

HNeRV: A Hybrid Neural Representation for Videos

Hao Chen¹ Matthew Gwilliam¹ Ser-Nam Lim² Abhinav Shrivastava¹

¹University of Maryland, College Park ²Meta AI

{chenh, mgwillia, abhinav}@cs.umd.edu, sernamlim@meta.com

<https://haochen-rye.github.io/HNeRV/>

Abstract

Implicit neural representations store videos as neural networks and have performed well for various vision tasks such as video compression and denoising. With frame index or positional index as input, implicit representations (NeRV, E-NeRV, etc.) reconstruct video frames from fixed and content-agnostic embeddings. Such embedding largely limits the regression capacity and internal generalization for video interpolation. In this paper, we propose a Hybrid Neural Representation for Videos (HNeRV), where a learnable encoder generates content-adaptive embeddings, which act as the decoder input. Besides the input embedding, we introduce HNeRV blocks, which ensure model parameters are evenly distributed across the entire network, such that higher layers (layers near the output) can have more capacity to store high-resolution content and video details. With content-adaptive embeddings and re-designed architecture, HNeRV outperforms implicit methods in video regression tasks for both reconstruction quality (+4.7 PSNR) and convergence speed (16× faster), and shows better internal generalization. As a simple and efficient video representation, HNeRV also shows decoding advantages for speed, flexibility, and deployment, compared to traditional codecs (H.264, H.265) and learning-based compression methods. Finally, we explore the effectiveness of HNeRV on downstream tasks such as video compression and video inpainting.

1. Introduction

Given the massive amount of videos generated every day, storing and transferring them efficiently is a key task in computer vision and video processing. Even for modern storage systems, the space requirements of raw video data can be overwhelming. Despite storage becoming cheaper, network speeds and I/O processing remain a bottleneck and make transferring and processing videos expensive.

Traditional video codecs, such as H.264 [47] and

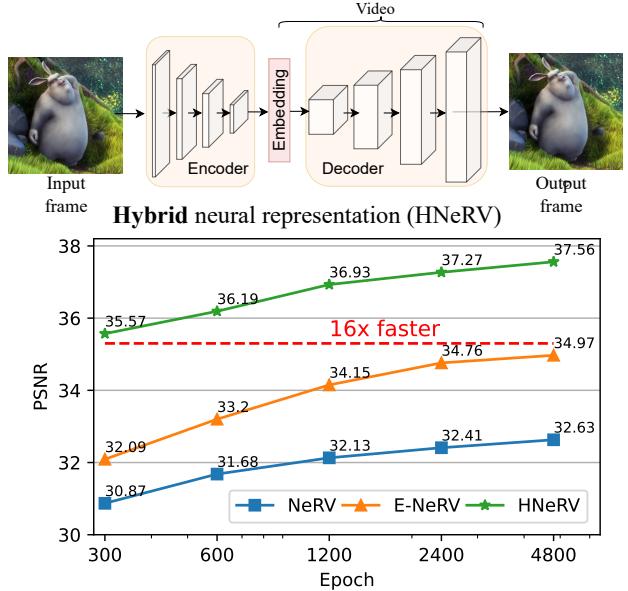


Figure 1. **Top:** hybrid neural representation with learnable and content-adaptive embedding (ours). **Bottom:** video regression for hybrid and implicit neural representations.

HEVC [41], rely on a manually-designed encoder and decoder based on discrete cosine transform [3]. With the success of deep learning, many attempts [2, 7, 11, 23, 24, 27, 36, 37, 49] have been made to replace certain components of existing compression pipelines with neural networks. Although these learning-based compression methods show high potential in terms of rate-distortion performance, they suffer from complex pipelines and expensive computation, not just to train, but also to encode and decode.

To address the complex pipelines and heavy computation, implicit neural representations [6, 33, 35, 38, 40] have become popular due to their simplicity, compactness, and efficiency. These methods show great potential for visual data compression, such as COIN [8] for image compression, and NeRV [4] for video compression. By representing videos as neural networks, video compression problems can

be converted to model compression problems, which greatly simplifies the encoding and decoding pipeline.

Implicit representation methods for video compression present a major trade-off: they embrace simplicity at the expense of generalizability. Given a frame index t as input, NeRV [4] uses a fixed position encoding function and a learnable decoder to reconstruct video frames from temporal embeddings. Another implicit representation, E-NeRV [22], takes a temporal embedding and spatial embedding to reconstruct video frames. Since the embeddings of NeRV and E-NeRV are based on spatial and/or temporal information only; without connection to the actual content of frames, they are content-agnostic. For decoding, NeRV-like models compute these embeddings using frame index alone, without access to the original frame. This is quite elegant for video compression, since instead of storing many frame embeddings, one would only need to store model weights and basic metadata (*e.g.*, number of frames).

However, this comes with some major disadvantages. Firstly, since embeddings are content-agnostic, and due to how the temporal embeddings are computed, there is no way to meaningfully interpolate between frames. Secondly, and more importantly, the positional embedding used by the fully-implicit models provides no visual prior and limits the regression capacity, since all the information needs to be learned by and stored in the video decoder.

In this paper, we propose a learnable encoder as a key component of hybrid neural representation for videos (HNeRV, Figure 1 (top)). Our proposed neural representation is a hybrid between implicit (network-centric) and explicit (embedding-centric) approaches since it stores videos in two parts: the tiny content-adaptive frame embeddings and a learned neural decoder. Besides the issue of content-agnostic embedding, prior work such as NeRV also suffers from an imbalance in the distribution of model parameters. In these decoders, later layers (closer to the output image) have much fewer parameters than earlier layers (closer to the embedding). This hinders NeRV’s ability to effectively reconstruct massive video content while preserving frame details. To rectify this, we introduce the HNeRV block, which increases kernel sizes and channel widths at later stages. With HNeRV blocks, we can build video decoders with parameters that are more evenly distributed over the entire network. As a hybrid method, HNeRV improves reconstruction quality for video regression and boosts the convergence speed by up to $16\times$ compared to implicit methods, shown in Figure 1 (bottom). With content-adaptive embeddings, HNeRV also shows much better internal generalization (ability to encode and decode frames from the video that were not seen during training), and we verify this by frame interpolation results in Section 4.2.

HNeRV only requires a network forward operation for video decoding, which offers great advantages over tradi-

tional codecs and prior deep learning approaches in terms of speed, flexibility, and ease of deployment. Additionally, most other video compression methods are auto-regressive and there is a high dependency on the sequential order of video frames. In contrast, there is no dependency on the sequential order of frames for HNeRV, which means it can randomly access frames efficiently to decode frames in parallel. Such simplicity and parallelism make HNeRV a good codec for further speedups, like a special neural processing unit (NPU) chip, or parallel decoding with huge batches.

HNeRV is still viable for video compression, while also showing promising performance for video restoration tasks. We design our encoder such that it can also be compressed; additionally, our HNeRV decoder blocks perform well in the model compression regime, such that HNeRV is competitive with state-of-the-art methods. We posit that neural representation can be robust to distortion in pixel space and therefore restore videos which have undergone distortions. We verify this observation on the video inpainting task.

In summary, we propose a hybrid neural representation for videos. With content-adaptive embedding and re-designed architecture, HNeRV shows much better video regression performance over implicit methods, in reconstruction quality (+4.7 PSNR), convergence speed ($16\times$ faster), and internal generalization. As an efficient video codec, HNeRV is easy to deploy, and is simple, fast, and flexible during video decoding. Finally, HNeRV shows good performance over downstream tasks like video compression and video inpainting.

2. Related Work

Neural Representation. Implicit representations fit to each individual test signal [29] where the model is regressed to a given image, scene, or video. Most implicit neural representations are coordinate-based. These coordinate-based implicit representations are used in image reconstruction [40, 42], shape regression [6, 33], and 3D view synthesis [31, 38]. NeRV [4] instead proposes an image-wise implicit representation, which takes frame indices as inputs and leverages neural representation for fast and accurate video compression. Relying only on index, and not coordinates, speeds up the encoding and decoding process compared to coordinate-based (pixel-wise) methods. Based on NeRV, CNeRV [5], DNeRV [14], E-NeRV [22], and NIRVANA [28] further improves the video regression performance. Traditional autoencoders would not be considered implicit representations since most information is stored in their large image-specific embeddings. Nevertheless, they are a form of neural representation, and HNeRV borrows the general concept for its encoder from standard U -shaped autoencoders [3, 18, 34, 44]. HNeRV keeps the embedding intentionally tiny and compact, so as to keep most of the representation implicit (stored in the decoder).

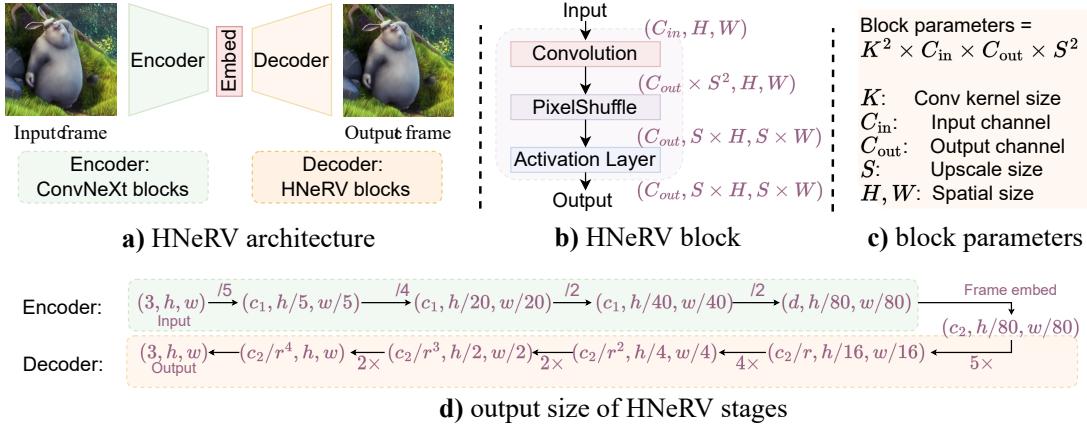


Figure 2. **a)** HNeRV uses ConvNeXt blocks to encode frames as tiny embeddings, which are decoded by HNeRV blocks. **b)** HNeRV blocks consist of three layers: convolution, PixelShuffle, and activation (with input/output size illustrated). **c)** We demonstrate how to compute parameters for a given HNeRV block. **d)** Output size of each stage with strides 5,4,2,2.

Video Compression. Traditional video compression methods such as MPEG [20], H.264 [48], and H.265 [41] achieve good reconstruction results with decent decompression speeds. Recently, deep learning techniques have been proposed for video compression. While these approaches focus on replacing the entire compression pipeline, they each borrow principles from the traditional handcrafted approaches. Some have framed the problem primarily as image compression and interpolation [7, 49], or attempt to solve this task with image compression via autoencoders [11], or focus purely on interpolation for the sake of compression [25]. Others essentially reformulate traditional video compression pipelines using deep learning tools [2, 23, 37], at varying levels of complexity. Recent approaches have focused on tackling the computational inefficiencies of existing art, including by fine-tuning traditional codecs [17], and by optimizing pieces of the compression pipeline [36]. The approach which inspired much of this work, NeRV, responds to these same inefficiencies by proposing a specialized architecture for video memorization [4]. Once video is represented as a neural network, the video compression problem can be converted to a model compression problem and achieve good bit-distortion performance. With learnable embeddings and re-designed decoder blocks, HNeRV improves the video regression capacity and convergence speed, while video compression is still viable by model compression.

Model Compression. NeRV formulated video compression as model compression [4], which is a diverse area. In this paper we apply only a small subset of possible methods. We use weight pruning [13] and weight quantization [10, 16, 19]. We also use entropy encoding for lossless compression after pruning and quantization [12, 15]. Note that many other model compression methods can be lever-

aged to further reduce the size and video neural representation can always benefit from developments in the model compression area.

Video Inpainting & Internal Learning. Video inpainting is typically framed as some combination of object removal and attempting to recreate missing regions of images. Whereas some methods rely on priors from training on large datasets [46], ours has more in common with a recent zero-shot fully-internal approach [32]. We define “Internal learning” in terms of exploiting recurrence of information within a single domain, like within an image [39] or within an video [50]. It can be thought of as a sort of DIP-for-video, a line of work that was started for images with DIP [43] and extended for video by double-DIP [9]. Other methods have embraced this paradigm partially, learning some priors from large external datasets, before learning video-specific priors via internal learning [46].

3. Method

We first give an overview of HNeRV (Sec. 3.1). We explain what makes HNeRV a “hybrid” representation, and the advantages that this offers. We provide architectural details, loss functions, and explain how we compute the size of our video representation. We then give particulars necessary for utilizing HNeRV on downstream tasks (Sec. 3.2). These include vector quantization for video compression and reconstruction loss for inpainting.

3.1. HNeRV overview

Compared with a simple auto-encoder which uses one model for *all* videos, and has *large* frame embeddings, HNeRV (Figure 2 (a)) fits a model on *each* video, and has *tiny* frame embeddings. We utilize ConvNeXt blocks [26] to

build the encoder, and propose novel HNeRV blocks (Figure 2 (b)) to build the decoder.

Hybrid Neural Representation. Compact video representations can be divided into two parts: explicit methods and implicit methods. Explicit methods use an autoencoder to encode and decode all videos, and store content *explicitly* as a latent embedding. Given a *video-specific embedding* as input, the decoder can reconstruct the video. Implicit methods use only a learnable decoder to represent the video. Given fixed frame index as input, the *video-specific decoder* can reconstruct the video. With content-adaptive embedding as input, explicit representation shows better generalization and compression performance, while implicit representations have a much simpler encoding/decoding pipeline and a high potential for compression (benefits from model compression techniques) and other downstream tasks (*e.g.* efficient video dataloader, video denoising, inpainting). In this paper, we propose a hybrid neural representation to combine the advantages of both explicit and implicit methods. Similar to implicit representation, we use a learnable decoder to model video separately and store most content implicitly in the video-specific decoder. To achieve better reconstruction, we use a learnable embedding as input and store information explicitly in these frame-specific embeddings, which is similar to explicit methods. Therefore, we can use any powerful encoder to generate tiny content-adaptive embeddings to boost the performance of implicit representation. Since these embeddings are quite small (*e.g.* a 128-d vector for a 640×1280 frame), our hybrid neural representation is as compact as implicit methods, but with stronger capacity, faster convergence, and better internal generalization, while keeping the full potential for downstream tasks.

Model Architecture. Similar to a NeRV block, an HNeRV block consists of three layers: convolution layer, pixelshuffle layer, and activation layer. Within each block, only the convolution layer has learnable parameters (Figure 2 (b)). Illustrated in Table 1, a NeRV block uses fixed kernel sizes for all stages $K = 3$, and reduces channel width by 2, $C_{\text{out}} = C_{\text{in}}/2$. Therefore, for blocks at later stages, the parameters are quite few and may not be strong enough to store video content at high resolution. In contrast, we increase the kernel size and channel width for later HNeRV blocks, where K increases from 1 (stage 1), to 3 (stage 2), to K_{\max} (5, *etc.* for later stages), and we decrease channel width by a reduction factor r (1.2, *etc.*). With kernel size 1, the first block has much fewer parameters; with larger kernel size and wider channels, HNeRV blocks at later stages are much stronger; and we therefore get a more even distribution of model parameters across layers. We list output sizes of various stages in Figure 2 (d), with embedding dimension d and channel reduction r . Each stage has one block, and we use a 1×1 convolution layer to get low-

Table 1. **HNeRV block vs. NeRV block.** k is kernel size for each stage, C_{out} and C_{in} are output/input channels for each block. We decrease parameters via a small $k = 1$ for first block, and increase parameters for later layers with a larger k and wider channels.

| | |
|--------------|---------------------------------------------------------------|
| NeRV blocks | $k = 3, C_{\text{out}} = C_{\text{in}}/2$ |
| HNeRV blocks | $k = 1, 3, \dots, K_{\max}, C_{\text{out}} = C_{\text{in}}/r$ |

dimension frame embeddings (channel width from c_1 to d), and a 3×3 convolution layer for final image predictions (channel width from c_2/r^4 to 3).

Loss Functions. Since HNeRV attempts to reconstruct video with high fidelity, we use the loss objective $L = \text{Loss}(x, p)$, where x is the input frame, p is the HNeRV prediction, and ‘Loss’ is any reconstruction loss function like L2, L1, or SSIM loss.

Total size. As a hybrid neural representation, we include both frame embedding and decoder parameters to compute the total size of our video representation: $\text{TotalSize} = \text{EmbedSize} + \text{DecoderSize}$.

3.2. Downstream tasks

Video Compression. We leverage both model compression and embedding quantization for video compression. Similar to NeRV, we apply global unstructured pruning, model quantization, and weight entropy encoding for model compression. *Note that unlike NeRV, which only stores non-zero parameters but no mechanism for mapping stored weights to their location in the network (and is thus impractical and unfair for compression comparison) we use entropy encoding to store the sparse weights (details can be found in the appendix).*

For quantization of a vector μ , we linearly map every element to the closest integer,

$$\begin{aligned} \mu_i &= \text{Round} \left(\frac{\mu_i - \mu_{\min}}{\text{scale}} \right) * \text{scale} + \mu_{\min}, \text{ where} \\ \text{scale} &= \frac{\mu_{\max} - \mu_{\min}}{2^b - 1}, \end{aligned} \quad (1)$$

μ_i is vector element, ‘Round’ is a function that rounds to the closest integer, b is the quantization bit length, μ_{\max} and μ_{\min} are the max and min value of vector μ , and ‘scale’ is the scaling factor.

Video Inpainting. For partially distorted video, we only compute loss for non-masked pixels,

$$L_{\text{inpainting}} = (1 - M) * \text{Loss}(x, p) \quad (2)$$

where M is the mask matrix where distorted pixels are 1 and other are 0. For inpainting output, following IIIVI [32], we fill the masked region with HNeRV’s output.

4. Experiments

We first provide information necessary for replicating our results, including datasets used and hyperparameter set-

Table 2. Video regression with different **sizes**

| Size | 0.35M | 0.75M | 1.5M | 3M | avg. |
|--------|--------------|--------------|--------------|--------------|--------------|
| NeRV | 26.99 | 28.46 | 30.87 | 33.21 | 29.88 |
| E-NeRV | 27.84 | 30.95 | 32.09 | 36.72 | 31.90 |
| HNeRV | 30.15 | 32.81 | 35.57 | 37.43 | 33.90 |

Table 3. Video regression with different **epochs**

| Epoch | 300 | 600 | 1200 | 1800 | 2400 | 3600 |
|--------|--------------|--------------|--------------|--------------|--------------|--------------|
| NeRV | 28.46 | 29.15 | 29.57 | 29.73 | 29.77 | 29.86 |
| E-NeRV | 30.95 | 32.07 | 32.79 | 33.1 | 33.36 | 33.67 |
| HNeRV | 32.81 | 33.89 | 34.51 | 34.73 | 34.88 | 35.03 |

Table 4. Video regression at resolution **960×1920**, PSNR↑ reported

| Video | beauty | swan | bmx | bosph | dance | camel | bee | jockey | ready | shake | yach | avg. |
|-------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|
| NeRV | 33.25 | 28.48 | 27.86 | 33.22 | 26.45 | 24.81 | 37.26 | 31.74 | 24.84 | 33.08 | 28.30 | 29.94 |
| HNeRV | 33.58 | 30.35 | 29.98 | 34.73 | 30.45 | 26.71 | 38.96 | 32.04 | 25.74 | 34.57 | 29.26 | 31.49 |

Table 5. Video regression at resolution **480×960**, PSNR↑ reported

| Video | beauty | swan | bmx | bosph | dance | camel | bee | jockey | ready | shake | yach | avg. |
|-------|--------------|--------------|--------------|--------------|--------------|--------------|-------|--------------|--------------|--------------|--------------|--------------|
| NeRV | 36.27 | 29.75 | 28.81 | 35.07 | 29.47 | 26.75 | 40.76 | 32.58 | 25.81 | 35.33 | 30.11 | 31.88 |
| HNeRV | 36.91 | 31.92 | 31.27 | 36.95 | 33.85 | 28.85 | 42.05 | 33.33 | 27.07 | 36.97 | 30.96 | 33.65 |

tings (Sec. 4.1), and then show main results for video regression (Sec. 4.2). We show the effectiveness of decoder-side parameter-redistribution for improving the appearance of high resolution frames (Sec. 4.3). Finally, we demonstrate compelling initial results for downstream tasks including decoding speed, video compression, internal generalization, and video inpainting (Sec. 4.4). We offer results for extensive ablation studies in Appendix A.

4.1. Dataset and Implementation Details

We use the Big Buck Bunny (Bunny) [1], UVG [30] and DAVIS [45] datasets. Bunny has 132 frames with resolution 720×1280 , and we center-crop 640×1280 to get tiny spatial size (*e.g.* 1×2) for embedding. UVG has 7 videos ¹ with size 1080×1920 at FPS 120 of 5s or 2.5s, and we center-crop 960×1920 . We also take 10 videos ² from the DAVIS validation subset (1080×1920 , 50-200 frames) and center crop the 960×1920 . Unless otherwise specified, we use the Adam optimizer, with beta as $(0.9, 0.999)$, weight decay as 0, and learning rate at 0.001 with cosine learning rate decay. We also use batch size as 2 and L2 loss as reconstruction loss function. K_{\max} is set as 5, reduction r is set as 1.2 in Table 1. We set stride list as $(5, 4, 4, 2, 2)$, $(5, 4, 3, 2, 2)$, and $(5, 4, 4, 3, 2)$ for video resolutions of 640×1280 , 480×960 , and 960×1920 respectively.

For evaluation metrics, we use PSNR and MS-SSIM to evaluate reconstruction quality, bits per pixel (bpp) for compression, and pixels per pixel (ppp) for model compactness. We conduct all experiments in Pytorch with RTX2080ti

¹Beauty, Bosphorus, HoneyBee, Jockey, ReadySetGo, ShakeNDry, YachtRide

²bike-packing, blackswan, bmx-trees, breakdance, camel, car-round, car-shadow, cows, dance-twirl, dog

GPUs, where it takes around 8s per epoch to train a 130 frame video of size 640×1280 . We choose HNeRV’s size to ensure the PSNR lies between 30-40 for fair video reconstruction. We provide more experiment details such as architecture details, qualitative results, exact numerical results corresponding to plots, and per-video compression results, in the supplementary material.

4.2. Main Results

We first compare HNeRV with implicit methods NeRV and E-NeRV on Bunny. For fair comparison, we scale channel width to make total size comparable as NeRV did. In Table 2, with the same size and 300 epochs, HNeRV outperforms both NeRV and E-NeRV. We also show comparison of different training time in Table 3 with 0.75M size and in Figure 1 (right) with 1.5M size, where HNeRV converges much faster compared to implicit methods. We show improvements qualitatively as well in Figure 3. As a compact representation, HNeRV reconstructs the video well with only 0.35M parameters, at 0.003 ppp. We also evaluate it on 7 UVG videos and 4 DAVIS videos, where HNeRV shows large improvements at resolution 960×1920 in Table 4, and its resized 480×960 version in Table 5, with size 3M and 300 epochs.

4.3. Parameter Distribution Analysis

As part of our novel architectural innovation, we balance the parameters in the HNeRV decoder. While NeRV-like architectures naturally have a vanishingly small number of parameters in the final layers, Fig. 4 shows how we adjust such that the initial and final layers have a roughly equal number of parameters. Table 6 demonstrates how more even parameter distribution achieves optimal results not only for HN-

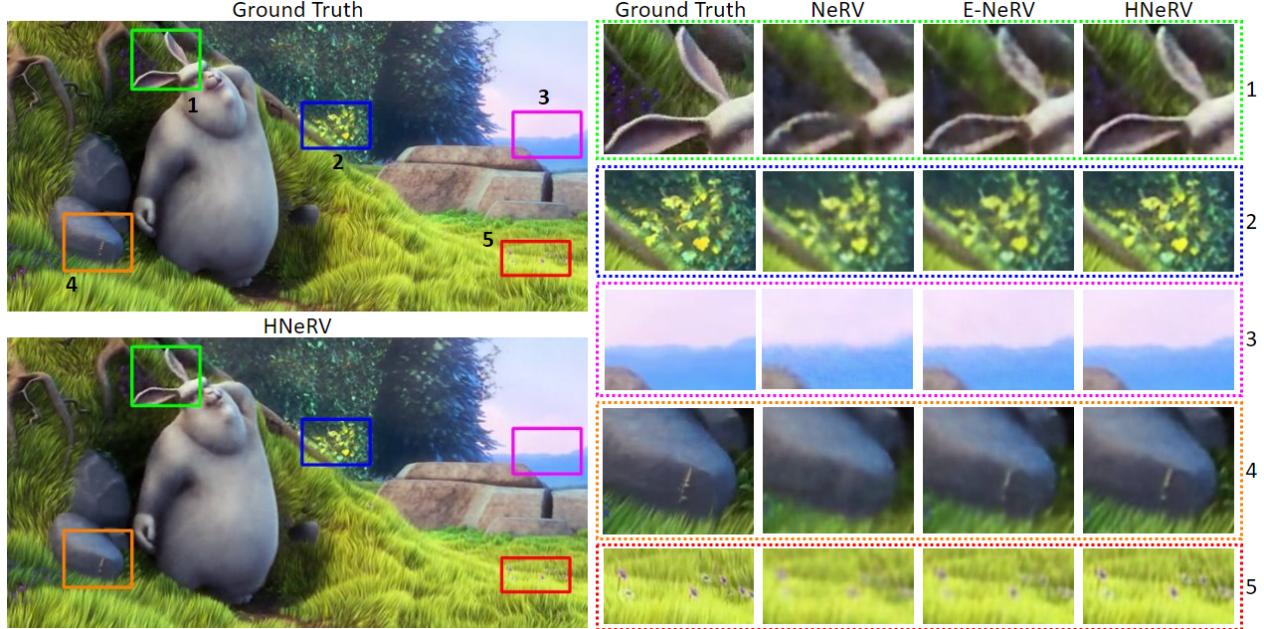


Figure 3. **Visualization of video neural representations** at 0.003 ppp, which means the total size is only about 0.3% of the original video size. On the **left**, we compare HNeRV to ground truth. On the **right**, we compare NeRV, E-NeRV, and HNeRV for 5 patches.

Table 6. Analysis of parameter rebalancing.

| K | r | NeRV | | HNeRV | |
|-----|-----|--------------|---------------|--------------|---------------|
| | | PSNR | MS-SSIM | PSNR | MS-SSIM |
| 3,3 | 2 | 30.87 | 0.9341 | 29.91 | 0.9203 |
| 3,3 | 1.2 | 32.27 | 0.9496 | 33.09 | 0.9587 |
| 1,5 | 2 | 31.34 | 0.9399 | 34.32 | 0.9715 |
| 1,5 | 1.2 | 33.03 | 0.9573 | 35.57 | 0.9773 |

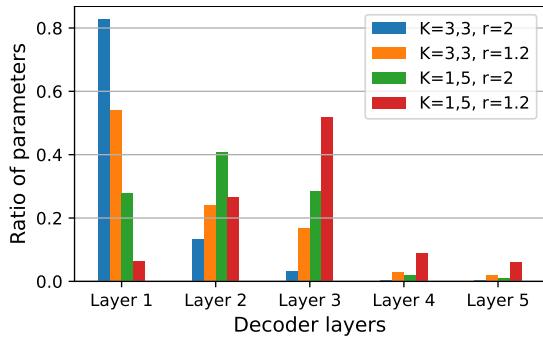


Figure 4. Parameter distributions for decoder blocks. See Tab. 6 for PSNR and MS-SSIM results with these 4 settings.

eRV, but also for NeRV³: setting K_{min} to 1 and K_{max} to 5, with r at 1.2, maximizes PSNR and MS-SSIM for both architectures. Tab. 9 and Tab. 10 further verify that these hyperparameters are optimal for HNeRV. We offer these results as strong evidence that our parameter balancing approach not only enables the success of HNeRV, but would be broadly applicable to other NeRV-like architectures.

³We label NeRV with $K_{max}=5$ as (1,5) since its FC layer can be seen as a 1×1 convolution layer.

4.4. Downstream Tasks

Video decoding. We evaluate video decoding on Bunny with channel reduction r as 1.5, where H.264 and H.265 are tested with 4 CPUs⁴, while DCVC [21] and HNeRV are tested with 1 GPU (RTX2080ti). We only measure the forward time for DCVC and HNeRV. We compare video decoding at various reconstruction qualities (PSNR at 32, 35, and 37) in Figure 5 (left), where HNeRV outperforms traditional codecs (H.264 and H.265) and learning-based DCVC. Note that although many prior learning-based compression methods show bit-distortion improvements, their decoding speeds *lag far behind* traditional codecs and neural representation. Besides, most compression methods encode and decode frames in an auto-regressive way and can not access frames randomly. Compared to these methods, the decoding of HNeRV is much simpler and can be deployed to any platform easily. We compare decoding time in Figure 5 (Middle) (PSNR at 35) where 100%, 50%, and 25% of frames (evenly sampled, to mimic *e.g.* a discrete reduction of FPS) are decoded. Since there is no dependency among video frames, HNeRV can decode them in parallel and decoding time decreases linearly with the number of frames decoded. In contrast, H.264 and H.265 still need to decode most frames, even though only some of them are needed. Finally, we compare with implicit methods in Figure 5 (right), where HNeRV is slightly slower than NeRV since the computation of later layers is more expensive due to large K and channel width. As a hybrid neural representation, HNeRV achieves much better trade-offs with respect

⁴Intel(R) Xeon(R) Silver 4216 CPU @ 2.10GHz

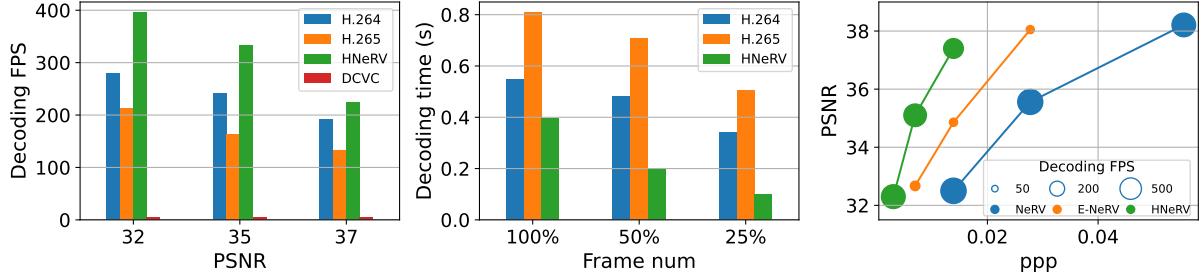


Figure 5. **Video decoding.** **Left:** HNeRV outperforms traditional video codecs H.264 and H.265, and learning-based compression method DCVC. **Middle:** HNeRV shows much better flexibility when decoding only a portion of video frames. **Right:** HNeRV performs well for compactness (ppp), reconstruction quality (PSNR), and decoding speed (FPS).

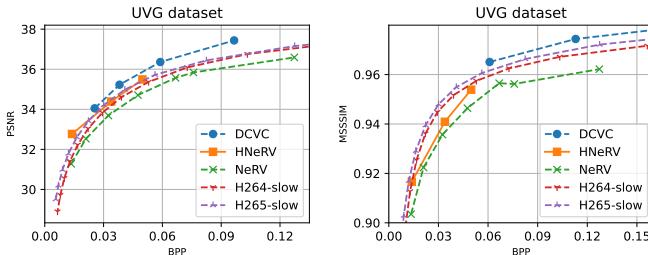


Figure 6. **Compression results on UVG dataset.**

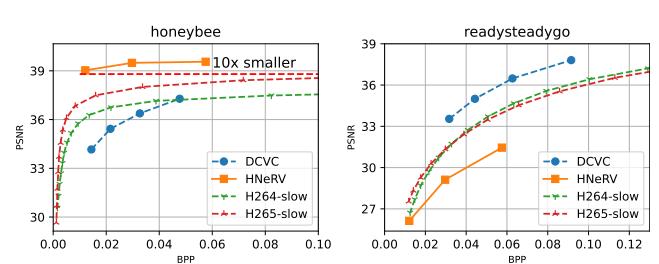


Figure 7. **Best/worst compression cases from UVG dataset.**

Table 7. **Internal generalization** results. NeRV, E-NeRV, and HNeRV use interpolated embeddings as input, HNeRV \dagger uses held-out frames as input. With content-adaptive embeddings as input, HNeRV shows much better reconstruction on held-out frames

| Method | beauty | swan | bmx | bosph | dance | camel | bee | jockey | ready | shake | yach | avg. |
|-----------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|-------|-------|--------------|
| NeRV | 28.05 | 17.94 | 15.55 | 30.04 | 16.99 | 14.83 | 36.99 | 20.00 | 17.02 | 29.15 | 24.50 | 22.82 |
| HNeRV | 30.97 | 21.44 | 17.35 | 34.38 | 20.2 | 19.93 | 38.83 | 23.67 | 20.90 | 32.69 | 27.30 | 26.15 |
| HNeRV \dagger | 31.10 | 21.97 | 18.29 | 34.38 | 20.29 | 20.64 | 38.83 | 23.82 | 20.99 | 32.61 | 27.24 | 26.38 |

Table 8. Video **inpainting** results (PSNR \uparrow) with 5 fixed box masks on input videos. ‘Input’ is the baseline of mask video and ground truth

| Video | bike | b-swan | bmx | b-dance | camel | c-round | c-shadow | cows | dance-twirl | dog | avg. |
|-------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|-------------|--------------|
| Input | 23.14 | 20.24 | 19.99 | 21.36 | 17.3 | 20.47 | 18.92 | 19.37 | 20.45 | 18.39 | 19.96 |
| NeRV | 30.94 | 33.43 | 32.07 | 27.82 | 31.99 | 29.09 | 31.63 | 30.08 | 30.45 | 33.85 | 31.14 |
| IIVI | 31.87 | 36.02 | 34.36 | 27.63 | 35.11 | 32.61 | 33.69 | 31.26 | 31.44 | 35.7 | 32.97 |
| HNeRV | <u>31.27</u> | <u>34.24</u> | <u>33.95</u> | <u>27.94</u> | <u>32.21</u> | <u>30.88</u> | <u>33.07</u> | <u>30.82</u> | <u>31.21</u> | <u>34.7</u> | <u>32.03</u> |

to compactness (ppp), reconstruction quality (PNSR), and decoding speed (FPS).

Video compression. With model pruning (10% pruned), embedding quantization (8 bits), model quantization (8 bits), and model entropy encoding (8% saved), we show video compression results on UVG in Figure 6. HNeRV outperforms the implicit method, NeRV, and traditional video codecs H.264 and H.265. Note that HNeRV achieves this using a small model for each video, while NeRV fits a big model on concat videos (for better compression) which greatly slows down the encoding and decoding speed. We also show the best and worst cases of HNeRV video compression for the ‘honeybee’ and ‘readysteadygo’ videos re-

spectively in Figure 7, where HNeRV achieves outstanding performance when the camera is not moving, such as with the ‘honeybee’ video (10 \times smaller than H.265 for equivalent PSNR). Given the limited performance on videos of highly dynamic scenes, we propose finding a good size and network architecture for such videos as future work.

Internal generalization. Since HNeRV leverages content-adaptive embeddings, we also evaluate it for the video interpolation task. Holding out every other frame as a test set, NeRV, E-NeRV, and HNeRV use interpolated embedding as input, while HNeRV \dagger uses the test frame as input. With learnable and content-adaptive embedding, our HNeRV shows much better generalization, quantitatively in Ta-

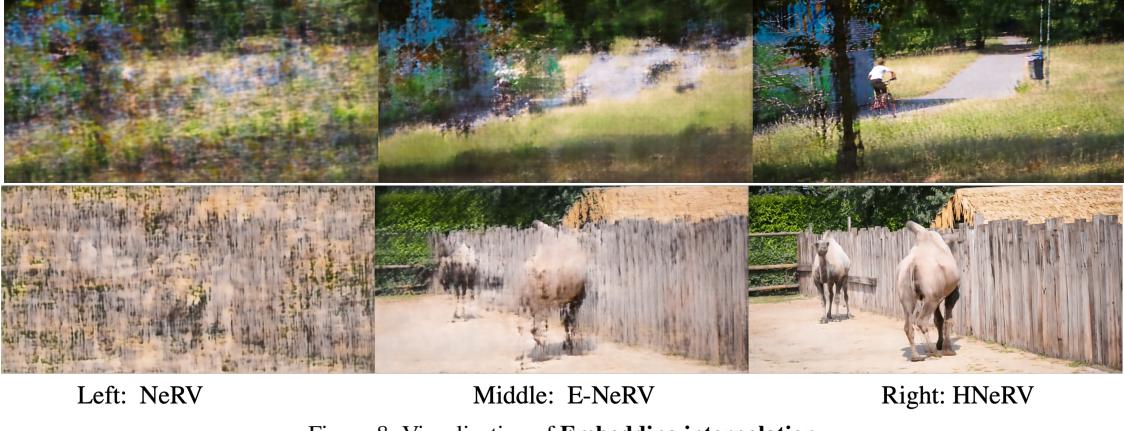


Figure 8. Visualization of **Embedding interpolation**.



Figure 9. **Inpainting** results of fixed masks and object masks. **Left**) input frame; **Right**) HNeRV output.

ble 7 and qualitatively in Figure 8.

Video inpainting. We also explore video inpainting with fixed and object masks. For fixed masks, we use 5 boxes of width 50 (Figure 9 (top)) and show quantitative results in Table 8 where HNeRV improves inpainting performance over the implicit method, NeRV. Although we do not have any specific design for the inpainting task, HNeRV even achieves comparable performance with an SOTA inpainting method, IIVI [32]. We show qualitative results in Figure 9.

5. Conclusion

In this paper, we propose a hybrid neural representation for videos (HNeRV). With content-adaptive embedding and evenly-distributed parameters, HNeRV improves video regression performance compared to implicit methods in terms of reconstruction quality, convergence speed, and in-

ternal generalization. As a video representation, HNeRV is also simple, fast, and flexible for video decoding, and shows good performance for video compression and inpainting.

There are many limitations of HNeRV as well. Firstly, as a neural representation, HNeRV stores each video as a neural network. Given a new video, HNeRV still needs time to train to fit the video. Secondly, although HNeRV can represent a video well, finding a best-fit embedding size, model size, and network architecture (K_{\max} , r , etc.) remains an open problem. Finally, although increasing kernel sizes and channel widths at later layers largely improves the regression performance, it slightly slows down the network, as shown in Figure 5 (right).

Acknowledgements. This project was partially funded by the DARPA SAIL-ON (W911NF2020009) program, an independent grant from Facebook AI, and Amazon Research Award to AS.

References

- [1] Big buck bunny, sunflower version. <http://bbb3d.renderfarming.net/download.html>. Accessed: 2010-09-30. 5
- [2] 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 *CVPR*, June 2020. 1, 3
- [3] N. Ahmed, T. Natarajan, and K.R. Rao. Discrete cosine transform. *IEEE Transactions on Computers*, C-23(1):90–93, 1974. 1, 2
- [4] Hao Chen, Bo He, Hanyu Wang, Yixuan Ren, Ser-Nam Lim, and Abhinav Shrivastava. NeRV: Neural representations for videos. In *NeurIPS*, 2021. 1, 2, 3
- [5] Hao Chen, A Gwilliam Matthew, Bo He, Ser-Nam Lim, and Abhinav Shrivastava. CNerV: Content-adaptive neural representation for visual data. In *BMVC*, 2022. 2
- [6] Zhiqin Chen and Hao Zhang. Learning implicit fields for generative shape modeling. In *CVPR*, 2019. 1, 2
- [7] Abdelaziz Djelouah, Joaquim Campos, Simone Schaub-Meyer, and Christopher Schroers. Neural inter-frame compression for video coding. In *ICCV*, 2019. 1, 3
- [8] Emilien Dupont, Adam Golinski, Milad Alizadeh, Yee Whye Teh, and Arnaud Doucet. Coin: Compression with implicit neural representations. In *ICLR workshop*, 2021. 1
- [9] Yosef Ganelsman, Assaf Shocher, and Michal Irani. “double-dip”: Unsupervised image decomposition via coupled deep-image-priors. In *CVPR*, June 2019. 3
- [10] Suyog Gupta, Ankur Agrawal, Kailash Gopalakrishnan, and Pritish Narayanan. Deep learning with limited numerical precision. *CoRR*, abs/1502.02551, 2015. 3
- [11] Amirhossein Habibian, Ties van Rozendaal, Jakub M. Tomczak, and Taco S. Cohen. Video compression with rate-distortion autoencoders. In *ICCV*, 2019. 1, 3
- [12] Song Han, Huizi Mao, and William J Dally. Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. *ICLR*, 2016. 3
- [13] Song Han, Jeff Pool, John Tran, and William Dally. Learning both weights and connections for efficient neural network. In *NeurIPS*, pages 1135–1143, 2015. 3
- [14] Bo He, Xitong Yang, Hanyu Wang, Zuxuan Wu, Hao Chen, Shuaiyi Huang, Yixuan Ren, SerNam Lim, and Abhinav Shrivastava. Towards scalable neural representation for diverse videos. In *CVPR*, 2023. 2
- [15] David A Huffman. A method for the construction of minimum-redundancy codes. *Proceedings of the IRE*, 40(9), 1952. 3, 1
- [16] Benoit Jacob, Skirmantas Kligys, Bo Chen, Menglong Zhu, Matthew Tang, Andrew Howard, Hartwig Adam, and Dmitry Kalenichenko. Quantization and training of neural networks for efficient integer-arithmetic-only inference. In *CVPR*, June 2018. 3
- [17] Mehrdad Khani, Vibhaalakshmi Sivaraman, and Mohammad Alizadeh. Efficient video compression via content-adaptive super-resolution. *ICCV*, 2021. 3
- [18] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 2
- [19] Raghuraman Krishnamoorthi. Quantizing deep convolutional networks for efficient inference: A whitepaper. *arXiv preprint arXiv:1806.08342*, 2018. 3
- [20] Didier Le Gall. Mpeg: A video compression standard for multimedia applications. *Commun. ACM*, 1991. 3
- [21] Jiahao Li, Bin Li, and Yan Lu. Deep contextual video compression. *NeurIPS*, 34, 2021. 6
- [22] Zizhang Li, Mengmeng Wang, Huaijin Pi, Kechun Xu, Jianbiao Mei, and Yong Liu. E-nerv: Expedite neural video representation with disentangled spatial-temporal context. *ECCV*, 2022. 2
- [23] Haojie Liu, Tong Chen, Ming Lu, Qiu Shen, and Zhan Ma. Neural video compression using spatio-temporal priors. *arXiv preprint arXiv:1902.07383*, 2019. 1, 3
- [24] Haojie Liu, Ming Lu, Zhan Ma, Fan Wang, Zhihuang Xie, Xun Cao, and Yao Wang. Neural video coding using multiscale motion compensation and spatiotemporal context model. *IEEE Transactions on Circuits and Systems for Video Technology*, 2021. 1
- [25] Jerry Liu, Shenlong Wang, Wei-Chiu Ma, Meet Shah, Rui Hu, Pranaab Dhawan, and Raquel Urtasun. Conditional entropy coding for efficient video compression. In *ECCV*, 2020. 3
- [26] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *CVPR*, 2022. 3
- [27] Guo Lu, Wanli Ouyang, Dong Xu, Xiaoyun Zhang, Chunlei Cai, and Zhiyong Gao. Dvc: An end-to-end deep video compression framework. In *CVPR*, 2019. 1
- [28] Shishira R Maiya, Sharath Girish, Max Ehrlich, Hanyu Wang, Kwot Sin Lee, Patrick Poirson, Pengxiang Wu, Chen Wang, and Abhinav Shrivastava. Nirvana: Neural implicit representations of videos with adaptive networks and autoregressive patch-wise modeling. In *CVPR*, 2023. 2
- [29] Ishit Mehta, Michaël Gharbi, Connelly Barnes, Eli Shechtman, Ravi Ramamoorthi, and Manmohan Chandraker. Modulated periodic activations for generalizable local functional representations. In *ICCV*, 2021. 2
- [30] 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*, 2020. 5
- [31] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020. 2
- [32] Hao Ouyang, Tengfei Wang, and Qifeng Chen. Internal video inpainting by implicit long-range propagation. In *ICCV*, 2021. 3, 4, 8
- [33] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. DeepSDF: Learning continuous signed distance functions for shape representation. In *CVPR*, 2019. 1, 2
- [34] Yunchen Pu, Zhe Gan, Ricardo Henao, Xin Yuan, Chunyuan Li, Andrew Stevens, and Lawrence Carin. Variational autoencoder for deep learning of images, labels and captions. In *NeurIPS*, 2016. 2

- [35] Nasim Rahaman, Aristide Baratin, Devansh Arpit, Felix Draxler, Min Lin, Fred Hamprecht, Yoshua Bengio, and Aaron Courville. On the spectral bias of neural networks. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *ICML*, volume 97 of *Proceedings of Machine Learning Research*, pages 5301–5310. PMLR, 09–15 Jun 2019. 1
- [36] Oren Rippel, Alexander G. Anderson, Kedar Tatwawadi, Sanjay Nair, Craig Lytle, and Lubomir Bourdev. Elf-vc: Efficient learned flexible-rate video coding. In *ICCV*, 2021. 1, 3
- [37] Oren Rippel, Sanjay Nair, Carissa Lew, Steve Branson, Alexander G. Anderson, and Lubomir Bourdev. Learned video compression. In *ICCV*, 2019. 1, 3
- [38] Katja Schwarz, Yiyi Liao, Michael Niemeyer, and Andreas Geiger. Graf: Generative radiance fields for 3d-aware image synthesis. In *NeurIPS*, 2020. 1, 2
- [39] Assaf Shocher, Nadav Cohen, and Michal Irani. “zero-shot” super-resolution using deep internal learning. In *CVPR*, June 2018. 3
- [40] Vincent Sitzmann, Julien N. P. Martel, Alexander W. Bergman, David B. Lindell, and Gordon Wetzstein. Implicit neural representations with periodic activation functions. In *NeurIPS*, 2020. 1, 2
- [41] 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*, 2012. 1, 3
- [42] Matthew Tancik, Pratul P. Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan T. Barron, and Ren Ng. Fourier features let networks learn high frequency functions in low dimensional domains. In *NeurIPS*, 2020. 2
- [43] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Deep image prior. In *CVPR*, pages 9446–9454, 2018. 3
- [44] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. ICML. Association for Computing Machinery, 2008. 2
- [45] Haiqiang Wang, Weihao Gan, Sudeng Hu, Joe Yuchieh Lin, Lina Jin, Longguang Song, Ping Wang, Ioannis Katsavounidis, Anne Aaron, and C-C Jay Kuo. Mcl-jcv: a jnd-based h.264/avc video quality assessment dataset. In *2016 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2016. 5
- [46] Tengfei Wang, Hao Ouyang, and Qifeng Chen. Image inpainting with external-internal learning and monochromic bottleneck. In *CVPR*, pages 5120–5129, 2021. 3
- [47] T. Wiegand, G.J. Sullivan, G. Bjontegaard, and A. Luthra. Overview of the h.264/avc video coding standard. *IEEE Transactions on Circuits and Systems for Video Technology*, 2003. 1
- [48] 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*, 2003. 3
- [49] Chao-Yuan Wu, Nayan Singhal, and Philipp Krahenbuhl. Video compression through image interpolation. In *ECCV*, 2018. 1, 3
- [50] Haotian Zhang, Long Mai, Ning Xu, Zhaowen Wang, John Collomosse, and Hailin Jin. An internal learning approach to video inpainting. In *ICCV*, pages 2720–2729, 2019. 3

HNeRV: A Hybrid Neural Representation for Videos

Supplementary Material

A. Ablation study

We show the effectiveness of even-distributed parameters in Table 9 and Table 10 by increasing kernel size and channel width of later layers. For the NeRV block, it uses fixed $K = 3$, and channel reduction factor $r = 2$. We also show an embeddings ablation study, for spatial size ($h \times w$) in Table 11 and embedding dimensions (d) in Table 12.

Table 9. Kernel size (K_{\min}, K_{\max}) ablation, (with $r=1.2$)

| K | PSNR | MS-SSIM |
|-----|--------------|---------------|
| 1,3 | 35.02 | 0.9752 |
| 1,5 | 35.57 | 0.9773 |
| 1,7 | 35.07 | 0.9757 |
| 3,3 | 33.09 | 0.9587 |

Table 10. Channel reduction r ablation, (with $K=1,5$)

| r | PSNR | MS-SSIM |
|-----|--------------|---------------|
| 1 | 34.96 | 0.9745 |
| 1.2 | 35.57 | 0.9773 |
| 1.5 | 34.98 | 0.9762 |
| 2 | 34.32 | 0.9715 |

Table 11. Embedding spatial size ablation

| $h \times w$ | PSNR | MS-SSIM |
|--------------|--------------|---------------|
| 1×2 | 34.79 | 0.9735 |
| 2×4 | 35.57 | 0.9773 |
| 4×8 | 35.12 | 0.9761 |

Table 12. Embedding dimension ablation

| d | PSNR | MS-SSIM |
|-----|--------------|---------------|
| 8 | 35.13 | 0.9770 |
| 16 | 35.57 | 0.9773 |
| 32 | 35.08 | 0.9758 |

B. Video decoding

We firstly show command to evaluate decoding speed of H.264 and H.265:

ffmpeg -threads ThreadsNum -i Video -preset medium -f null -benchmark -

And we also show quantitative decoding results in Table 13, 14, and Table 15. In Table 15, we can further increase video decoding speed with a smaller channel width (*i.e.* a big reduction factor $r = 2$).

C. Video compression

Then we show the details for downstream tasks of video compression, which can be divided into three steps: global unstructure pruning, quantization, and entropy encoding.

1) *Model Pruning*. Given a pre-trained model, we use global unstructured pruning to reduce the model size, where parameters below a threshold are pruned and set as zero.

For a model parameter θ_i , $\theta_i = \begin{cases} \theta_i, & \text{if } \theta_i \geq \theta_q \\ 0, & \text{otherwise,} \end{cases}$ where θ_q is the q percentile value for all model parameters θ . As a

Table 13. Decoding FPS \uparrow

| PSNR | 32 | 35 | 37 |
|-------|--------------|--------------|--------------|
| H.264 | 279.7 | 240.9 | 192.7 |
| H.265 | 211.9 | 163.2 | 132.5 |
| DCVC | 4.7 | 4.6 | 4.5 |
| HNeRV | 395.9 | 332.7 | 224.8 |

Table 14. Decoding time (s) \downarrow

| # Frames | 100% | 50% | 25% |
|----------|--------------|--------------|--------------|
| H.264 | 0.548 | 0.480 | 0.343 |
| H.265 | 0.809 | 0.708 | 0.506 |
| DCVC | 27.913 | 24.424 | 17.446 |
| HNeRV | 0.397 | 0.198 | 0.099 |

Table 15. HNeRV Decoding FPS

| PSNR | 32 | 35 | 37 |
|--------|--------------|--------------|--------------|
| r=1.5 | 395.9 | 332.7 | 224.8 |
| r=1.75 | 397.4 | 373.8 | 320.7 |
| r=2 | 405.5 | 383.3 | 350.5 |

normal practice, we fine-tune the model to regain the representation after pruning.

2) *Model and embedding quantization*. Model quantization and embedding quantization follow the same scheme. Given an vector μ , we linearly map every element to the closest integer,

$$\mu_i = \text{Round} \left(\frac{\mu_i - \mu_{\min}}{\text{scale}} \right) * \text{scale} + \mu_{\min}, \text{ where} \quad (3)$$

$$\text{scale} = \frac{\mu_{\max} - \mu_{\min}}{2^b - 1}$$

μ_i is one vector element, ‘Round’ is a function that rounds to the closest integer, ‘ b ’ is the bit length for quantization, μ_{\max} and μ_{\min} are the max and min value of vector μ , and ‘scale’ is the scaling factor. For scaling factor and zero points at this step, we can also try other methods instead of current min-max one, like choosing 2^b evenly-distributed values to minimum the mean square error.

3) *Entropy encoding*. Finally, we use entropy encoding to further reduce the size. Specifically, we leverage Huffman coding [15] for quantized weights and get lossless compression.

D. Weight Pruning for Model Compression.

We appreciate this concern, which has been unresolved since the original NeRV paper. By applying entropy encoding (assigning fewer bits for frequent symbols), we can store pruned weights with limited bits, since all pruned weights share a frequent symbol: 0. We provide corrected model compression results in Tab. 16, and will update the paper accordingly. We use 2 baselines (models with no

pruning) – one where the model is only quantized, and another where we apply entropy encoding after quantization. As we prune more parameters, entropy encoding enables us to use fewer bits to store the sparse model weights.

Table 16. Compression results. “Size ratio” compares to model with quant. only, and “Sparsity” indicates amount of weights pruned.

| Compression | Quant | Prune + Quant + Entropy coding | | | | |
|-------------|--------|--------------------------------|--------|--------|-------|-------|
| | | 0% | 10% | 15% | 20% | 25% |
| PSNR | 37.61 | 37.56 | 37.51 | 37.32 | 37.02 | 36.61 |
| Size (bits) | 11.54M | 10.94M | 10.41M | 10.09M | 9.77M | 9.36M |
| Size ratio | 100% | 94.8% | 90.2% | 87.4% | 84.7% | 81.1% |

E. HNeRV architecture details

We also provide architecture details for HNeRV models in various tasks and datasets in Table 17, with total size, strides list, encoder dimension $c1$, embedding dimension d , channel width of decoder input $c2$, channel reduction r , lowest channel width Ch_{min} , min and max kernel size K_{min} , K_{max} .

Table 17. HNeRV architecture details

| Video size | size | strides | c1 | d | c2 | r | Ch _{min} | K _{min} , K _{max} |
|------------|------|-----------|----|----|-----|-----|-------------------|-------------------------------------|
| 640×1280 | 0.35 | 5,4,4,2,2 | 64 | 16 | 32 | 1.2 | 12 | 1,5 |
| 640×1280 | 0.75 | 5,4,4,2,2 | 64 | 16 | 48 | 1.2 | 12 | 1,5 |
| 640×1280 | 1.5 | 5,4,4,2,2 | 64 | 16 | 68 | 1.2 | 12 | 1,5 |
| 640×1280 | 3 | 5,4,4,2,2 | 64 | 16 | 97 | 1.2 | 12 | 1,5 |
| 480×960 | 3 | 5,4,3,2,2 | 64 | 16 | 110 | 1.2 | 12 | 1,5 |
| 960×1920 | 3 | 5,4,4,3,2 | 64 | 16 | 92 | 1.2 | 12 | 1,5 |

F. Per-video compression results

We also show video compression results for **UVG** videos in Figure 10.

G. More visualizations

We show more visualizations for video regression (Figure 11), video interpolation (Figure 12), and video inpainting (Figure 13).

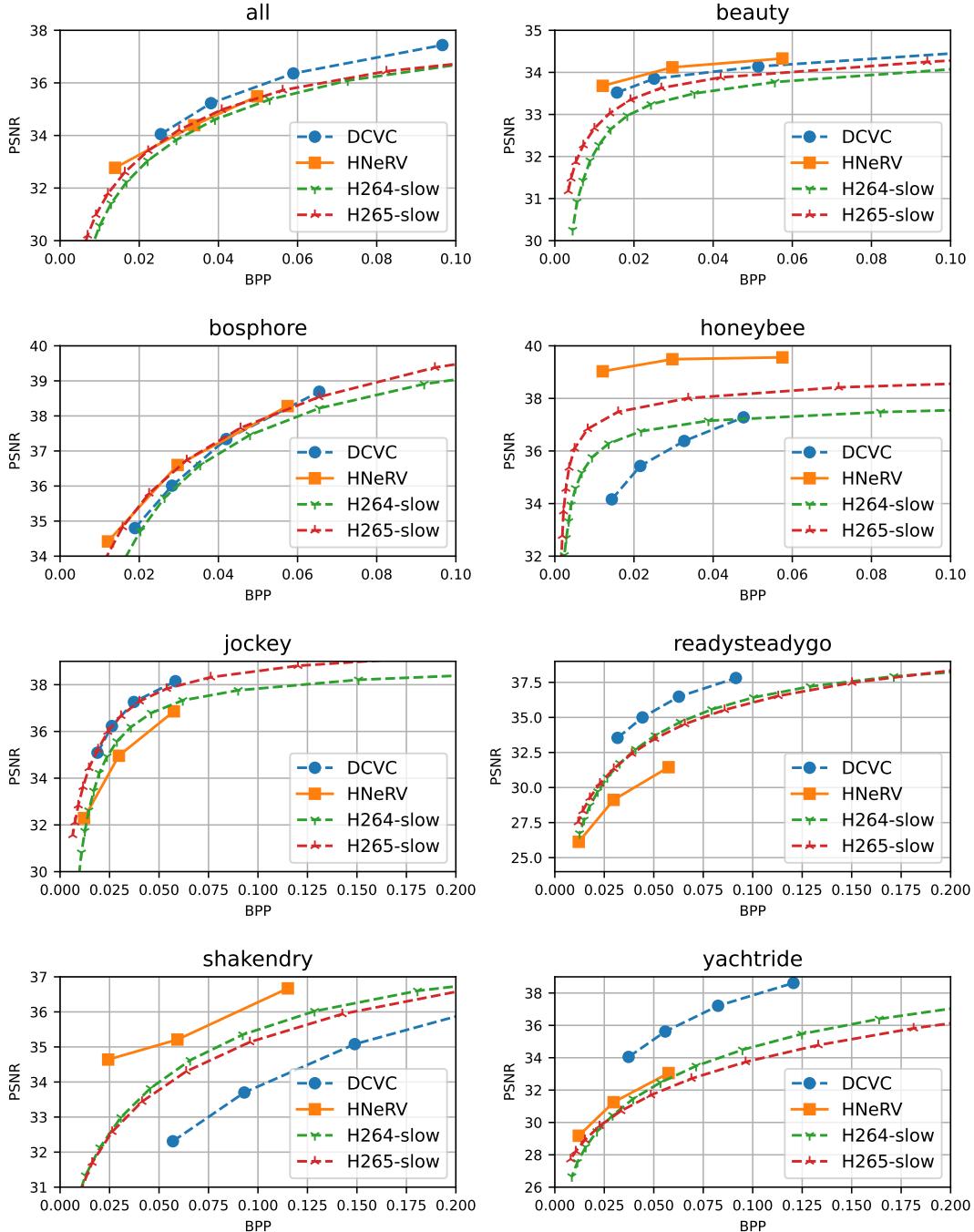


Figure 10. **Compression** results averaged across all **UVG** videos, and for each specific videos.

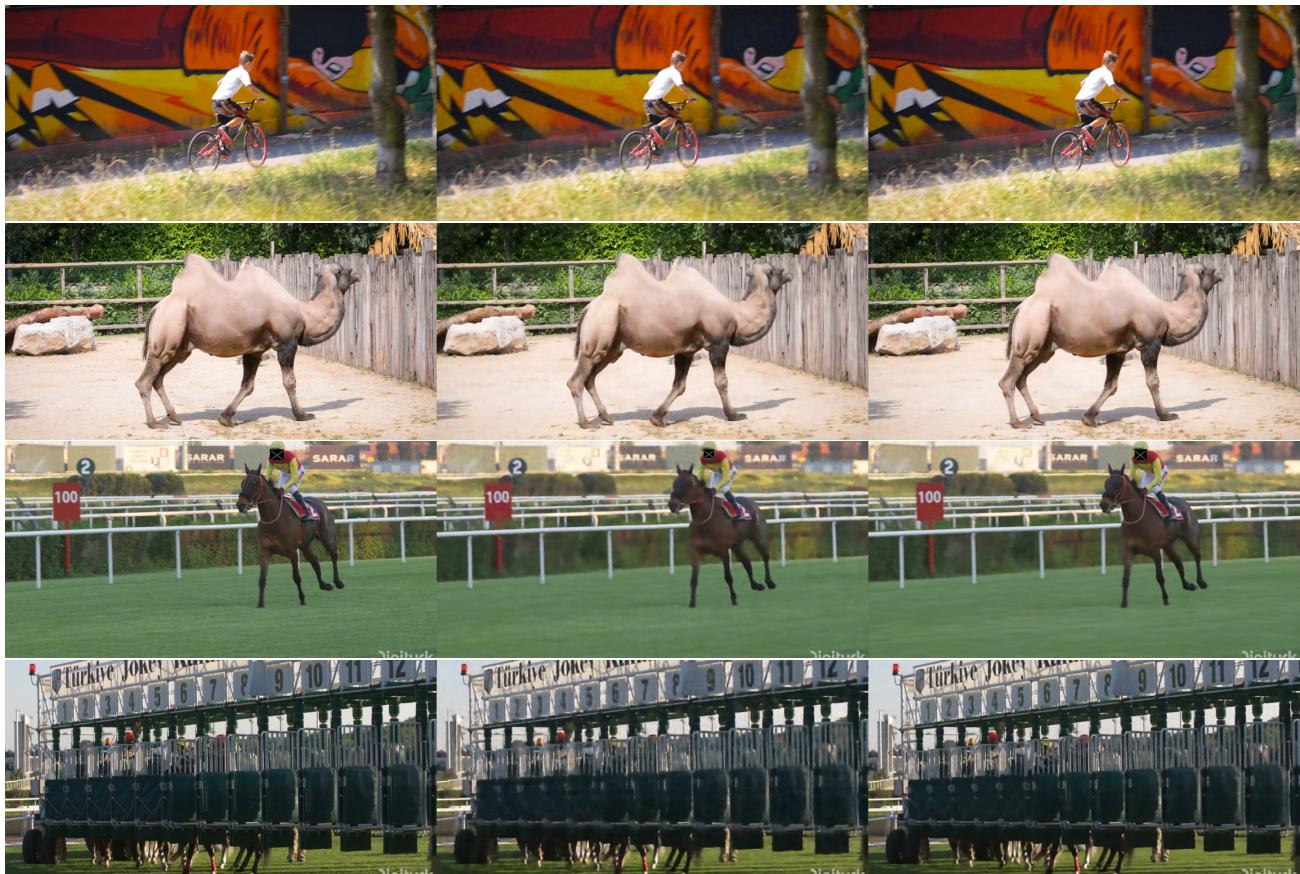


Figure 11. **Video regression** results. **Left)** ground truth. **Middle)** NeRV output. **Right)** HNeRV output.

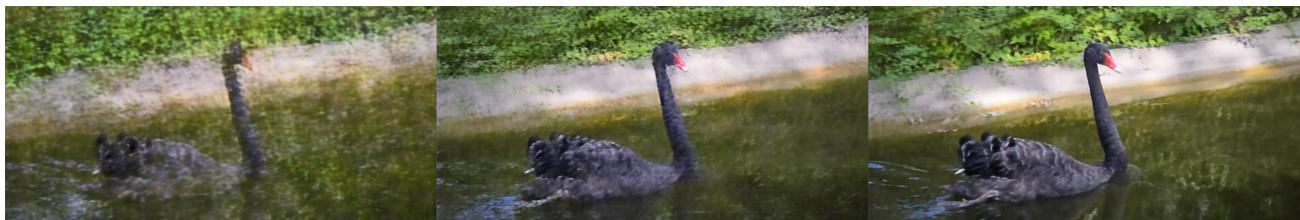


Figure 12. **Interpolation** results.



Figure 13. **Inpainting** results.