

Supplementary Materials: QS-NeRV: Real-Time Quality-Scalable Decoding with Neural Representation for Videos

Anonymous Authors

1 ARCHITECTURE DETAILS OF SDMN

1.1 Decoder

The detailed network settings of SDMN decoder are displayed in Table 3. Assuming that the shape of the embedding E_t^{L0} is $C_0 \times h_0 \times w_0$, the channel will be expanded by a convolutional layer to C_{init} at first. Then, the expanded features are decoded progressively by five reparameterization-base decoding blocks. After being processed by the i -th block, where the kernel size and stride of convolution layers is $Ks[i]$ and $Strides[i]$, the channel dimension of the output feature is $Channels[i]$, and the resolution will be enlarged by a factor of $Strides[i]$.

1.2 Invertible Skip Connection

The denseblock of invnets consists of five 3×3 convolution layers. Each convolutional layer takes the concatenated result of the input of the block and all outputs of its previous convolution layers as input. It's worth noting that we only establish connections between the two layers with the smallest scales in the encoder and decoder to save the bitstream overhead. The shapes of the two frame-specific features \hat{f}_t transmitted in the connection for videos of different resolutions are presented in Table 1.

Table 1: The shape of \hat{f}_t for videos of different resolutions.

Video resolution	Shape ($[c \times h \times w]$)
960×480	$\begin{bmatrix} 1 \times 8 \times 16 \\ 1 \times 4 \times 8 \end{bmatrix}$
1920×960	$\begin{bmatrix} 1 \times 12 \times 24 \\ 1 \times 4 \times 8 \end{bmatrix}$
1280×640	$\begin{bmatrix} 1 \times 8 \times 16 \\ 1 \times 4 \times 8 \end{bmatrix}$

2 IMPLEMENT DETAILS OF VIDEO COMPRESSION

We use SHM 12.4, which is the official coding test tool of SHVC, to perform three-layer scalable video coding by following the command:

```
TAppEncoderStatic -c encoder_lowdelay_P_scalable.cfg -c layers.cfg -c VIDEO.cfg -q0 37 -q1 32 -q2 27 -b bitstream.bin
```

where $q0$, $q1$, and $q2$ denote the quantization parameters (QPs) set for BL, the first EL, and the second EL, respectively.

We follow the steps in [1] to generate the videos compressed by X264 and X265. We assign corresponding constant bitrates for different quality levels.

```
ffmpeg -i VIDEO/f%05d.png -c:v libx264 -preset medium -bf 0
```

```
-b:v BITRATE VIDEO.h264
```

```
ffmpeg -i VIDEO/f%05d.png -c:v libx265 -preset medium -x265-params bframes=0 -b:v BITRATE VIDEO.h265
```

where $BITRATE$ is set to 5000k, 7500k, and 10000k corresponding to 3M, 4.5M, and 6M model sizes of INR-based approaches.

3 MORE ABLATION STUDY

3.1 Reparameterization-base Decoding Block

We compare the performance of our proposed reparameterization-base decoding block and HNeRV decoding block. As shown in Table 2, the results illustrate that reparameterization allows the network to get a stronger representation capability.

Table 2: Reparameterization-base vs. HNeRV decoding block in SDMN, *Beauty*.

Decoding block	Reparameterization-base	HNeRV
PSNR (dB)	33.92	33.78

3.2 Model Quantization

We analyze the impact of model quantization on performance, and the results are shown in Table 4.

4 ADDITIONAL QUANTITATIVE RESULTS

Here, we provide video regression and interpolation results for all 22 sequences of the DAVIS dataset. The results are shown in Table 5.

5 ADDITIONAL QUALITATIVE RESULTS

We provide more video regression qualitative and video interpolation results of UVG dataset in Fig. 1 and 2.

REFERENCES

- [1] Hao Chen, Bo He, Hanyu Wang, Yixuan Ren, Ser Nam Lim, and Abhinav Shrivastava. 2021. Nerv: Neural representations for videos. *Advances in Neural Information Processing Systems* 34 (2021), 21557–21568.

Table 3: Architecture of SDMN decoder.

Resolution	Size	C_0	C_{init}	Channels	Strides	Ks
960×480	3M	16	98	[82, 68, 57, 48, 40]	[5, 4, 3, 2, 2]	[1, 3, 5, 5, 5]
1920×960	3M	16	84	[70, 58, 48, 40, 33]	[5, 4, 4, 3, 2]	[1, 3, 5, 5, 5]
1280×640	0.75M	12	43	[36, 30, 24, 20, 17]	[5, 4, 4, 2, 2]	[1, 3, 5, 5, 5]

Table 4: Quantization ablations on 1920×960 UVG in PSNR

Video	Size	Beauty	Bosph	Bee	Jockey	Ready	Shake	Yacht	avg.
N/A	3M	33.92	35.74	39.27	33.83	27.67	35.04	29.97	33.63
8-bit Quant		33.88	35.68	39.11	33.72	27.63	35.03	29.95	33.57
N/A	4.5M	34.11	36.69	39.38	34.89	28.75	35.68	30.94	34.35
8-bit Quant		34.06	36.67	39.31	34.83	28.74	35.65	30.90	34.31

Table 5: PSNR (dB) and SSIM on DAVIS at resolution 1920×960 . Bold means best results.

Video	Video Regression				Video Interpolation		
	HNeRV	DNeRV	FFNeRV	QS-NeRV	DNeRV	FFNeRV	QS-NeRV
Blackswan	30.35/0.891	30.20/0.898	30.64/0.940	32.78/0.942	27.40/0.834	26.30/0.786	30.31/0.902
Bmx-bumps	29.98/0.872	31.38/0.903	32.00/ 0.937	33.21/0.932	25.17/0.755	26.02/ 0.788	26.18/0.779
Bmx-trees	28.76/0.861	28.64/0.858	29.06/0.912	31.08/0.917	24.95/0.755	25.18/0.725	25.65/0.755
Breakdance	30.45/0.961	30.01/0.962	30.84/0.976	31.40/0.969	25.48/0.918	26.21/0.919	27.02/0.933
Camel	26.71/0.844	26.51/0.862	27.16/0.911	27.35/0.879	25.14/0.831	25.20/ 0.857	25.82/0.844
Car-round	27.75/0.912	28.04/0.917	29.43/0.952	29.11/0.930	24.70/0.859	27.72/0.932	25.11/0.863
Car-shadow	31.32/0.936	30.23/0.924	33.06/ 0.964	33.89/0.954	26.29/0.872	27.63/0.883	29.70/0.927
Car-turn	29.65/0.879	29.67/0.878	30.30/ 0.915	31.12/0.910	26.91/0.813	27.00/0.829	28.17/0.833
Cows	24.11/0.792	24.29/0.798	22.36/0.707	25.13/0.842	23.06/0.752	22.36/0.707	24.36/0.821
Dance-twirl	28.19/0.845	28.31/0.847	28.09/ 0.896	29.23/0.869	23.50/0.717	24.07/0.761	25.42/0.788
Dog	30.96/0.898	31.10/0.900	31.22/0.936	33.35/0.943	26.36/0.724	26.95/0.757	28.69/0.798
Dog-ag	28.75/0.893	29.91/0.921	34.57/0.982	32.97/0.960	23.99/0.825	29.77/0.935	25.09/0.835
Drift-straight	30.80/0.932	30.56/0.928	31.29/0.962	34.24/0.968	23.72/0.726	24.90/0.730	25.63/0.738
Drift-turn	29.72/0.834	29.99/0.844	30.43/0.813	31.76/0.885	24.00/0.696	23.36/0.683	25.56/0.728
Goat	26.62/0.858	26.90/0.863	25.62/0.875	28.61/0.906	22.27/0.634	22.06/0.650	24.65/0.761
Libby	32.69/0.917	33.17/0.925	33.03/0.948	35.11/0.952	27.35/0.761	28.04/0.830	27.54/0.699
Mallard-fly	29.22/0.848	28.98/0.839	29.89/ 0.915	30.88/0.897	24.58/0.684	25.14/ 0.724	25.51/0.711
Mallard-water	29.08/0.908	29.32/0.912	28.87/0.931	30.90/0.936	23.19/0.717	23.72/0.742	25.10/0.769
Parkour	26.56/0.851	26.69/0.853	26.97/0.874	28.01/0.891	23.31/0.739	23.85/0.774	24.55/0.791
Rollerblade	32.19/0.935	32.32/0.938	34.77/0.971	34.54/0.959	26.11/0.813	27.81/ 0.849	27.87/0.834
Scooter-black	27.38/0.923	27.86/0.932	27.75/0.954	31.26/0.961	20.69/0.708	21.90/0.719	22.92/0.750
Stroller	31.31/0.894	31.64/0.906	32.13/0.945	33.08/0.934	26.76/0.789	26.93/0.795	28.01/0.818
Average	29.21/0.886	29.35/0.891	29.98/0.919	31.31/0.924	24.77/0.769	25.58/0.790	26.31/0.804

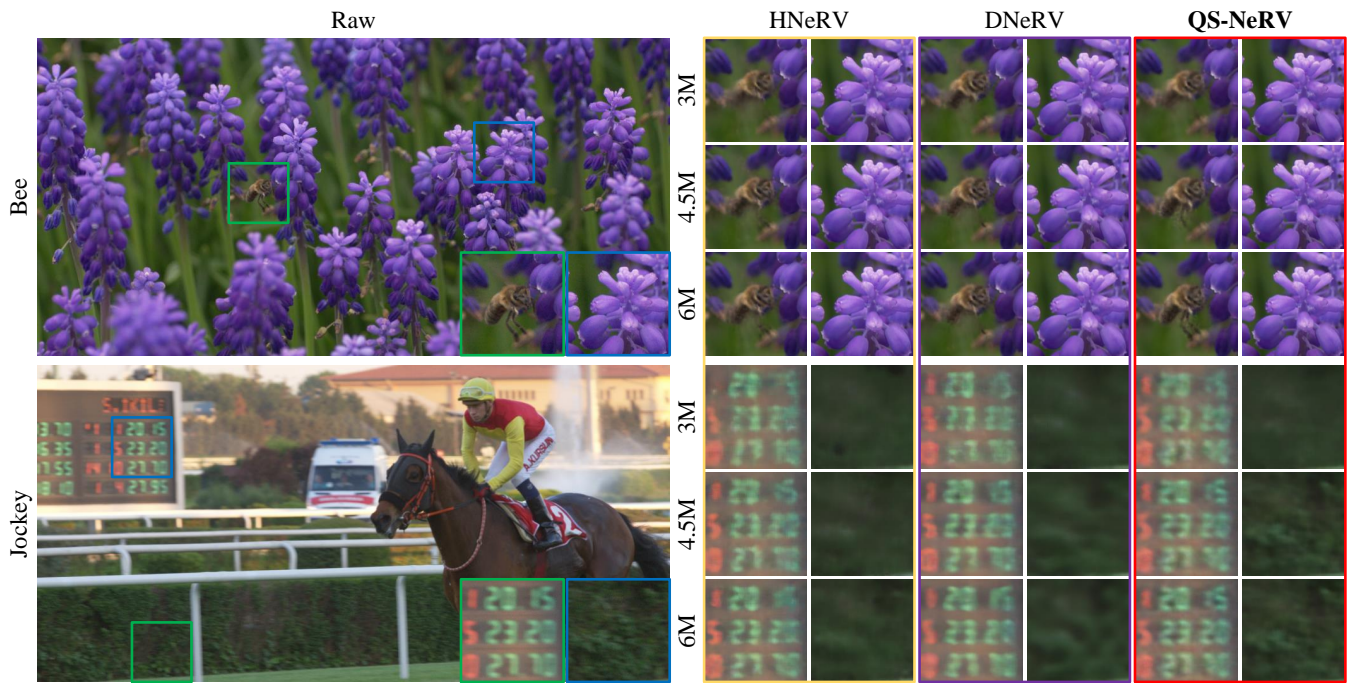


Figure 1: More visualization of video neural representations at various model sizes.

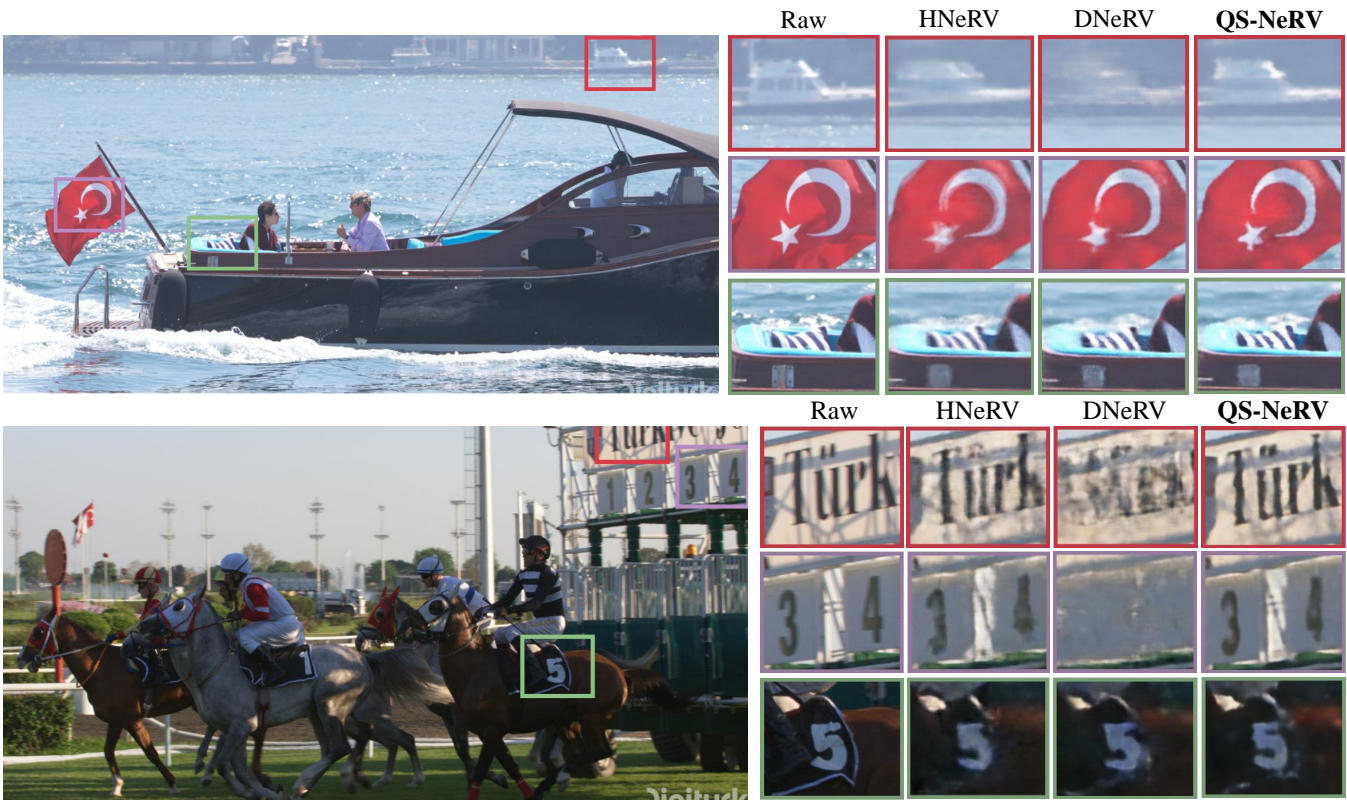


Figure 2: More visualization of video interpolation at 3M model size.