15. Additional examples for video interpolation on Camel, Car-shadow and Car-turn.

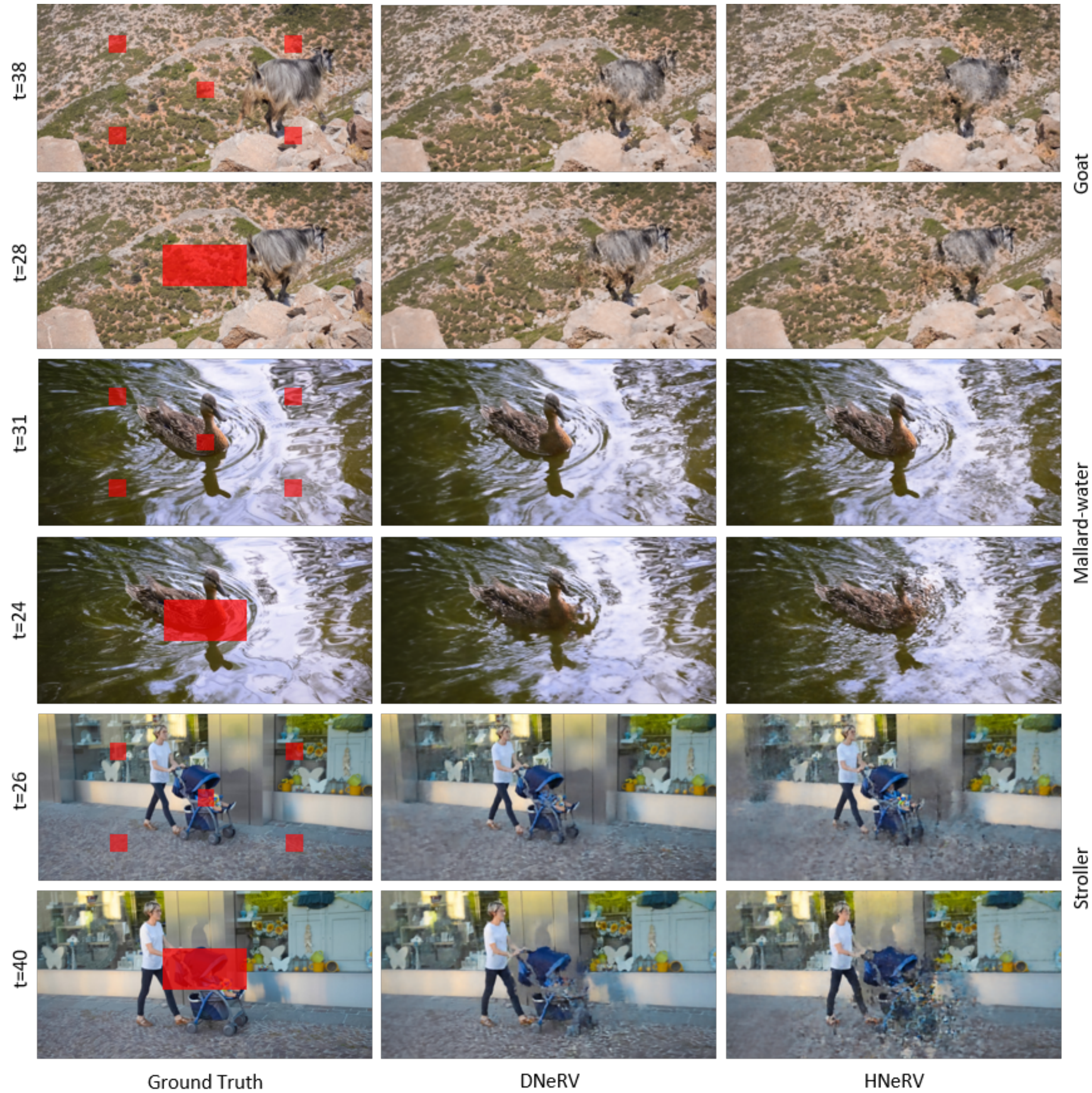

Figure 16. Additional examples for video interpolation on Goat, Mallard-water and Stroller.