Efficient Neural Video Representation via Structure-Preseving Patch Decoding

Taiga Hayami
Graduate School of FSE,
Waseda University
Tokyo, Japan
hayatai17@fuji.waseda.jp

Kakeru Koizumi
Graduate School of FSE,
Waseda University
Tokyo, Japan
kkeverio@ruri.waseda.jp

Hiroshi Watanabe
Graduate School of FSE,
Waseda University
Tokyo, Japan
hiroshi.watanabe@waseda.jp

***Abstract*—Implicit Neural Representations (INRs) have attracted significant interest for their ability to model complex signals by mapping spatial and temporal coordinates to signal values. In the context of neural video representation, several decoding strategies have been explored to balance compactness and reconstruction quality, including pixel-wise, frame-wise, and patch-wise methods. Patch-wise decoding aims to combine the flexibility of pixel-based models with the efficiency of frame-based approaches. However, conventional uniform patch division often leads to discontinuities at patch boundaries, as independently reconstructed regions may fail to form a coherent global structure. To address this limitation, we propose a neural video representation method based on Structure-Preserving Patches (SPPs). Our approach rearranges each frame into a set of spatially structured patch frames using a PixelUnshuffle-like operation. This rearrangement maintains the spatial coherence of the original frame while enabling patch-level decoding. The network learns to predict these rearranged patch frames, which supports a global-to-local fitting strategy and mitigates degradation caused by upsampling. Experiments on standard video datasets show that the proposed method improves reconstruction quality and compression performance compared to existing INR-based video representation methods.**

***Index Terms*—Implicit neural representation, video compression, video representation.**

I. Introduction

Implicit Neural Representations (INRs) have emerged as a compact, resolution-independent alternative to conventional explicit representations for modeling complex signals. By expressing data as continuous functions of spatial and temporal coordinates through neural networks, INRs significantly reduce memory usage. They also support desirable properties such as infinite-resolution rendering and smooth interpolation. These advantages make them well-suited for modeling fine-grained signals in domains such as novel view synthesis [1]–[4], image representation [5]–[8], and increasingly, video representation [9]–[11].

INR-based methods for neural video representation can be categorized by their decoding granularity: coordinate-based (pixel-wise) [12], [13], frame-based (image-wise) [9]–[11], and patch-based approaches [14], [15]. Each approach presents a different trade-off between flexibility, computational cost, and structural consistency. Coordinate-based approaches predict individual pixel values based on spatial and temporal inputs, offering unmatched resolution flexibility and the ability to capture intricate local variations. However, they are often computationally intensive and limited in capturing global context due to their inherently localized nature. Frame-based approaches, such as Neural Representations for Videos (NeRV) [9], mitigate this issue by predicting entire frames from compact embeddings, such as time indices or learned content vectors. These models benefit from faster inference and improved global structure modeling, but tend to struggle with high-frequency detail due to upsampling artifacts and spectral bias (favoring low-frequency components) [16]–[19]. Patch-based methods lie between these extremes. By reconstructing each frame from smaller subregions, they aim to combine the detail-preserving nature of pixel-wise decoding with the structural coherence of frame-wise decoding. Nonetheless, conventional patch partitioning via uniform spatial division can disrupt continuity across patch boundaries. Independently reconstructed patches often fail to align perfectly, resulting in visible seams and reduced visual quality. This limitation motivates the exploration of patch-based methods that maintain global spatial coherence without sacrificing local detail fidelity.

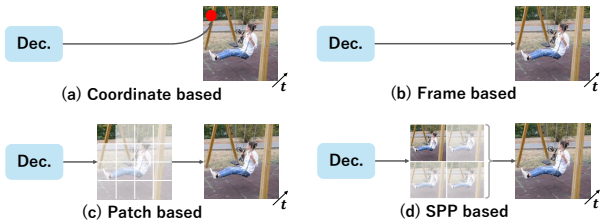

Fig. 1. Comparison of decoding strategies for neural video representation. (a) Coordinate-based: predicts pixels from spatio-temporal coordinates. (b) Frame-based: generates entire frames from embeddings. (c) Patch-based: reconstructs frames from uniformly divided patches. (d) SPP-based (ours): rearranges pixels into spatially aligned patches for structure-preserving, global-to-local reconstruction.

To overcome these limitations, we propose the Structure-Preserving Patch (SPP) method for neural video representation. Instead of independently decoding spatially divided patches, our method rearranges each frame into multiple sub-images through a pixel redistribution operation inspired by PixelUnshuffle. This operation preserves the relative spatial structure of the original frame across all patches. The model is trained to predict these structurally aligned patch images,

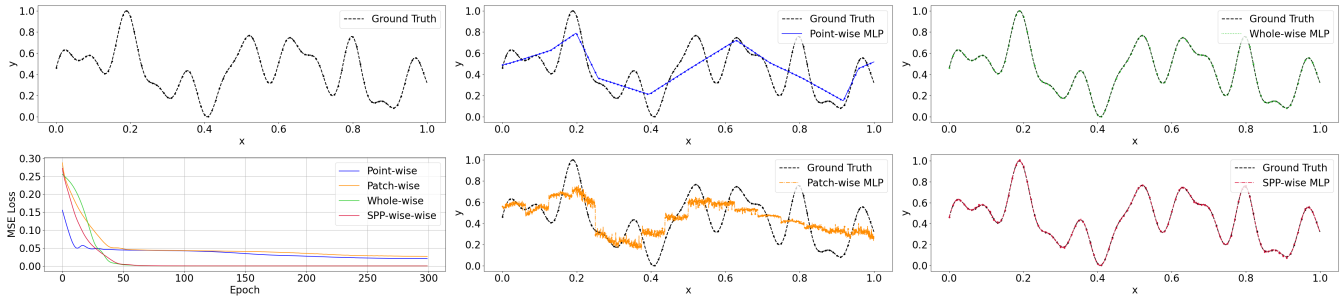

Fig. 2. Toy experiment comparing the fitting performance of one-dimensional signals. Top row (left to right): ground-truth signal, point-wise fitting, and whole-wise fitting. Bottom row (left to right): training loss curves, patch-wise fitting, and the proposed Structure-Preserving Patch (SPP)-wise fitting. While conventional patch-wise approaches suffer from boundary discontinuities, the proposed SPP-wise approach maintains structural continuity and enables accurate local fitting.

enabling it to first capture global structure and then refine local content. This global-to-local fitting process reduces the risk of discontinuities at patch boundaries and leads to improved visual quality. We validate the effectiveness of our method through experiments on several benchmark video datasets. The results show that our structure-aware patch decoding improves reconstruction quality and compression performance compared to prior INR-based video representation models.

II. Related Works

A. Implicit Neural Representation

The central concept of INR is to model a signal as a continuous function implemented by a neural network. Formally, a signal is defined as:

$$y = \Phi_\theta(x), \tag{1}$$

where x is a continuous coordinate, y is the signal value at that location, and θ denotes the learnable parameters of the network. The function Φ_θ is trained to fit a given target signal through optimization. A seminal work in this field is Neural Radiance Fields (NeRF) [1], which models a 3D scene as a radiance field defined over spatial coordinates and viewing directions. NeRF demonstrated that photo-realistic view synthesis is possible by optimizing a fully connected network to predict color and density values along camera rays. Although computationally demanding, NeRF highlights the potential of neural functions to model complex continuous structures without relying on explicit mesh or voxel representations. To enhance the capability of INRs in modeling high-frequency details, SIREN [12] introduced a sinusoidal-activated network capable of modeling high-frequency content with improved gradient flow. By leveraging periodic activation functions, SIREN reduces the spectral bias inherent in ReLU-based networks and improves the network's ability to capture fine details and textures. These foundational works have inspired the application of INRs to spatio-temporal domains, including video data. Their success shows that compact and continuous neural representations can be extended from static to dynamic signals, motivating recent research into INR-based video modeling.

B. Neural Video Representation

INRs have recently been explored as an alternative paradigm for video representation and compression. A representative example is NeRV [9], which encodes an entire video in the weights of a neural network that maps a frame index to its corresponding RGB frame. Unlike conventional INR with coordinate mapping, NeRV treats the video as a global function of learned time, effectively embedding the video content into the network parameters. Video decoding then requires only a forward pass for a given frame index. This formulation means that compressing the network (through techniques like weight pruning or quantization) effectively compresses the video. Building on this idea, E-NeRV [10] improves NeRV by disentangling spatial and temporal components. It learns separate embeddings for each domain and combines them during decoding via Adaptive Instance Normalization (AdaIN) [20], thereby improving both representational efficiency and compression performance. HNeRV [11] further builds on this concept by replacing content-agnostic embeddings derived from the frame index with content-adaptive embeddings. This adjustment allocates greater modeling capacity to complex or high-frequency frames. More recently, BoostingNeRV [21] introduced a conditional decoding framework that can enhance various existing INR-based video models. It modulates intermediate features using a temporal affine transformation based on frame indices, enabling improved reconstruction performance across different architectures.

In addition, patch-wise and hybrid methods have been proposed to improve reconstruction quality and scalability. For instance, PS-NeRV [14] reconstructs videos on a patch-by-patch basis instead of generating full frames in one shot. This patch-wise process offers flexibility in handling higher resolutions and can distribute computational load, albeit with the risk of patch boundary artifacts. To better capture video dynamics, DNeRV [22] adopts a dual-stream architecture. One stream models the static background, while the other captures inter-frame differences, allowing motion to be represented implicitly. Several other approaches introduce hierarchical model structures [23]–[25], incorporating explicit residual connections [26], [27], or emphasizing the preservation of

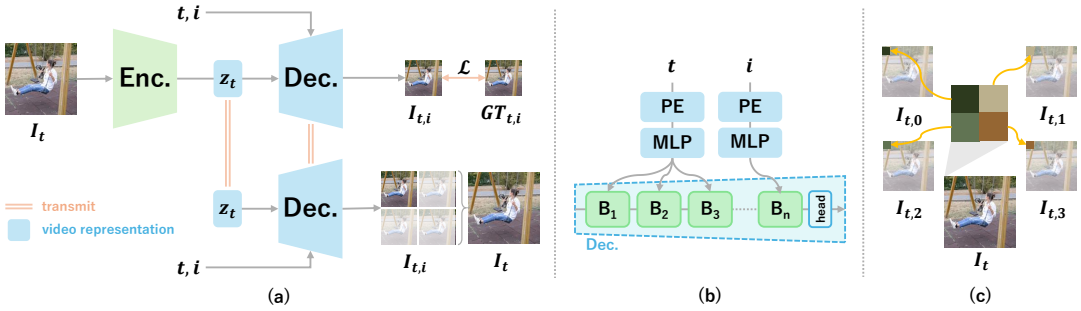

Fig. 3. Overview of the proposed Structure-Preserving Patchs (SPPs) based video representation framework. (a) Overall architecture. An input frame I_t is encoded into a frame-level embedding z_t, which is then decoded into patch images $I_{t,i}$ using the temporal index t and patch index i. The full frame is reconstructed by spatially rearranging the decoded patches. (b) Decoder details. Temporal and patch indices are embedded using positional encodings and MLPs. Early decoder layers are conditioned on t to capture global structure, while later layers use i to refine local patch details. (c) Patch construction via SPPs. Each frame is rearranged into multiple patch images (e.g., $I_{t,0}$, $I_{t,1}$, $I_{t,2}$, $I_{t,3}$) that preserve spatial consistency. This design supports a smooth global-to-local reconstruction process.

frequency-domain characteristics [28]–[30]. Each of these innovations addresses specific shortcomings of basic INR video models, such as temporal flicker and poor high-frequency reconstruction. Our work aligns with this line of research, focusing on patch-based representation. We specifically aim to preserve global image structure in a patch-wise model to avoid the common pitfalls of patch division.

III. Proposed Method

A. Overview

We propose a neural video representation method based on Structure-Preserving Patches (SPPs), as illustrated in Fig. 3. In this framework, each frame is decoded at the patch level. Unlike conventional patch-based methods, however, the patches are constructed to retain the original spatial layout of the frame. The key idea is to avoid treating patches as independent tiles and to generate patch images in a way that preserves the layout of the frame. These patch images tend to share similar structures and contain redundant information, which the network can exploit to improve representation efficiency. The decoding process is designed to follow a global-to-local fitting strategy. The model first learns a coarse global representation of the frame and then progressively refines local details within each patch. This hierarchical fitting process improves the network's ability to reconstruct both global context and fine details. By preserving spatial continuity across patch boundaries, our method mitigates the quality degradation commonly observed in naive patch-based upsampling or reconstruction.

B. Motivation

To highlight the importance of the global-to-local strategy, we conduct a simple one-dimensional signal-fitting experiment (Fig. 2). We compare four strategies: (1) *Point-wise*, where the signal is fitted at each coordinate independently; (2) *Patch-wise*, where the signal is divided into fixed-length segments and each segment is fitted separately; (3) *Whole-wise*, where the entire signal is fitted as a single input; and (4) *SPP-wise*, our proposed approach, which fits a rearranged version of the signal based on fixed-position samples, mimicking the structure-preserving patch mechanism. The results show that the whole-wise and SPP-wise strategies achieve superior fitting accuracy by preserving the overall structure of the signal. In contrast, point-wise and patch-wise methods prioritize local fitting and fail to maintain global consistency. These observations suggest that modeling the global structure first, followed by local refinement, is essential for accurate signal reconstruction. This insight forms the foundation of our global-to-local decoding framework used for video representation.

C. Structure-Preseving Patch Decoding

Our framework predicts each video frame at the patch level using a neural network that follows a global-to-local decoding strategy, as illustrated in Fig.3. For each input frame I_t at time t, we divide it conceptually into N spatially structured patches. To implement this, we apply a pixel rearrangement operation, similar to PixelUnshuffle, which transforms the frame into N patch images: $I_{t,0}, I_{t,1}, \ldots, I_{t,N-1}$. This rearrangement preserves the original spatial relationships across patches. The encoder, implemented using ConvNeXt blocks [11], [31], first processes the full frame I_t to produce a compact frame-level embedding z_t. This embedding is passed to a decoder, which generates each patch image $I_{t,i}$ under the influence of two indices: the temporal index t and the patch index i. In the decoder, early layers are modulated only by t, allowing the model to capture the global structure shared across all patches in the frame. Later layers are additionally conditioned on i, enabling the network to refine local details specific to each patch. Once all patch outputs $I_{t,i}$ are obtained, we reconstruct the full-resolution frame $\hat{I}_t$ by spatially rearranging them. This hierarchical conditioning mechanism allows the decoder to first build a coherent global structure, then specialize to local content, resulting in improved reconstruction quality. In addition, the structure-preserving nature of our patch generation helps avoid boundary discontinuities typically associated with naive patch-based decoding. The decoder is composed of multiple blocks B_n, each of which integrates the SNeRV block and the TAT Residual Block introduced in BoostingNeRV [21].

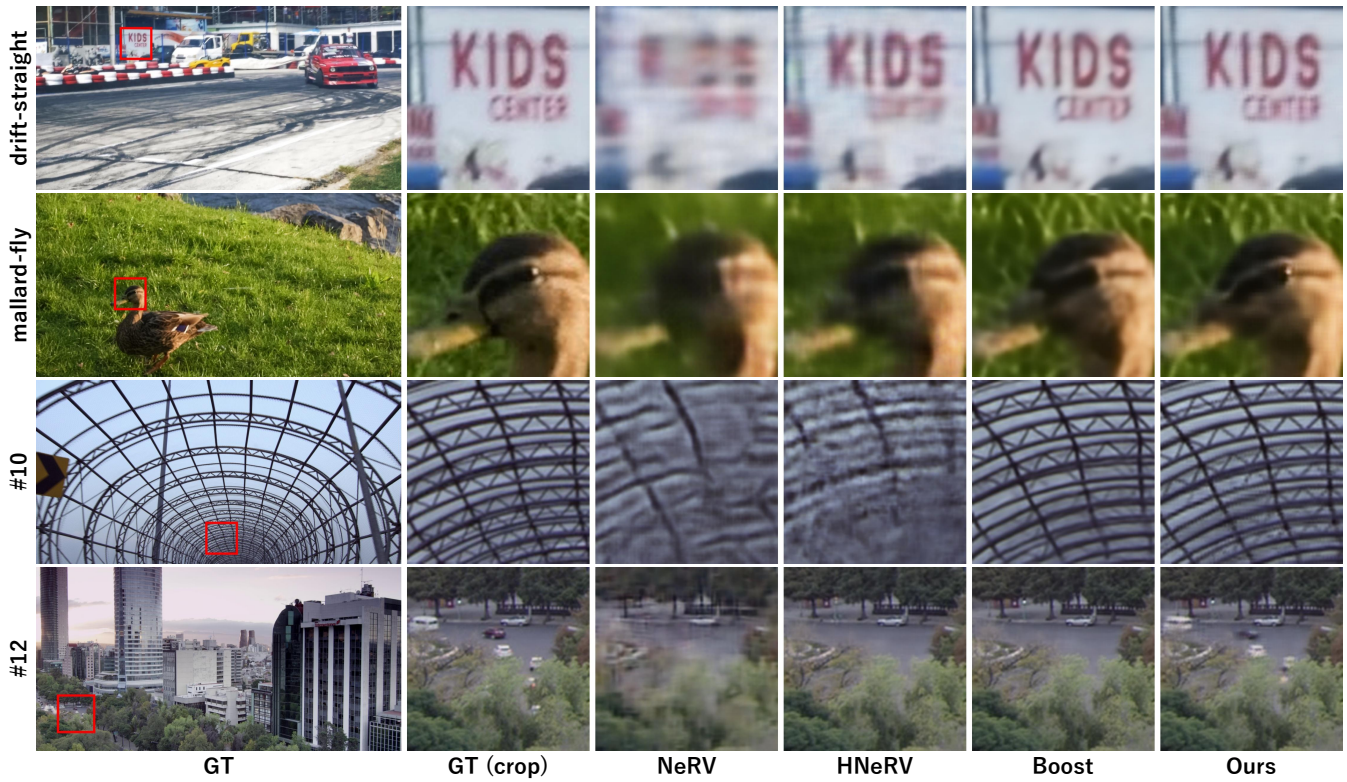

Fig. 4. Visualization of reconstructed frames from different sequences. From top to bottom: "drift-straight" and "mallard-fly" from the DAVIS dataset, and sequences #10 and #12 from the MCL-JCV dataset. Our structure-preserving patch (SPP) based method maintains sharp edges and consistent spatial structure across frames, demonstrating effective reconstruction quality across diverse scenes.

TABLE I
COMPARISON OF RECONSTRUCTED VIDEO QUALITY ON DAVIS DATASET (1.5M, 640 × 1280)

Method	PSNR ↑	MS-SSIM ↑	LPIPS ↓
NeRV	28.60	0.8811	0.4150
HNeRV	30.69	0.9146	0.3476
Boost	<u>33.53</u>	<u>0.9604</u>	<u>0.2674</u>
Ours	**34.23**	**0.9643**	**0.2539**

TABLE II
COMPARISON OF RECONSTRUCTED VIDEO QUALITY ON MCL-JCV DATASET (1.5M, 640 × 1280)

Method	PSNR ↑	MS-SSIM ↑	LPIPS ↓
NeRV	31.64	0.9217	0.4126
HNeRV	33.47	0.9417	0.3571
Boost	<u>35.60</u>	<u>0.9638</u>	<u>0.3039</u>
Ours	**35.94**	**0.9654**	**0.2936**

D. Loss Function

For each patch i of a frame, we define a composite loss $\mathcal{L}_i$ that accounts for spatial accuracy, perceptual quality, and frequency-domain consistency:

$$\mathcal{L}_i = w_i(\alpha\mathcal{L}_1(x_i,\hat{x}_i)+\beta\mathcal{L}_{\text{MS-SSIM}}(x_i,\hat{x}_i))+\mathcal{L}_{\text{freq}}(x_i,\hat{x}_i), \quad (2)$$

$$\mathcal{L}_{\text{freq}}(x,\hat{x}) = \mathcal{L}_1(\text{FFT}(x),\text{FFT}(\hat{x})), \quad (3)$$

where x_i and $\hat{x}_i$ denote the ground-truth and predicted patch images, respectively. The term $\mathcal{L}1$ is the standard L1 pixel loss, and $\mathcal{L}$MS-SSIM is a multi-scale structural similarity loss, which promotes perceptual quality. The coefficients α and β balance their relative contributions. In addition, The frequency-domain term $\mathcal{L}_{\text{freq}}$ is computed by applying the Fast Fourier Transform (FFT) to both images and measuring their L1 difference. In our SPP-based decoding scheme, patch-specific variations are adjusted only at the later layers of the decoder. This localized modulation makes it difficult for the model to learn patches that differ significantly from their neighbors. To address this, we introduce an adaptive patch weighting factor w_i as follows:

$$w_i = \frac{\sum_{j\neq i}\|x_i - x_j\|_1}{\sum_k\sum_{j\neq k}\|x_k - x_j\|_1 + \epsilon}, \quad (4)$$

where ϵ is a small constant to avoid division by zero. This formulation normalizes the weights across all patches. The model focuses more on learning consistent structures while maintaining flexibility for difficult regions.

IV. EXPERIMENT

We evaluate the effectiveness of the proposed SPP-based video representation method across four benchmark datasets: Bunny [32], DAVIS [33], MCL-JCV [34], and UVG [35]. Our

TABLE III
COMPARISON OF RECONSTRUCTED VIDEO QUALITY ON UVG DATASET IN PSNR/MS-SSIM (3M, 1080 × 1920)

Method	Beauty	Bosphorus	HoneyBee	Jockey	ReadySetGo	ShakeNDry	YachtRide	Average
NeRV	32.93 / 0.8843	33.09 / 0.9288	37.08 / 0.9780	31.06 / 0.8767	24.66 / 0.8205	32.67 / 0.9264	27.75 / 0.8573	31.22 / 0.8937
HNeRV	33.18 / 0.8876	17.57 / 0.5885	38.97 / 0.9838	31.75 / 0.8884	25.09 / 0.8358	33.73 / 0.9337	27.95 / 0.8546	29.44 / 0.8470
Boost	33.70 / 0.8980	35.81 / 0.9631	39.52 / 0.9852	33.84 / **0.9253**	27.72 / **0.9070**	35.55 / 0.9532	29.03 / **0.8967**	33.45 / 0.9311
Ours	**33.82 / 0.8992**	**36.08 / 0.9644**	**39.60 / 0.9854**	**34.21** / 0.9249	**28.03** / 0.9054	**35.96 / 0.9584**	**29.33** / 0.8945	**33.70 / 0.9312**

TABLE IV
PSNR OVER TRAINING EPOCHS ON THE BUNNY VIDEO
(1.5M, 640 × 1280)

Epoch	50	100	150	200	250	300
NeRV	26.95	29.40	30.43	30.96	31.14	31.33
HNeRV	29.93	33.32	34.86	35.61	36.06	36.35
Boost	34.90	37.35	38.12	38.62	38.70	39.03
Ourts	**37.80**	**38.62**	**39.00**	**39.18**	**39.31**	**39.40**

method is compared with existing implicit neural video representation baselines, including NeRV, HNeRV, and Boosting-HNeRV ("Boost"). The Bunny dataset consists of a single sequence with a resolution of 720 × 1280 and 132 frames. To ensure stable training, the frames are cropped to 640 × 1280. This adjustment is made because HNeRV failed to converge when trained on high-resolution input, as observed in the "Bosphorus" sequence in Table III. The DAVIS dataset consists of 50 natural video sequences with resolution 1080×1920 (cropped 640 × 1280). Each sequence is relatively short (25 to 104 frames). Similarly, MCL-JCV contains 30 sequences at 1080 × 1920 (cropped to 640 × 1280), with lengths ranging from 120 to 150 frames. The UVG dataset comprises 7 long sequences at full HD resolution (1080×1920), each containing either 300 or 600 frames.

We adjust the stride configuration in the decoder to match the spatial resolution: the stride list is set to [5, 4, 2, 2, 2] for 640 × 1280, and [5, 3, 2, 2] for 1080 × 1920. The number of patches is fixed at $N = 4$, and we empirically set the loss weighting coefficients to $\alpha = 42$ and $\beta = 18$. The default model size is 1.5M parameters, except for the UVG dataset at 1080 × 1920 resolution, where a 3.0M parameter model is used. For quantitative evaluation, we use Peak Signal-to-Noise Ratio (PSNR), Multi-Scale Structural Similarity (MS-SSIM), and Learned Perceptual Image Patch Similarity (LPIPS). In the video compression setting, we use the UVG dataset with a resolution of 1080 × 1920 and subsample each sequence to 60 or 120 frames. Each video is encoded into multiple models of varying sizes by adjusting the network width. We apply quantization and entropy coding to these models to simulate real-world compression, enabling rate-distortion (RD) analysis.

Quantitative results on DAVIS and MCL-JCV are summarized in Tables I and II, respectively. These tables show the average reconstruction quality across each datasets. Table III reports PSNR and MS-SSIM scores for each sequence in the UVG dataset. Our method consistently achieves higher scores

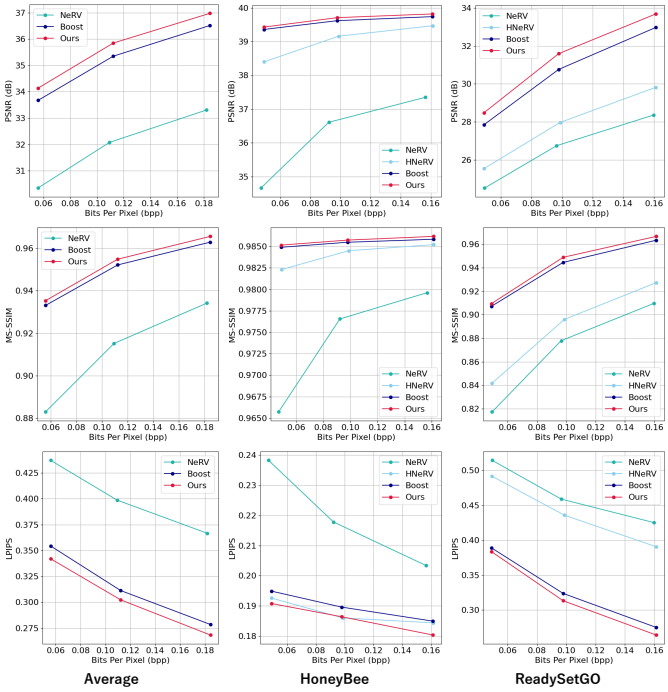

Fig. 5. Compression result on UVG dataset.

in most sequences, demonstrating superior reconstruction quality compared to baseline methods. In particular, Table IV presents PSNR values at each training epoch for the Bunny dataset, demonstrating not only improved reconstruction performance, but also faster convergence compared to competing methods. Qualitative comparisons are shown in Fig. 4, which visualizes representative frames from DAVIS and MCL-JCV. Our method better preserves sharp edges and spatial coherence across a variety of scenes. Finally, Fig. 5 presents the rate-distortion curves obtained from UVG. The proposed method achieves superior compression performance relative to the baselines. Note that HNeRV is excluded from the average curve due to repeated convergence failures during training.

V. CONCLUSION

We propose a neural video representation method based on structure-preserving patches (SPPs). In our approach, each video frame of a video is rearranged into spatially consistent patch images, enabling reconstruction that preserves global structure. This strategy mitigates the degradation commonly seen in conventional patch-based decoding. To facilitate

smooth reconstruction, we designed a global-to-local decoder architecture. Early decoding layers are conditioned on the temporal index to model the global structure, while later layers incorporate patch-specific information to enhance local details. This separation allows the network to balance structural coherence and fine-grained fidelity effectively. Experimental results demonstrate that our method achieves higher reconstruction quality than existing NeRV-based baselines. These findings highlight the potential of structure-aware patch design for improving the fidelity and efficiency of neural video representations. In future work, we plan to explore more flexible adaptive patch layouts to further enhance spatial coherence and compression performance.

REFERENCES

[1] B. Mildenhall, P. Srinivasan, M. Tancik, J. Barron, R. Ramamoorthi, and R. Ng, "NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis," European Conference on Computer Vision (ECCV), 2020, pp. 405-421.

[2] J. Barron, B. Mildenhall, M. Tancik, P. Hedman, R. Martin-Brualla, and P. Srinivasan, "Mip-NeRF: A Multiscale Representation for Anti-Aliasing Neural Radiance Fields," IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 5855-5864.

[3] A. Pumarola1, E. Corona1, G. Pons-Moll, and F. Moreno-Noguer, "D-NeRF: Neural Radiance Fields for Dynamic Scenes," IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 10318-10327.

[4] R. Martin-Brualla, N. Radwan, M. Sajjadi, J. Barron, A. Dosovitskiy, and D. Duckworth, "NeRF in the Wild: Neural Radiance Fields for Unconstrained Photo Collections," IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 7210-7219.

[5] E. Dupont, A. Goliński, M. Alizadeh, Y. Teh, and A. Doucet, "COIN: COmpression with Implicit Neural representations," arXiv preprint arXiv:2103.03123, 2021.

[6] E. Dupont, H. Loya, M. Alizadeh, A. Goliński, Y. Teh, and A Doucet, "COIN++: Neural Compression Across Modalities," arXiv preprint arXiv:2201.12904, 2022.

[7] T. Ladune, P. Philippe, F. Henry, G. Clare, and T Leguay, "COOL-CHIC: Coordinate-based Low Complexity Hierarchical Image Codec," IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 13515-13522.

[8] H. Kim, M. Bauer, L. Theis, J. Schwarz, and E. Dupont, "C3: High-performance and low-complexity neural compression from a single image or video,"

[9] H. Chen, B. He, H. Wang, Y. Ren, S. Lim, and A. Shrivastava, "NeRV: Neural Representations for Videos," Advances in Neural Information Processing Systems (NeurIPS), 2021, pp. 21557-21568.

[10] L. Zizhang, W. Mengmeng, P. Huaijin, X. Kechun, M. Jianbiao, and L. Yong, "E-NeRV: Expedite Neural Video Representation with Disentangled Spatial-Temporal Context," European Conference on Computer Vision (ECCV), 2022, pp. 267-284.

[11] H. Chen, M. Gwilliam, S. Lim, and A. Shrivastava, "HNeRV: A Hybrid Neural Representation for Videos," IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 10270-10279.

[12] S. Vincent, M. Julien, B. Alexander, L. David, and W. Gordon, "Implicit Neural Representations with Periodic Activation Functions," Advances in Neural Information Processing Systems (NeurIPS), 2020, pp. 7462-7473.

[13] A. Kayabasi, A. Vadathya, G. Balakrishnan, and V. Saragadam, "Bias for Action: Video Implicit Neural Representations with Bias Modulation," the Computer Vision and Pattern Recognition Conference (CVPR), 2025, pp. 27999-28008.

[14] Y. Bai, C. Dong, and C. Wang, "PS-NeRV: Patch-wise Stylized Neural Representations for Videos," IEEE International Conference on Image Processing (ICIP), 2023, pp. 41-45.

[15] S. Maiya, S. Girish, M. Ehrlich, H. Wang, K. Lee, P. Poirson, P. Wu, C. Wang, and A. Shrivastava, "NIRVANA: Neural Implicit Representations of Videos with Adaptive Networks and Autoregressive Patch-wise Modeling," IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 14378-14387.

[16] N. Rahaman, A. Baratin, D. Arpit, F. Draxler, M. Lin, F. Hamprecht, Y. Bengio, and A. Courville, "On the Spectral Bias of Neural Networks," Proceedings of Machine Learning Research, 5301-5310, Jun. 2019.

[17] Y. Cao, Z. Fang, Y. Wu, D. Zhou, and Q. Gu, "Towards Understanding the Spectral Bias of Deep Learning," arXiv preprint arXiv:1912.01198, 2020.

[18] Z. Cai, H. Zhu, Q. Shen, X. Wang, and X. Cao, "Batch Normalization Alleviates the Spectral Bias in Coordinate Networks," IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 25160-25171, Jun. 2024.

[19] Z. Liu, H. Zhu, Q. Zhang, J. Fu, W. Deng, Z. Ma, Y. Guo, X. Cao, "FINER: Flexible Spectral-bias Tuning in Implicit NEural Representation by Variable-periodic Activation Functions," IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 2713-2722.

[20] X. Huang, and S. Belongie, "Arbitrary Style Transfer in Real-time with Adaptive Instance Normalization," IEEE International Conference on Computer Vision (ICCV), 2017, pp. 1501-1510.

[21] X. Zhang, R. Yang, D. He, X. Ge, T. Xu, Y. Wang, H. Qin, and J. Zhang, "Boosting Neural Representations for Videos with a Conditional Decoder," IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 2556-2566.

[22] Q. Zhao, M. S. Asif, and Z. Ma, "DNeRV: Modeling Inherent Dynamics via Difference Neural Representation for Videos," IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 2031-2040.

[23] Y. Hao, K. Zhihui, Z. Xiaobo, Q. Tie, S. Xidong, and J. Dadong, "DS-NeRV: Implicit Neural Video Representation with Decomposed Static and Dynamic Codes," IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 23019-23029.

[24] H. Kwan, G. Gao, F. Zhang, A. Gower, and D. Bull, "HiNeRV: Video Compression with Hierarchical Encoding-based Neural Representation," Advances in Neural Information Processing Systems (NeurIPS), 2023, pp. 72692-72704.

[25] Q. Zhao, M. Asif, and Z. Ma, "PNeRV: Enhancing Spatial Consistency via Pyramidal Neural Representation for Videos," IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2024, pp. 19103-19112.

[26] M. Tarchouli, T. Guionnet, M. Riviere, W. Hamidouche, M. Outtas, and O. Deforges, "Res-NeRV: Residual Blocks For A Practical Implicit Neural Video Decoder," IEEE International Conference on Image Processing (ICIP), 2024, pp. 3751-3757.

[27] T. Hayami and H. Watanabe, "Implicit Neural Representation for Videos Based on Residual Connection," IEEE Global Conference on Consumer Electronics (GCCE), 2024, pp. 317-318.

[28] T. Hayami, T. Shindo, S. Akamatsu, and H. Watanabe, "Neural Video Representation for Redundancy Reduction and Consistency Preservation," IEEE International Conference on Consumer Electronics (ICCE), 2025, pp. 1-6.

[29] J. Kim, J. Lee, and J. Kang, "SNeRV: Spectra-preserving Neural Representation for Video," European Conference on Computer Vision (ECCV) 2024, pp. 332-348.

[30] T. Hayami, K. Kakeru, H. Watanabe, "SR-NeRV: Improving Embedding Efficiency of Neural Video Representation via Super-Resolution," arXiv preprint arXiv:2505.00046, 2025.

[31] Z. Liu, H. Mao, C. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A ConvNet for the 2020s," IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 11976-11986.

[32] T. Roosendaal, "Big buck bunny," ACM SIGGRAPH ASIA 2008 computer animation festival. 2008, pp. 62-62.

[33] F. Parazzi, J. Pont-Tuset, B. McWilliams, L. Van Gool, M. Gross, and A. Sorkine-Hornung, "A Benchmark Dataset and Evaluation Methodology for Video Object Segmentation," IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 724-732.

[34] H. Wang, W. Gan, S. Hu, J. Lin, L. Jin, L. Song, P. Wang, I. Katsavounidis, A. Aaron, and C. Kuo, "MCL-JCV: A JND-based H.264/AVC video quality assessment dataset," IEEE International Conference on Image Processing (ICIP), 2016, pp. 1509-1513.

[35] A. Mercat, M. Viitanen, and J. Vanne, "UVG dataset: 50/120fps 4K sequences for video codec analysis and development," ACM Multimedia Systems Conference, 2020, pp. 297-302.