# Towards Scalable Neural Representation for Diverse Videos
## *Supplementary Material*

Bo He[1] Xitong Yang[2] Hanyu Wang[1] Zuxuan Wu[3] Hao Chen[1]
Shuaiyi Huang[1] Yixuan Ren[1] Ser-Nam Lim[2] Abhinav Shrivastava[1]

[1]University of Maryland, College Park [2]Meta AI [3]Fudan University

Sec. 1 reports additional results on AVA [1] and KTH Action Recognition [2] datasets for the video compression task. Sec. 2 presents more comparison with the state-of-the-art INR-based video representation method E-NeRV [3]. Sec. 3 shows more dataset-specific implementation details. We also show more qualitative results on the UVG, UCF101, Davis datasets in Sec. 4. Sec. 5 provides more comparisons and discussions with learning-based video compression methods. Finally, we discuss the limitation and some future work of our paper in Sec. 6.

## 1. Additional Video Compression Results

To further demonstrate the effectiveness of D-NeRV, we conduct additional experiments on the AVA Actions [1] dataset and KTH Action Recognition [2] dataset for the video compression task.

**AVA Actions Dataset** For the AVA Actions dataset, each original video is a full movie lasting about 1-2 hours, which is much longer than short action videos (around 10 seconds) from the UCF101 and UVG datasets. We sample 10 videos with a spatial size 256×384 and a frame rate of 1 fps. The PSNR and MS-SSIM results under different compression ratios (indicated with S / M / L) are shown in the Table 11. We can see that D-NeRV consistently outperforms NeRV [4] and H.264 [5] when encoding especially long videos.

**KTH Action Recognition Dataset** The KTH Action Recognition [2] dataset consists of grayscale video sequences of 25 people performing six different actions: walking, jogging, running, boxing, hand waving, and hand clapping. The background is uniform and a single person performs actions in the foreground. The videos have 120×160 spatial size and 25 fps frame rates. Similar to the results on other datasets, our D-NeRV achieves the best performances when comparing to H.264 and NeRV in Table 12.

## 2. Additional Comparison with E-NeRV

We conduct an additional comparison with E-NeRV on the UVG dataset by following the same experimental setting as Table 1 from E-NeRV [3]. The original E-NeRV paper uniformly samples 150 frames from each video and resizes the input video from 1080×1920 to 720×1280, and fits each video with a much larger model size (12.5M). The results of NeRV and E-NeRV in Table 13 are the reported performance in Table 1 from the original E-NeRV paper. As we can see from Table 13, when using a much larger model size to fit each downsampled video, the PSNR scores of both NeRV and E-NeRV are higher and the performance gap between E-NeRV and NeRV becomes greater, comparing to the results of Table 1 in our main paper. However, our D-NeRV still outperforms E-NeRV by 0.82 dB and achieves the best result. It proves the superior advantages of D-NeRV over the state-of-the-art INR-based video representation method E-NeRV.

Table 11. Video compression results on the AVA dataset.

| Model | PSNR | | | MS-SSIM | | |
|---|---|---|---|---|---|---|
| | S | M | L | S | M | L |
| H.264 | 27.32 | 28.91 | 30.49 | 0.853 | 0.897 | 0.923 |
| NeRV | 26.48 | 27.28 | 28.21 | 0.840 | 0.868 | 0.893 |
| D-NeRV | **28.77** | **29.57** | **30.60** | **0.886** | **0.903** | **0.924** |

Table 12. Video compression results on the KTH dataset.

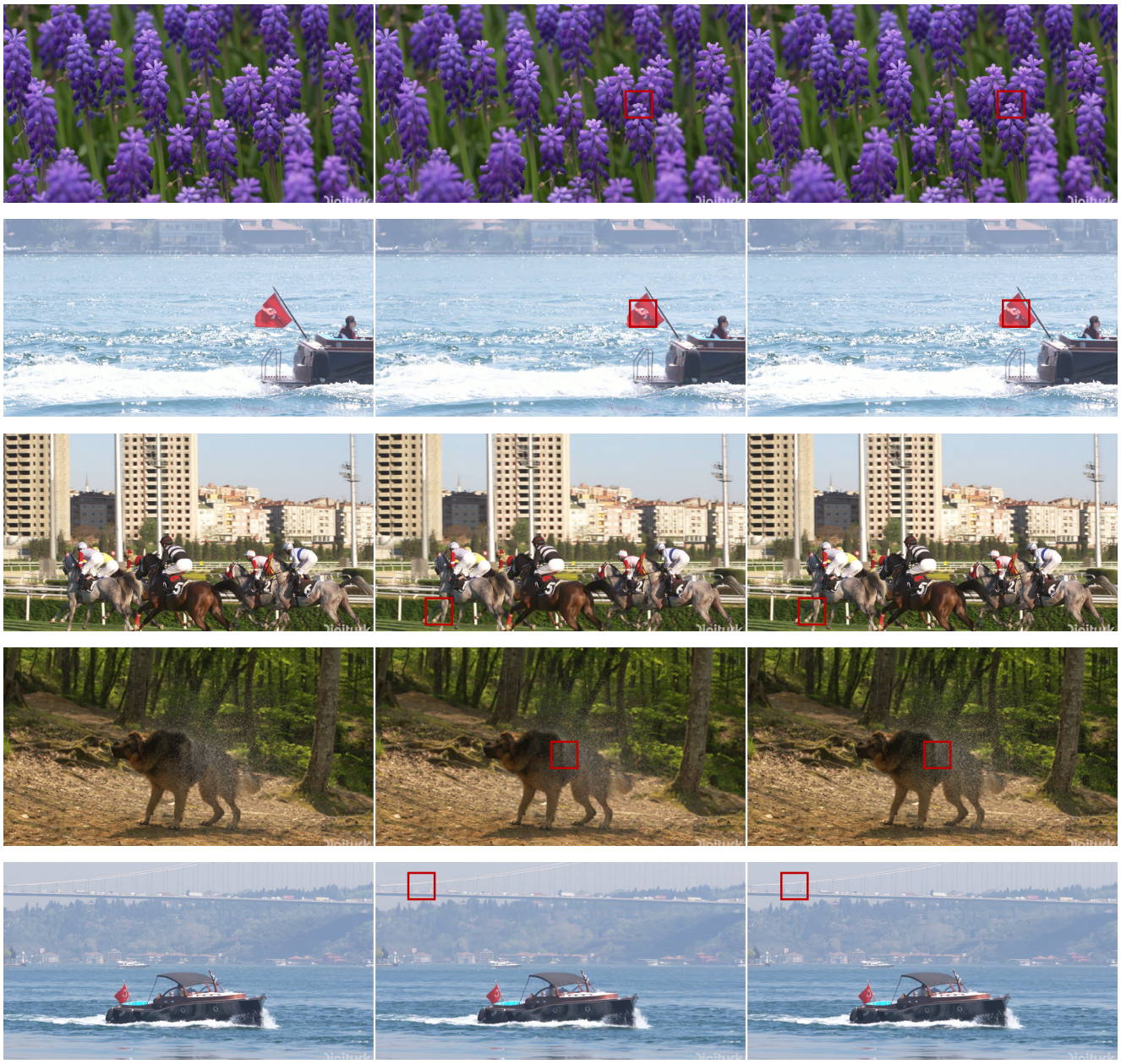
| Model | PSNR | | | MS-SSIM | | |
|---|---|---|---|---|---|---|
| | S | M | L | S | M | L |
| H.264 | 29.61 | 32.72 | 34.51 | 0.691 | 0.801 | 0.860 |
| NeRV | 30.56 | 32.14 | 33.31 | 0.701 | 0.748 | 0.784 |
| D-NeRV | **31.90** | **34.46** | **36.15** | **0.745** | **0.849** | **0.892** |

## 3. Experiment Details

On the UVG dataset, to compare with state-of-the-art video compression methods, we run experiments with 1600 epochs and a batch size of 32 and a learning rate of 5e-4. Due to the GPU memory limitation, we split the 1024×1920 input video frames into 256×320 image patches. We re-

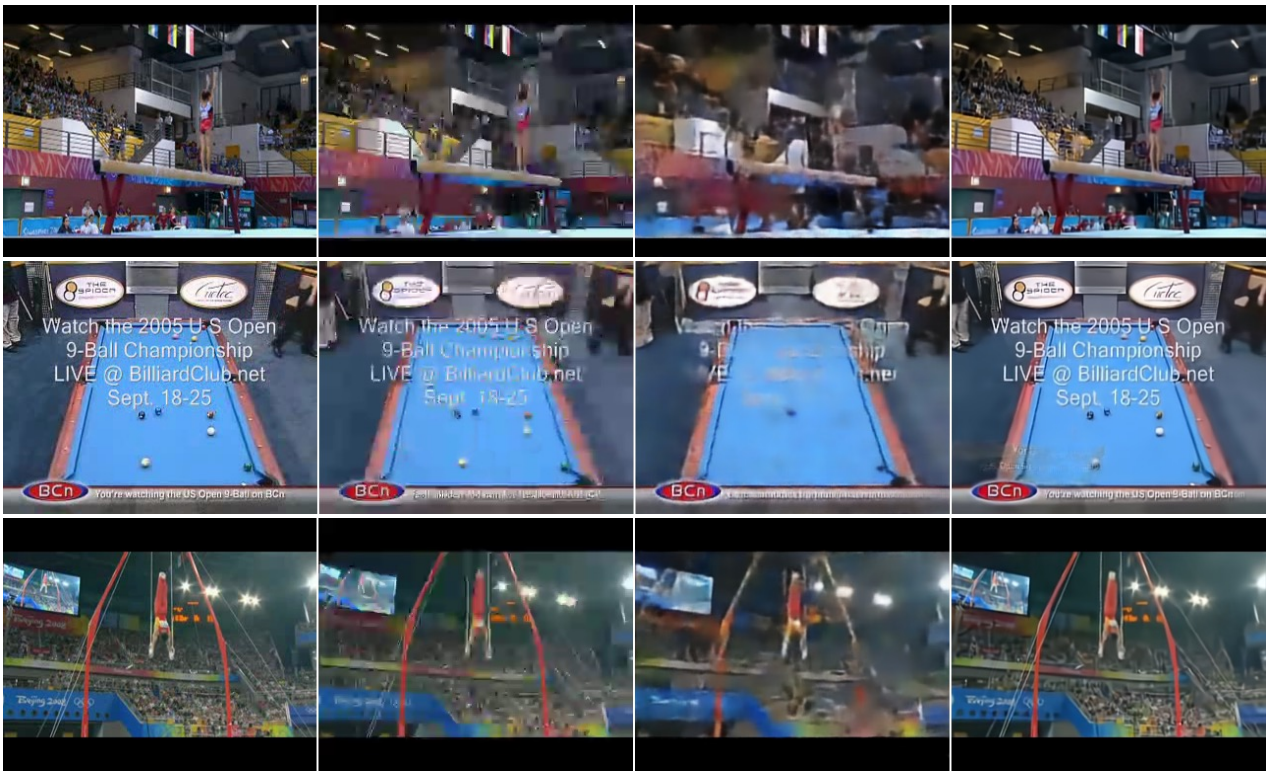
Table 13. Video reconstruction comparison between our D-NeRV, NeRV [4] and E-NeRV [3] on 7 videos from the UVG dataset. We follow the same setting as E-NeRV, which uniformly samples 150 frames from each video, resizes the input video from 1080×1920 to 720×1280 and trains models for 300 epochs.

| Video | Beauty | Bosphorus | Bee | Jockey | SetGo | Shake | Yacht | avg. |
|---|---|---|---|---|---|---|---|---|
| NeRV | 36.06 | 37.35 | 41.23 | 38.14 | 31.86 | 37.22 | 32.45 | 36.33 |
| E-NeRV | 36.72 | 40.06 | 41.74 | 39.35 | 34.68 | 39.32 | 35.58 | 38.21 |
| D-NeRV | 37.53 | 40.74 | 39.89 | 39.94 | 37.51 | 38.85 | 38.63 | **39.03** |



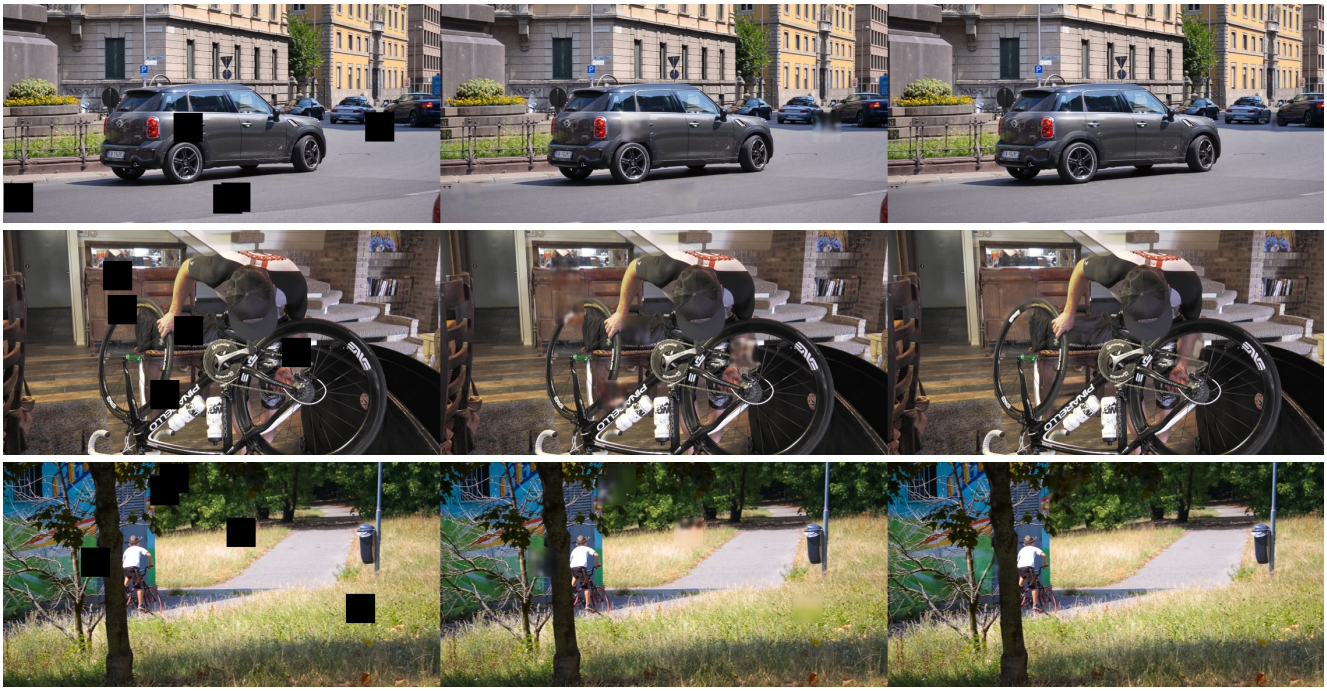
(a) Ground Truth (b) HEVC (0.029 BPP) (c) D-NeRV (0.026 BPP)

Figure 7. Visualization of ground-truth, HEVC, and D-NeRV for the video compression task on the UVG dataset. Red rectangles highlight the regions that HEVC fails to synthesize correctly and faithfully while our D-NeRV succeeds. Please zoom in to see the details.




(a) Ground Truth (b) H.264-S (c) NeRV-S (d) D-NeRV-S

Figure 8. Visualization of ground-truth, H.264, NeRV and D-NeRV for the video compression task on the UCF101 dataset.



(a) Ground Truth (b) NeRV (c) D-NeRV

Figure 9. Visualization of ground-truth, NeRV and D-NeRV for the video inpainting task on the Davis dataset. Please zoom in to see the details.

gard the patches at the same spatial location across different timesteps as a single video. On the AVA and KTH datasets, we run experiments with 800 epochs, a batch size of 32, and a learning rate of 5e-4. In our experiments, we set upscale factors 4, 2, 2, 2, 2 for each block. For the video compression task, following NeRV [4], we perform the model quantization and weight encoding steps but without the extra model pruning step to expedite the training process. And the quantization bit is set to 8 for all the datasets.

For the keyframe image compression, we use pre-trained [6] models to compress and decode the keyframes. Different pre-trained image compression models can compress keyframes with different compressed ratios.

## 4. More Qualitative Results

D-NeRV can produce clearer frames with less noise. Figure 7 displays a few samples from the UVG dataset. The red rectangles show the regions where our D-NeRV outperforms HEVC [7], for example, the flower, the flag, and the leg of the horse.

D-NeRV also achieves better qualitative results on UCF101 dataset as shown in Figure 8. For example, the athlete is more clear than NeRV and H.264 in row 1 and row 3. In row 2, D-NeRV also distinguishes from other methods when showing the foreground texts.

We also show more qualitative results on the Davis dataset for the video inpainting task in Figure 9. D-NeRV can inpaint the mask area more faithfully and naturally without blurry effects.

## 5. More Discussion

In this section, we compare and discuss our D-NeRV with existing learning-based video compression methods in more detail.

The key significant difference between D-NeRV and existing learning-based video compression methods is the way compressed videos are represented – neural network *vs*. latent codes, respectively. Since INR-based D-NeRV represents videos as a neural network, it can ***implicitly*** estimate flow and interpolate keyframes. In contrast, other learning-based methods that ***explicitly*** represent videos as latent codes generated by compressing flows and residuals for each frame. Due to the implicit design, D-NeRV is a more general architecture that can be applied to video compression and other video tasks such as video inpainting. On the other hand, these learning-based methods, including interpolating images (*e.g*., VCII [8]) and predicting flow estimation (*e.g*., LVC [9], SSF [10], FVC [11]), all decode video frames sequentially because of reliance on previous frames, which leads to a much worse decoding speed. In contrast, based on INR design, D-NeRV can reconstruct video frames parallelly with a faster speed.

## 6. Limitations and Future Work

INR-based methods often require longer training iterations than learning-based compression methods to better capture the high-frequency details. And they can not be generalized to unseen videos, which means they can only be trained and tested on the same videos. We believe more exploration of the generalization ability can be a good research direction for the INR-based video representation models. In addition, our current D-NeRV design still encodes the sampled keyframes by image compression techniques, however, encoding the sampled keyframes by a separate implicit neural network can make the whole pipeline more unified and may achieve better performances.

## References


[1] Chunhui Gu, Chen Sun, David A Ross, Carl Vondrick, Caroline Pantofaru, Yeqing Li, Sudheendra Vijayanarasimhan, George Toderici, Susanna Ricco, Rahul Sukthankar, et al. Ava: A video dataset of spatio-temporally localized atomic visual actions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6047–6056, 2018. 1

[2] Christian Schuldt, Ivan Laptev, and Barbara Caputo. Recognizing human actions: a local svm approach. In *Proceedings of the 17th International Conference on Pattern Recognition, 2004. ICPR 2004.*, volume 3, pages 32–36. IEEE, 2004. 1

[3] Zizhang Li, Mengmeng Wang, Huaijin Pi, Kechun Xu, Jianbiao Mei, and Yong Liu. E-nerv: Expedite neural video representation with disentangled spatial-temporal context. In *European Conference on Computer Vision*, pages 267–284. Springer, 2022. 1, 2

[4] Hao Chen, Bo He, Hanyu Wang, Yixuan Ren, Ser Nam Lim, and Abhinav Shrivastava. Nerv: Neural representations for videos. *Advances in Neural Information Processing Systems*, 34, 2021. 1, 2, 4

[5] Thomas Wiegand, Gary J Sullivan, Gisle Bjontegaard, and Ajay Luthra. Overview of the h. 264/avc video coding standard. *IEEE Transactions on circuits and systems for video technology*, 13(7):560–576, 2003. 1

[6] Zhengxue Cheng, Heming Sun, Masaru Takeuchi, and Jiro Katto. Learned image compression with discretized gaussian mixture likelihoods and attention modules. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7939–7948, 2020. 4

[7] Gary J Sullivan, Jens-Rainer Ohm, Woo-Jin Han, and Thomas Wiegand. Overview of the high efficiency video coding (hevc) standard. *IEEE Transactions on circuits and systems for video technology*, 22(12):1649–1668, 2012. 4

[8] Chao-Yuan Wu, Nayan Singhal, and Philipp Krahenbuhl. Video compression through image interpolation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 416–431, 2018. 4

[9] Oren Rippel, Sanjay Nair, Carissa Lew, Steve Branson, Alexander G Anderson, and Lubomir Bourdev. Learned video compression. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3454–3463, 2019. 4

[10] Eirikur Agustsson, David Minnen, Nick Johnston, Johannes Balle, Sung Jin Hwang, and George Toderici. Scale-space flow for end-to-end optimized video compression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8503–8512, 2020. 4

[11] Zhihao Hu, Guo Lu, and Dong Xu. Fvc: A new framework towards deep video compression in feature space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1502–1511, 2021. 4