HFS-HNeRV: High-Frequency Spectrum Hybrid Neural Representation for Videos

Jianhua Zhao
Department of Electrical and Electronic Engineering, Auckland University of Technology
NZ
jianhua.zhao@autuni.ac.nz

Xue Jun Li
Department of Electrical and Electronic Engineering, Auckland University of Technology
NZ
xuejun.li@aut.ac.nz

Peter Han Joo Chong
Department of Electrical and Electronic Engineering, Auckland University of Technology
NZ
peter.chong@aut.ac.nz

Abstract

Implicit neural representations have recently demonstrated considerable potential in various applications, including video compression and reconstruction, owing to their rapid decoding speed and high adaptability. Based on the most advanced neural representation for Videos (NeRV), Expedite Neural Representation for Videos (E-NeRV) and Hybrid Neural Representation for Videos (H-NeRV) primarily boost performance by enhancing and broadening the NeRV network's embedded input, whereas the NeRV module—the central component involved in video reconstruction—has attracted less attention. With a focus on high-frequency data in the frequency domain, this paper proposes a High-frequency Spectrum Hybrid Network (HFS-HNeRV), which adopts effective high-frequency data from the frequency domain to generate image details. Its core, HFS-HNeRV block, is a novel NeRV module, which adds the high-frequency spectrum convolution module (HFSCM) to the original one. This module extracts and emphasizes high-frequency features through the frequency domain attention mechanism, which not only provides superior performance, but also enhances the local detail recovery in video images. As an upgrade of the NeRV module, it has exceptional performance in terms of adaptability and versatility. It can conveniently substitute in a variety of current NeRV designs without requiring significant alterations to attain enhanced performance. Furthermore, this paper also introduces the High-frequency Spectrum (HFS) loss function to further mitigate the blurriness issue caused by the loss of high-frequency information during image generation. In the video compression task, the proposed HFS-HNeRV network outperformed NeRV, E-NeRV and HNeRV with an improvement of +5.68 dB, +4.46 dB, and +0.98 dB in reconstruction quality (PSNR), respectively.

CCS Concepts

• **Information systems**; • **Computing methodologies**;

Keywords

Video Compression; Artificial Intelligence; Implicit Neural Representation; Video Reconstruction

MMASIA '24, December 03–06, 2024, Auckland, New Zealand

ACM ISBN 979-8-4007-1273-9/24/12
https://doi.org/10.1145/3696409.3700250

ACM Reference Format:
Jianhua Zhao, Xue Jun Li, and Peter Han Joo Chong. 2024. HFS-HNeRV: High-Frequency Spectrum Hybrid Neural Representation for Videos. In *ACM Multimedia Asia (MMASIA '24), December 03–06, 2024, Auckland, New Zealand.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3696409.3700250

1 Introduction

Recently worldwide global Internet traffic has been growing at an average rate of 22%, and it surpasses 33 exabytes per day. This growth is particularly due to the increasing popularity of high definition video across multiple applications like video conferencing, security monitoring, medical care, agriculture and forestry, and online video streaming (e.g., Youtube and Netflix). Uncompressed raw video files require massive storage space and network bandwidth, despite ongoing advancements in hardware storage and network transmission technologies. Consequently, a crucial field of research is video compression, which requires applying efficient methods to minimize the amount of video data while maintaining the visual quality of the video as much as possible after reconstruction.

Traditionally, video encoding compresses data mainly through the application of the discrete cosine transform (DCT) [25] and predictive coding in the spatial and temporal domains. Deep learning-based video compression algorithms provide notable benefits in terms of end-to-end optimisation, quality retention and compression ratio. Existing popular works include learning-based modules for conventional codec adaption [9, 11, 27, 32, 33, 36] and end-to-end video compression models [1, 8, 10, 19, 22]. Furthermore, neural representation for video (NeRV) models [3–5, 18], based on implicit neural representation, have gained widespread attention owing to their straightforward design, high-level adaptability and blazingly fast decoding speed. Currently, E-NeRV [18] and HNeRV [4] are notable examples of cutting-edge works. Compared to NeRV [5], these two methods can reconstruct video frames efficiently with higher quality.

Although E-NeRV [18] and HNeRV [4] achieved satisfactory results, the research on NeRV still face several restrictions and challenges.

First of all, although both E-NeRV [18] and HNeRV [4] slightly adjust the number of channels in the NeRV blocks, they mainly achieve superior performance by optimizing input embeddings of the NeRV network. In [5], Chen et al. only utilized frame indices, which are real numbers, as the temporal input embeddings. E-NeRV [18] further adds spatial coordinates as spatial embeddings. Moreover, HNeRV [4] enriches the spatial embeddings by extracting

feature maps from the ground-truth video images employing ConvNeXt [21] (a regular Convolutional Neural Network (CNN)) as an encoder. Enhancing the quality of input embeddings is indeed a highly effective method to enhance the performance of models. Nonetheless, enhancing the NeRV block's effectiveness remains an essential issue.

Second, the best performing model currently, HNeRV [4], generates images with abrupt color pixels and missing edges, which not only blurs the image but also destroys the overall color coherence. In particular, HNeRV [4] failed to generate the edge details of the nose and mouth when reconstructing the character's face. Also, it introduced many noise points that disrupt color consistency on the face. We found that the NeRV module relies entirely on each convolution kernel in the convolution layer to generate new features by referencing the values of the elements adjacent to its center point. This presents two potential issues:

- Small convolutional kernels can only reference a restricted range of features, resulting in incorrect pixel values generated by the network. While enlarging the convolutional kernel can effectively expand the reference information to improve the performance, performing so will quadratically raise the total amount of parameters.
- Convolution is a weighted summation operation, which means that rather than generating high-frequency information with distinct local variations, it typically preserves or generates low-frequency information that is smooth and continuous over a wide range. Consequently, the network struggles to retain or generate high-frequency information, whether the encoder is processing deep feature maps (input embeddings) or the decoder is generating images. This causes the network cannot rebuild object edges and texture details.

In response to the above challenges, we propose an innovative high-frequency spectrum hybrid neural representation for video (HFS-HNeRV). Fig. 1 (a) and Fig. 1 (b) show the main architecture of HFS-HNeRV and its main differences from existing NeRV-based works. To address the first challenge, we propose the HFS-HNeRV block. It enhances the existing basic NeRV module with a high-frequency spectrum convolution module (HFSCM), providing more effective video frame reconstruction capabilities while maintaining a stable parameter count. Moreover, it has excellent compatibility and generalizability, as well as simplicity to integrate into an extensive variety of NeRV networks without requiring significant modifications to the original network architecture.

For the second challenge, the proposed HFSCM contains a new high-frequency enhancement attention mechanism. It introduces Haar wavelet transform to strengthen the high-frequency components, which can effectively capture the high-frequency features in the feature map to restore the edge details and texture of the image, thereby enhancing the quality of image generation. Also, the application of the attention mechanism promotes the module's ability to extract and fuse global information, partially alleviating the issue of insufficient receptive field. Furthermore, HFSCM adopts a dual convolutional layer structure, which can further fuse the features enhanced by the high-frequency spectrum attention mechanism (HFSAM) to generate richer feature representations. Moreover, we propose a high-frequency spectrum loss function to assist in the training of the model. This loss function extracts high-frequency signals of the predicted images and the ground-true images through Fourier transform and high-pass filters, then determining the mean square error between them. We add the high-frequency spectrum (HFS) loss to the loss calculation formula based on mean square error (MSE) loss and controllably raise its weight on the overall error through a hyperparameter, which can appropriately mitigate the excessive impact of low-frequency components on the overall error to force the model to focus on the generation of the image details such as edges and textures.

In summary, our work makes the following contributions:

- We propose a novel NeRV module, HFS-HNeRV block, which can be easily integrated into various NeRV networks without substantial modifications on the network architecture.
- We introduce a new loss function specifically designed for high-frequency information generation, enhancing the model's capacity to reconstruct image details.

2 Related Work

2.1 Implicit neural representations

Implicit neural representations [23], which are frequently employed in image [6, 30] or scene reconstruction [13, 24], are methods that employ neural networks to represent geometric forms or sceneries. For instance, NeRF [24] can reconstruct a 3D scene from supplied 3D coordinates. Additionally, NeRV [5] rebuilds the video rapidly by generating the current frame from frame indices. Compared to explicit representations, implicit representations primarily store information in the parameters of the network, resulting in lower storage requirements. However, implicit representations also have several limitations. It is challenging to employ neural networks for implicit representations in real-time applications since they need a great deal of resources during the training phase, such as training time, dataset and computing resources. Furthermore, the model's high complexity may lead to instability during the training phase.

2.2 Video super-resolution

Video super-resolution is a technique that is frequently employed in applications like remote sensing and telemedicine to enhance the resolution and visual clarity of video frames. Its core mainly involves upsampling methods such as interpolation, pixel shuffle and deconvolution. The deep learning approaches utilized in video super-resolution can be classified into single-frame type [7, 14, 17, 28, 29] and multi-frame type [2, 16, 26, 31, 34, 35]. Since video is a sequence of consecutive images forming a dynamic visual record, single-frame super-resolution networks are essentially extensions of image super-resolution methods. Super-resolution Convolutional Neural Network (SRCNN) [7], Very Deep Super-Resolution (VDSR) [14], and Super-Resolution Generative Adversarial Network (SRGAN) [17] are commonly applied methods. In contrast to single-frame networks, multi-frame super-resolution networks make full use of inter-frame information in videos, which can be classified into methods that align video frames and those that do not. Techniques based on optical flow estimation [16, 34] and deformable convolution [31, 35] are representative works in the former category; techniques based on 3D convolution [15, 20] and

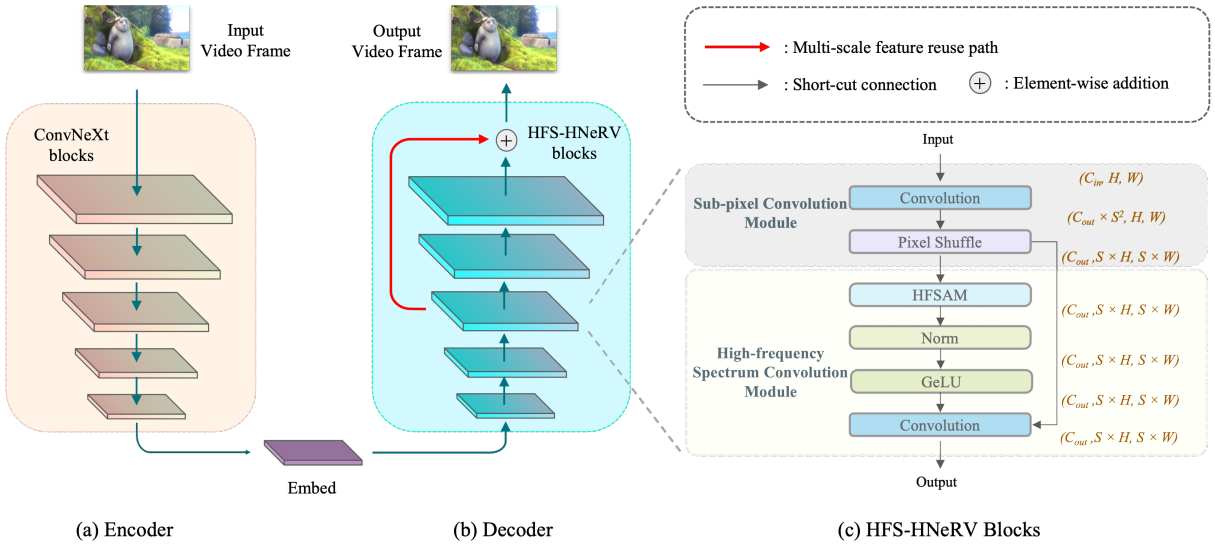

Figure 1: The structure of HFS-HNeRV. (a) In the encoding process, we employ ConvNeXt to downsample the input video frames into smaller embeddings. (b) In the decoding process, in addition to utilizing the unique HFS-HNeRV blocks, we also reuse the output feature maps from the third layer. (c) In HFS-HNeRV, we introduce a residual structure with dual convolutional layers, which further enriches the generated features by leveraging the attention maps produced by HFSAM.

recurrent convolutional neural networks [12, 37] are representative works in the latter. The primary goal of video super-resolution is to upsample low-resolution videos to high-resolution videos. This process is quite similar to how NeRV [5] progressively upsample an embedding into a complete video image. Therefore, there is some overlap between the methods employed in these two domains. Moreover, videos with low quality can be considered as compressed versions of those with high resolution.

3 Proposed Method

3.1 HFS-HNeRV Blocks

As can be seen in Fig. 1 (c), HFS-HNeRV Block is composed of sub-pixel convolution module and HFSCM.

Sub-pixel Convolution Module. This module is a combination of the convolution layer and the pixel shuffle layer. In the convolution layer as represented in Fig. 1 (c), the input feature map satisfies $X \in R^{H\times W\times C}$, and the output feature map is dimensioned to $Y \in R^{H\times W\times S^2C}$. The dimensionality enhancement process can be viewed as the network layer attempting to extract features from the current feature map, which will be used as references to generate more relevant features. Lowering the amount of input or output channels will cause this network layer's performance to drastically suffer. Furthermore, expanding the convolutional kernel size can enhance the network's efficiency, but it also leads to a substantial growth in model parameters. Consequently, in order to maintain the stability of the parameter count, we have kept the original settings of kernel size and channel.

High-frequency Spectrum Convolution Module. HFSCM is mainly divided into a high-frequency spectrum attention mechanism and an additional convolutional layer. As shown in Fig. 1 (c), the overall structure of the module is a residual block.

HFSAM consists of two parts: the channel attention layer and the frequency-domain spatial attention layer. The channel attention layer utilizes a double multilayer perceptron (MLP) layer to generate a channel attention map by extracting the global information from the feature vectors of all $H \times W$ positions in the feature map F_1, as illustrated in Fig. 2. This can be represented by the following formula:

$$C_{Atten} = \sigma(MLP(GeLU(MLP(F_1)))) \tag{1}$$

$$F_2 = (F_1 \otimes C_{Atten}) + F_1 \tag{2}$$

Where σ denotes the sigmoid function. $GeLU$ represents the GeLU activation function.

In the frequency-domain spatial attention layer as shown in Fig. 3, the first convolution layer would reduce the number of channels of the input feature map F_2. This is mainly to avoid the large number of parameters of the attention layer as much as possible. Subsequently, the feature map is decomposed into different frequency component sub-maps by Haar wavelet transform, namely low frequency-low frequency (LL) map, low frequency-high frequency (LH) map, high frequency-low frequency (HL) map and high frequency-high frequency (HH) map. Following an upsample to match the original feature map's size, we multiply these four sub-maps by various enhancement weights and perform element-wise addition of each sub-feature map with the input feature map, respectively. Then, these four enhanced sub-maps are concatenated

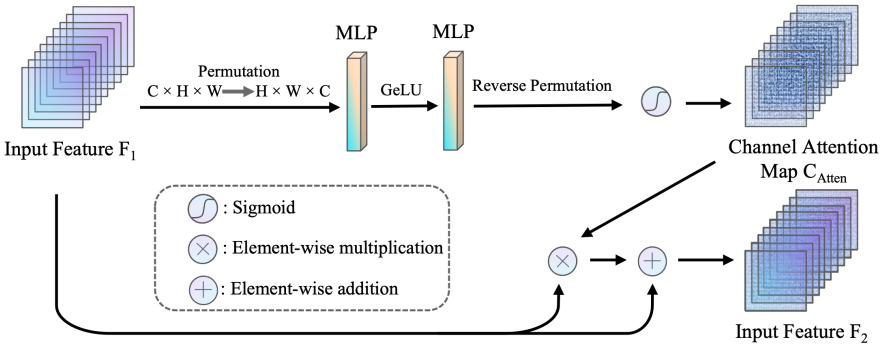

Figure 2: Channel Attention. This part employs dual MLP layers to integrates the intra-channel contextual information of input feature, computing the output feature map through a residual structure.

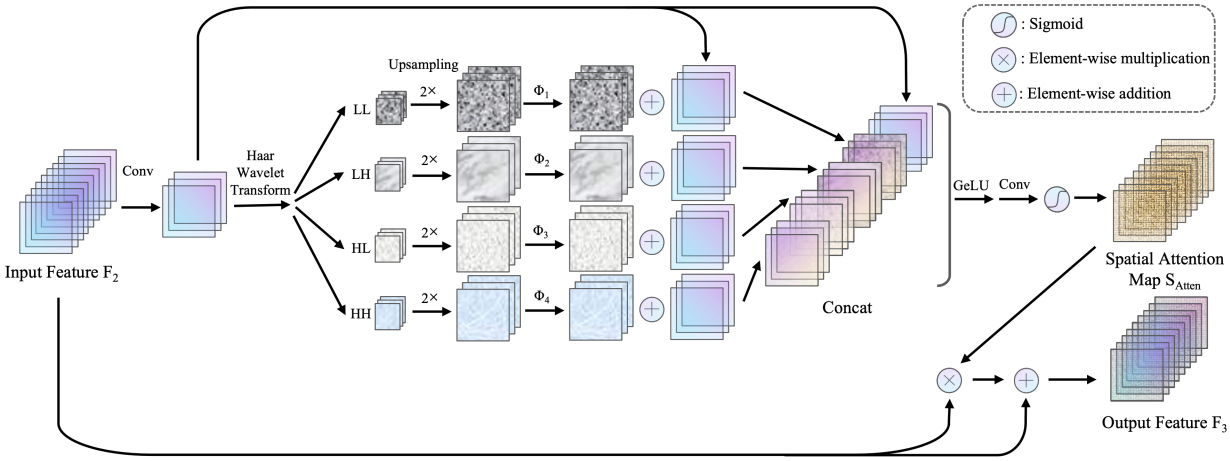

Figure 3: Spatial Attention. We enhance high-frequency information by incorporating Haar wavelet transform into the spatial attention mechanism. Note that 2x symbolizes the two-fold upsampling operation.

with the input feature map. By introducing the Haar wavelet transform to analyze the frequency domain information, HFSCM can more effectively capture the high-frequency features in the feature map to restore the edge details and texture of the image, thereby enhancing the quality of image generation. Moreover, the application of the attention mechanism promotes the module's ability to extract and fuse global information, partially alleviating the issue of insufficient receptive field. The process of spatial attention calculation can be illustrated as follows:

$$F_{LL}, F_{HL}, F_{LH}, F_{HH} = Up(DWT_{Haar}(Conv(F2))) \tag{3}$$

$$F_{LL2} = \Phi_1 F_{LL} + F2 \tag{4}$$

$$F_{HL2} = \Phi_2 F_{HL} + F2 \tag{5}$$

$$F_{LH2} = \Phi_3 F_{LH} + F2 \tag{6}$$

$$F_{HH2} = \Phi_4 F_{HH} + F2 \tag{7}$$

$$S_{Atten} = \sigma(Conv(GeLU(Concat(F_2, F_{LL2}, F_{HL2}, F_{LH2}, F_{HH2})))) \tag{8}$$

$$F_3 = (F_2 \otimes S_{Atten}) + F_2 \tag{9}$$

Where Up denotes the two-fold upsampling operation, and DWT_{Haar} represents the haar wavelet transform. Φ represents the enhancement factor of the frequency maps.

3.2 High-frequency Spectrum Loss

The MSE loss function is widely employed in a series of downstream tasks in computer vision. To further enhance the model's attention to high-frequency features in the image, we propose HFS loss based on Fourier transform and high-pass filters, which is added to the total loss.

Specifically, we first convert the generated image and the ground-true image into frequency domain representation through Fourier

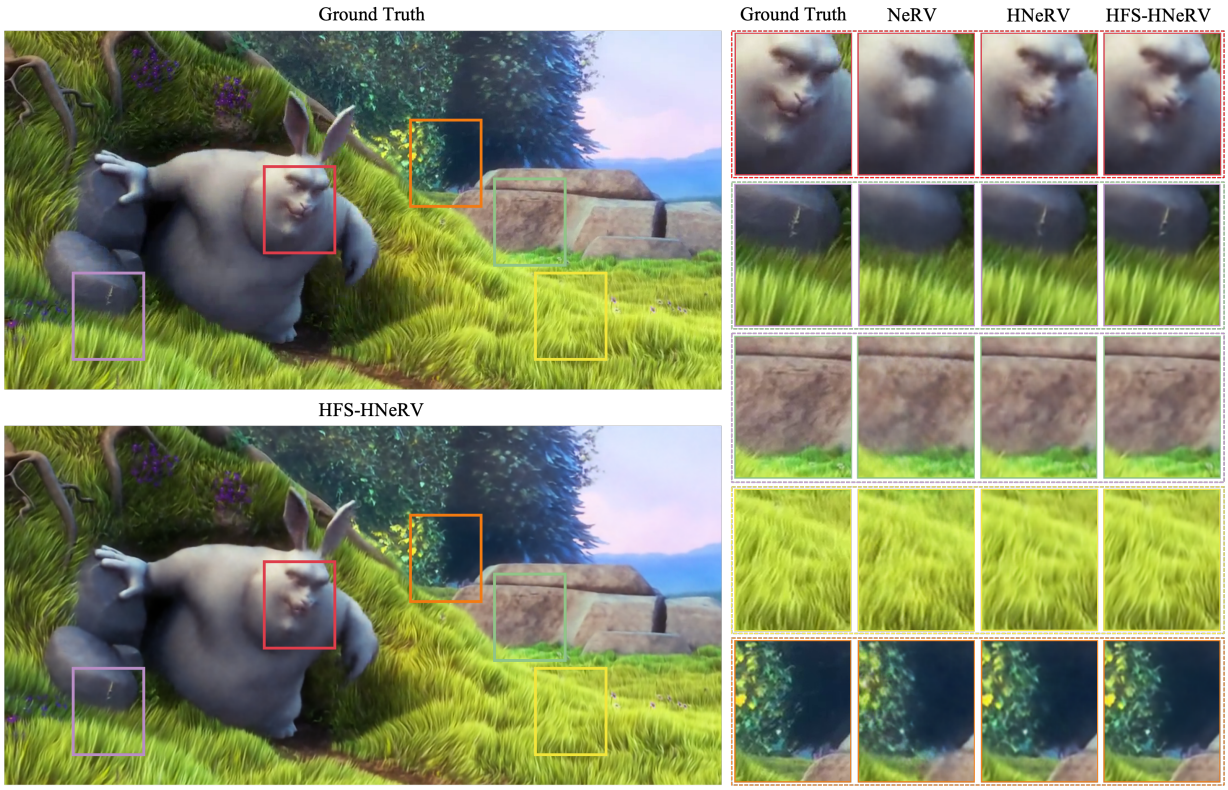

Figure 4: Visual quality comparison of videos. On the left, we compare one overall video frame generated by HFS-HNeRV with the ground truth. On the right, we compare NeRV, HNeRV and HFS-HNeRV by extracting and analyzing five patches from the images. It can be observed that HFS-HNeRV consistently outperforms in various aspects, including facial details, small objects (such as the contour of a small blade of grass), local region details (such as the texture of rocks and grass), and low-contrast objects (such as a leaf in darkness).

transform. Then, the low-frequency components below the threshold are removed by the created high-pass filter mask, and the remaining high-frequency information is enhanced by the enhancement factor. Finally, we employ the inverse Fourier transform to convert the frequency domain data of the two images back to the spatial domain representation and calculate their mean square error.

The formulae for MSE loss and HFS Loss are expressed as:

$$\mathcal{L}_{\text{MSE}} = \frac{1}{M \times N \times C} \sum_{i=1}^{M} \sum_{j=1}^{N} \sum_{c=1}^{C} \left(PRED_{i,j,c} - GT_{i,j,c}\right)^2 \tag{10}$$

$$\mathcal{L}_{\text{HFS}} = \mathcal{L}_{\text{MSE}}(iFFT(HPF(FFT(PRED, GT)))) \tag{11}$$

Where M and N represent the height and width of the image, respectively. C represents the number of channels of the image. $PRED$ and GT represent the predicted image and ground-truth image, respectively.

The total loss function can be expressed as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{MSE}}(PRED_i, GT_i) + \lambda \mathcal{L}_{\text{HFS}}(PRED_i, GT_i) \tag{12}$$

Where λ is the weight that controls the influence of HFS loss.

Table 1: PSNR(dB) results on Bunny with different model size.

Size	0.75M	1.5M	3.0M
NeRV	28.46	30.87	33.21
E-NeRV	30.95	32.09	36.72
HNeRV	32.81	35.57	37.43
HFS-HNeRV	34.17	36.55	38.82

4 Experiments

4.1 Dataset and Implementation

In this paper, we utilized the hybrid representation network architecture of HNeRV. Therefore, most of the experimental settings are consistent with HNeRV. For the dataset, we continued to deploy the Big Buck Bunny, cropped to the center with a resolution of 640 × 1280. For performance metrics, we retained the same settings as those in HNeRV [4], specifically peak-signal-to-noise ratio (PSNR) and multi-scale structural similarity index measure (MS-SSIM). Moreover, we selected multiple regions within the images to conduct a human visual quality comparison. All experiments were conducted on one RTX 3060 GPU.

Table 2: PSNR(dB) results on Bunny with different trainning epochs.

Epoch	300	600	1200
NeRV	30.87	31.68	32.13
E-NeRV	32.09	33.2	34.15
HNeRV	35.57	36.19	36.93
HFS-HNeRV	36.55	37.37	37.89

4.2 Results

Compared with existing work, we have enhanced both the performance metrics of the model and the actual human visual quality. All experiments were conducted on the Big Buck Bunny dataset. As shown in Table 1, under the condition of a model size of 1.5M, HFS-HNeRV with different training epochs outperforms NeRV, E-NeRV, and HNeRV. Similarly, setting the training cycle to 300 as the benchmark, HFS-HNeRV still achieves superior performance under the same model size (Table 2). In terms of actual visual quality, we set the experimental benchmark to 300 training epochs and a model size of 1.5M. As can be seen in Fig. 4, it is evident that the textures near the edges of small objects in the image are more distinct and complete. Moreover, the images generated by HFS-HNeRV exhibit significantly fewer pixels with abrupt color changes, resulting in an overall more harmonious and natural visual appearance.

5 Conclusion

In this paper, we propose a NeRV network based on high-frequency feature learning in the frequency domain, namely HFS-HNeRV. A function for calculating the loss of high-frequency features is proposed to assist its training, so as to further enhance the model's capacity to generate details such as edges and textures in the image. Specifically, we design a novel high-frequency spectrum convolution module (HFSCM) and a high-frequency spectrum loss function (HFS loss) to allow the model to learn and concentrate on high-frequency information in the frequency domain. Quantitative data results show that compared with other NeRV-based networks, such as NeRV, E-NeRV and HNeRV, HFS-HNeRV has achieved +5.95 dB, +4.73 dB and +0.98 dB improvements in PSNR, respectively. Regarding the actual quality of visual representation, HFS-HNeRV is more effective in reconstructing the edge textures of image objects, and the images that emerge have a more natural and seamless color distribution. In particular, both HFSCM and HFS loss are highly flexible and they can be easily integrated into various types of NeRV networks, which can play a positive role in video compression and reconstruction.

Acknowledgments

This work was funded by the Vice-Chancellor's Doctoral Scholarship at Auckland University of Technology, New Zealand.

References

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

[2] Kelvin CK Chan, Xintao Wang, Ke Yu, Chao Dong, and Chen Change Loy. 2021. Basicvsr: The search for essential components in video super-resolution and beyond. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4947–4956.

[3] Hao Chen, Matt Gwilliam, Bo He, Ser-Nam Lim, and Abhinav Shrivastava. 2022. CNeRV: Content-adaptive Neural Representation for Visual Data. arXiv:2211.10421 [cs.CV] https://arxiv.org/abs/2211.10421

[4] Hao Chen, Matthew Gwilliam, Ser-Nam Lim, and Abhinav Shrivastava. 2023. Hnerv: A hybrid neural representation for videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10270–10279.

[5] Hao Chen, Bo He, Hanyu Wang, Yixuan Ren, Ser-Nam Lim, and Abhinav Shrivastava. 2021. NeRV: Neural Representations for Videos. arXiv:2110.13903 [cs.CV] https://arxiv.org/abs/2110.13903

[6] Yinbo Chen, Sifei Liu, and Xiaolong Wang. 2021. Learning continuous image representation with local implicit image function. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8628–8638.

[7] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. 2014. Learning a deep convolutional network for image super-resolution. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part IV 13*. Springer, 184–199.

[8] Liangwei Fu, Ping Wang, and Xinhong Wang. 2024. An Improved Neural Network Approach to End-to-end Video Compression. In *Proceedings of the 5th International Conference on Computer Information and Big Data Applications*. 57–61.

[9] Man M Ho, Jinjia Zhou, Gang He, Muchen Li, and Lei Li. 2020. SR-CL-DMC: P-frame coding with super-resolution, color learning, and deep motion compensation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*. 124–125.

[10] Zhihao Hu, Dong Xu, Guo Lu, Wei Jiang, Wei Wang, and Shan Liu. 2022. Fvc: An end-to-end framework towards deep video compression in feature space. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 45, 4 (2022), 4569–4585.

[11] Zhijie Huang, Xiaopeng Guo, Mingyu Shang, Jie Gao, and Jun Sun. 2021. An efficient qp variable convolutional neural network based in-loop filter for intra coding. In *2021 Data Compression Conference (DCC)*. IEEE, 33–42.

[12] Takashi Isobe, Xu Jia, Shuhang Gu, Songjiang Li, Shengjin Wang, and Qi Tian. 2020. Video super-resolution with recurrent structure-detail network. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII 16*. Springer, 645–660.

[13] Chiyu Jiang, Avneesh Sud, Ameesh Makadia, Jingwei Huang, Matthias Nießner, Thomas Funkhouser, et al. 2020. Local implicit grid representations for 3d scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6001–6010.

[14] Jiwon Kim, Jung Kwon Lee, and Kyoung Mu Lee. 2016. Accurate image super-resolution using very deep convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1646–1654.

[15] Soo Ye Kim, Jeongyeon Lim, Taeyoung Na, and Munchurl Kim. 2019. Video super-resolution based on 3D-CNNS with consideration of scene change. In *2019 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2831–2835.

[16] Tae Hyun Kim, Mehdi SM Sajjadi, Michael Hirsch, and Bernhard Scholkopf. 2018. Spatio-temporal transformer network for video restoration. In *Proceedings of the European conference on computer vision (ECCV)*. 106–122.

[17] Christian Ledig, Lucas Theis, Ferenc Huszar, Jose Caballero, Andrew Cunningham, Alejandro Acosta, et al. 2017. Photo-realistic single image super-resolution using a generative adversarial network [C]. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, Vol. 2017. 4681–4690.

[18] Zizhang Li, Mengmeng Wang, Huaijin Pi, Kechun Xu, Jianbiao Mei, and Yong Liu. 2022. E-NeRV: Expedite Neural Video Representation with Disentangled Spatial-Temporal Context. arXiv:2207.08132 [cs.CV] https://arxiv.org/abs/2207.08132

[19] Bowen Liu, Yu Chen, Rakesh Chowdary Machineni, Shiyu Liu, and Hun-Seok Kim. 2023. Mmvc: Learned multi-mode video compression with block-based prediction mode selection and density-adaptive entropy coding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 18487–18496.

[20] Hongying Liu, Peng Zhao, Zhubo Ruan, Fanhua Shang, and Yuanyuan Liu. 2021. Large motion video super-resolution with dual subnet and multi-stage communicated upsampling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 2127–2135.

[21] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. 2022. A ConvNet for the 2020s. arXiv:2201.03545 [cs.CV] https://arxiv.org/abs/2201.03545

[22] Guo Lu, Xiaoyun Zhang, Wanli Ouyang, Li Chen, Zhiyong Gao, and Dong Xu. 2020. An end-to-end learning framework for video compression. *IEEE transactions on pattern analysis and machine intelligence* 43, 10 (2020), 3292–3308.

[23] Ishit Mehta, Michaël Gharbi, Connelly Barnes, Eli Shechtman, Ravi Ramamoorthi, and Manmohan Chandraker. 2021. Modulated periodic activations for generalizable local functional representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 14214–14223.

[24] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. 2021. Nerf: Representing scenes as neural radiance fields for view synthesis. *Commun. ACM* 65, 1 (2021), 99–106.

[25] William K Pratt, Julius Kane, and Harry C Andrews. 1969. Hadamard transform image coding. *Proc. IEEE* 57, 1 (1969), 58–68.

[26] Mehdi SM Sajjadi, Raviteja Vemulapalli, and Matthew Brown. 2018. Frame-recurrent video super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 6626–6634.

[27] Jens Schneider, Johannes Sauer, and Mathias Wien. 2017. Dictionary learning based high frequency inter-layer prediction for scalable HEVC. In *2017 IEEE Visual Communications and Image Processing (VCIP)*. IEEE, 1–4.

[28] Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. 2016. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1874–1883.

[29] Assaf Shocher, Nadav Cohen, and Michal Irani. 2018. "zero-shot" super-resolution using deep internal learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3118–3126.

[30] Yannick Strümpler, Janis Postels, Ren Yang, Luc Van Gool, and Federico Tombari. 2022. Implicit neural representations for image compression. In *European Conference on Computer Vision*. Springer, 74–91.

[31] Yapeng Tian, Yulun Zhang, Yun Fu, and Chenliang Xu. 2020. Tdan: Temporally-deformable alignment network for video super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3360–3369.

[32] Yang Wang, Xiaopeng Fan, Shaohui Liu, Debin Zhao, and Wen Gao. 2019. Multi-scale convolutional neural network-based intra prediction for video coding. *IEEE Transactions on Circuits and Systems for Video Technology* 30, 7 (2019), 1803–1815.

[33] Zhao Wang, Changyue Ma, Ru-Ling Liao, and Yan Ye. 2021. Multi-density convolutional neural network for in-loop filter in video coding. In *2021 Data Compression Conference (DCC)*. IEEE, 23–32.

[34] Tianfan Xue, Baian Chen, Jiajun Wu, Donglai Wei, and William T Freeman. 2019. Video enhancement with task-oriented flow. *International Journal of Computer Vision* 127 (2019), 1106–1125.

[35] Xinyi Ying, Longguang Wang, Yingqian Wang, Weidong Sheng, Wei An, and Yulan Guo. 2020. Deformable 3d convolution for video super-resolution. *IEEE Signal Processing Letters* 27 (2020), 1500–1504.

[36] Zheng-Teng Zhang, Chia-Hung Yeh, Li-Wei Kang, and Min-Hui Lin. 2017. Efficient CTU-based intra frame coding for HEVC based on deep learning. In *2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*. IEEE, 661–664.

[37] Xiaobin Zhu, Zhuangzi Li, Xiao-Yu Zhang, Changsheng Li, Yaqi Liu, and Ziyu Xue. 2019. Residual invertible spatio-temporal network for video super-resolution. In *Proceedings of the AAAI conference on artificial intelligence*, Vol. 33. 5981–5988.