PNeRV: Enhancing Spatial Consistency via Pyramidal Neural Representation for Videos

Qi Zhao
Nanjing University
qizhao@smail.nju.edu.cn

M. Salman Asif
University of California Riverside
sasif@ucr.edu

Zhan Ma*
Nanjing University
mazhan@nju.edu.cn

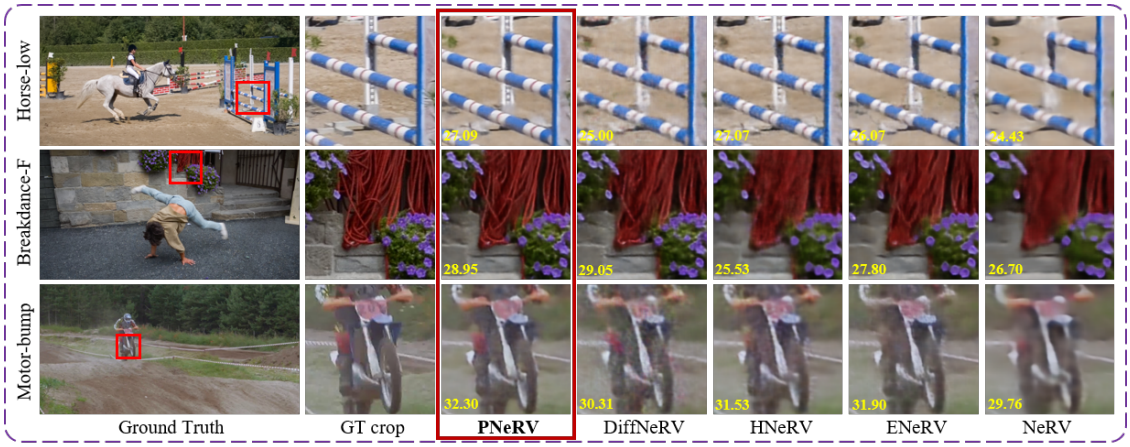

Figure 1. High-quality video (1920 × 960) reconstruction comparisons between the proposed **P**yramidal **NeRV** and other models, PSNR in yellow. PNeRV outperforms other models on perceptual quality with less noise and artifacts, maintaining spatial consistency.

Abstract

The primary focus of Neural Representation for Videos (NeRV) is to effectively model its spatiotemporal consistency. However, current NeRV systems often face a significant issue of spatial inconsistency, leading to decreased perceptual quality. To address this issue, we introduce the Pyramidal Neural Representation for Videos (PNeRV), which is built on a multi-scale information connection and comprises a lightweight rescaling operator, Kronecker Fully-connected layer (KFc), and a Benign Selective Memory (BSM) mechanism. The KFc, inspired by the tensor decomposition of the vanilla Fully-connected layer, facilitates low-cost rescaling and global correlation modeling. BSM merges high-level features with granular ones adaptively. Furthermore, we provide an analysis based on the Universal Approximation Theory of the NeRV system and validate the effectiveness of the proposed PNeRV. We conducted comprehensive experiments to demonstrate that PNeRV surpasses the performance of contemporary NeRV models, achieving the best results in video regression on UVG and DAVIS under various metrics (PSNR, SSIM, LPIPS, and FVD). Compared to vanilla NeRV, PNeRV achieves a +4.49 dB gain in PSNR and a 231% increase in FVD on UVG, along with a +3.28 dB PSNR and 634% FVD increase on DAVIS.

1. Introduction

In recent years, Implicit Neural Representation (INR) has emerged as a pivotal area of research across various vision tasks, including neural radiance fields modeling [35, 51], 3D vision [6, 40, 45] and multimedia neural coding [7, 42]. INR operates on the philosophy that target implicit mapping will be encoded into a learnable neural network through end-to-end training. By leveraging the modeling capabilities of neural nets, INR can approximate a wide range of complex nonlinear or high-dimensional mappings.

However, when considering the video coding task, extant NeRV systems exhibit a notable deficiency in perceptual quality. The reconstructions of foreground subjects, which are obscured by high-frequency irrelevant details or blurring, prove challenging for current NeRV models. This issue of spatial inconsistency is primarily attributed to se-

[1]Corresponding author: Zhan Ma (mazhan@nju.edu.cn)

mantic uncertainty, causing the model to struggle with discerning whether two long-range pixels pertain to the same objects or constitute part of a noisy background. We postulate that this predicament stems from the absence of **global receptive field** and **multi-scale information communication**. Inspired by existing empirical evidence from other vision research, we speculate that if the dense prediction could leverage the high-level information learned from raw input, it would substantially alleviate both the semantic uncertainty and spatial inconsistency (as illustrated in Fig. 1).

In practice, introducing multi-scale structures into NeRV poses a significant and non-trivial challenge. Existing NeRV models typically resort to cascaded upsampling layers (the so-called "mainstream") for decoding fine video, striking a compromise between performance and efficiency. However, layers that use subpixel-based operators [41, 53] can hardly maintain a balance between the increasing receptive field, parameter demand, and performance (more discussions in Sec. **??** and visualization in Fig. 2). Additionally, these decoding layers are solely receptive to features from the previous layer, ignoring information from other preceding layers. Moreover, the design of multi-scale structures in NeRV remains unguided by either practical or theoretical principles due to constraints on parameter quantities compared with methods for other vision tasks.

To address this issue, we propose the **P**yramidal **Ne**ural **R**epresentation for **V**ideos (**PNeRV**) based on hierarchical information interaction via a low-cost upscaling operator, *Kronecker Fully-connected* (KFC) layer, and a gated mechanism, *Benign Selective Memory* (BSM), which aims at adaptive feature merging. Utilizing these modules, PNeRV can fuse the high-level features directly into each underlying fine-grained layer via shortcuts, thereby creating a pyramidal structure. Further, we introduce *Universal Approximation Theory* (UAT) into the NeRV system for the first time and provide an analysis of existing NeRV models, revealing the superiority of our proposed pyramid structure. Our main contributions are summarized as follows.

- Towards the poor perceptual quality of NeRV systems, we propose PNeRV to enhance spatial consistency via multi-scale feature learning.
- In pursuit of model efficiency pursuit, we propose the KFC, which realizes low-cost upsampling with a global receptive field and BSM for adaptive feature fusion, thus forming an efficient multi-level pyramidal structure.
- We introduce the first UAT analysis in NeRV research. Using UAT, we describe NeRV-based video neural coding as the Implicit Video Neural Coding problem, clarifying and defining some fundamental concepts within this framework.
- We confirm the superiority of PNeRV against other models on two datasets (UVG and DAVIS) using four video quality metrics (PSNR, SSIM, LPIPS, and FVD).

2. Related Work

Implicit Neural Representation for Videos. In recent years, INR has gained increasing attention in various vision areas, such as neural radiance fields modeling [9, 13, 35, 36], novel view synthesis [22, 34], and multimedia neural coding [7, 8, 11, 24, 55]. For INR-based neural video coding, NeRV [7] first uses index embeddings as input and then decodes back to high-resolution videos via cascaded PixelShuffle [41] blocks. ENeRV [24] aims to reallocate the parameter quantity between different modules for better performance. Unlike the above index-based methods, HNeRV [8] employs ConvNeXT [30] blocks as an encoder and provides *content-aware embeddings*, improving the performance. Furthermore, apart from content embedding, DiffNeRV [55] inputs the difference between adjacent frames as *temporal embeddings*, enhancing temporal consistency. The major distinction between PNeRV and DiffNeRV is that the latter does not refer to multi-scale spatial information, resulting in spatial discontinuity.

Multi-scale Hierarchy Structure for Dense Prediction. In previous CV research, there have emerged numerous studies on multi-scale vision [5, 18, 27–29, 31, 39, 48, 56]. UNet [39] aimed to improve accuracy by combining contextual information from features at different resolutions. FPN [27] developed a top-down architecture with high-level semantic feature maps at all scales, showing significant improvements in dense prediction tasks. PANet [28] followed the idea of multi-level information fusion and proposed adaptive feature pooling to leverage useful information from each level. PVT [48] introduced the pyramidal architecture into vision transformers. The success of pyramidal structure lies in multi-level feature fusion, and detailed predictions should be guided by high-level context features.

Video Coding Pipelines and Theories. Video coding has been studied for several decades based on handcrafted design and domain transformation [2, 14, 44, 50]. Furthermore, *neural video coding* [21, 23, 26, 32] aims to replace some components in the traditional pipeline, but they suffer from high computational complexity and slow decoding speeds. Beyond Rate-Distortion Optimization (RDO) [43], [3] reveals the importance of perceptual quality and proposes the Perception-Distortion Optimization (PDO). [4] defines the Rate-Distortion-Perception Optimization (RDPO). Different from those pipelines, we reinterpret the INR-based video coding [7, 8, 55] with UAT framework, and more details are in Sec. 4.2 and Sec. A.1.

Universal Approximation Theory (UAT). One of the pursuits of UAT analysis on the deep neural net is to estimate the minimal width of a model to approximate continuous functions under certain errors and fixed lengths. [16] provides the estimation of minimal width w^* of a RELU net as $d_{in}+1 \leq w^* \leq d_{in}+d_{out}$ in Theorem 1. [37] provides the first definitive result for deep ReLU nets, and the minimum

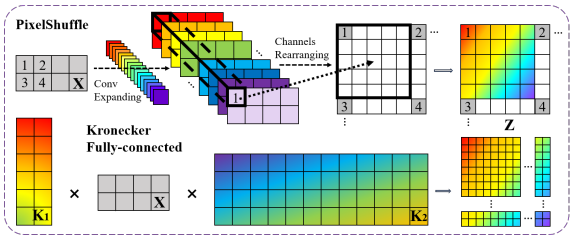

Figure 2. Visualized comparison between PixelShuffle and KFc, where $\times$ denotes matrix multiplication and black box is the subpixel area. PixelShuffle fills the subpixels using a local receptive field, lacking long-range relationship modeling ability, while KFc calculates the correlation between every position.

width required for the universal approximation of the L^p functions is exactly $\max\{d_{in}+1, d_{out}\}$. [25] demonstrates that a deep ReLU ResNet with one neuron per hidden layer can uniformly approximate any Lebesgue integrable function. More discussions are given in Sec. 4 and Sec. A.1.

3. Pyramidal Neural Representation for Videos

As analyzed before, pursuing spatial consistency leads to the communication of multi-scale information via a global receptive field. Fine-grained reconstruction requires high-level information as guidance and a low-cost upsampling operator is crucial for creating multi-level shortcuts.

Therefore, we propose **P**yramidal **NeRV** (**PNeRV**) consisting of a learnable encoder and a novel pyramidal decoder. The main innovation in the decoder is a low-cost global-wise upscaling operator, Kronecker Fully-connected (KFc) layer, and a gated memory unit, Benign Selective Memory (BSM) for disentangled feature fusion. The overall structure of PNeRV is shown in Fig. 3.

3.1. Kronecker Fully-connected Layer

NeRV aims to decode high-resolution videos from tiny embeddings. Therefore, Conv-based upsampling operators [41, 53] are not efficient enough due to the huge upscaling ratio, which differs from previous visual tasks. The parameter quantity will grow sharply due to increased channels or kernel size. However, NeRV aims to encode videos with as few parameters as possible, namely *model efficiency pursuit*.

In contrast to this goal, subpixel-based upscaling operators fail to form shortcuts and a pyramidal structure. Once upscaling from given embeddings F_0 ($16 \times 2 \times 4$) to fine-grained features F_n ($16 \times 320 \times 640$), there is an intolerable increase in parameters ($25600\times$) to fill in the target subpixels. Even when the kernel size is only 1×1, a single PixelShuffle [41] layer requires 6.96M parameters from F_0 to F_n, regardless of the size of videos or model structure.

Towards this dilemma, we propose the Kronecker Fully-connected layer (KFc), given as

$$\mathbf{Z} = \underset{i}{\text{CONCAT}} \left(\mathbf{K}_1^{(i)} \mathbf{X}^{(i)} \mathbf{K}_2^{(i)} \right) + \mathbf{b}_c \otimes \mathbf{b}_h \otimes \mathbf{b}_w, \quad (1)$$

where $\mathbf{X}^{(i)} \in \mathbb{R}^{H_{in} \times W_{in}}$ are input features, $\mathbf{Z}^{(i)} \in \mathbb{R}^{H_{out} \times W_{out}}$ are output features, $\mathbf{K}_{1,2}$ are two kernels which $\mathbf{K}_1^{(i)} \in \mathbb{R}^{H_{out} \times H_{in}}$ and $\mathbf{K}_2^{(i)} \in \mathbb{R}^{W_{in} \times W_{out}}$ in channel i. Each feature map is calculated channel-wise and will be concatenated in the channel. $\mathbf{b}_{c,h,w}$ are three vectors and they output the BIAS via kronecker product $\otimes$ where $\mathbf{b}_c \in \mathbb{R}^{C \times 1}$, $\mathbf{b}_h \in \mathbb{R}^{H_{out} \times 1}$ and $\mathbf{b}_w \in \mathbb{R}^{W_{out} \times 1}$.

Motivation. KFc is motivated by the fact that, *the subpixels of one position are related to every other position in current feature maps*. The dilemma between local and global feature learning is an enduring issue in deep learning [31, 47, 49, 52]. Unlike the local prior in the CONV layer, FC is more effective, especially for the top embeddings containing semantic features with little local spatial structure. The calculation between $\mathbf{K}_1$, $\mathbf{X}$ and $\mathbf{K}_2$ is actually the product between vectorized input features $\text{vec}(\mathbf{X}) \in \mathbb{R}^{H_{in}W_{in} \times 1}$ and hybrid weight matrix $\mathbf{K}_\otimes \in \mathbb{R}^{H_{out}W_{out} \times H_{in}W_{in}}$, where $\mathbf{K}_\otimes = \mathbf{K}_1 \otimes \mathbf{K}_2^\top$. Compared with the vanilla FC layer, two low-rank matrices $\mathbf{K}_1$ and $\mathbf{K}_2$ come from the Kronecker decomposition, while the bias term $\mathbf{b}_{c,h,w}$ is the CP decomposition of the original ones.

Besides, KFc is also inspired by LoRA [20], which uses adaptive weights in low "intrinsic dimension" [1] for PEFT. Visualization is shown in Fig. 2. For the same F_0 and F_n mentioned above, parameters needed by KFc is 0.05M, only **0.7%** of that required by PixelShuffle. Detailed comparisons of parameters and FLOPs are given in Fig. 3.

3.2. Benign Selective Memory

Using KFc as the basic operator for shortcuts, PNeRV realizes efficient multi-scale feature learning. Also, adaptive feature fusion between different levels is quite important.

Therefore, we propose the Benign Selective Memory (BSM). BSM is inspired by the gated mechanism in RNN research [10, 19], treating features in different streams as input and cell states. We follow the convention in RNN, where lowercase represents hidden states. For the high-level feature z on the top and the fine-grained feature h_{l-1} in the l-th layer, BSM is given as follows:

$$\begin{aligned} n_l &= W_n * z, && \text{KOWNLEDGE} \\ m_l &= W_m * h_{l-1} && \text{MEMORY} \\ s_l &= \sigma(W_s * \text{RELU}(n_l + m_l)), && \text{DECISION} \\ h_l &= h_{l-1} \odot (1 - s_l) + n_l \odot s_l, && \text{BEHAVIOUR} \end{aligned}$$

where $*$ is convolution with weights $W_{n,m,s}$,$\odot$ is hadamard product and σ is the sigmoid activation.

BSM is an imitation of the human learning and decision-making process. The high-level z is regarded as external Knowledge, while h_{l-1} from the previous block in the mainstream is the inheriting Memory. The model should learn from Knowledge and integrate it with Memory to guide the Behaviors (reconstruction). That is the so-called Benign Selective Memory.

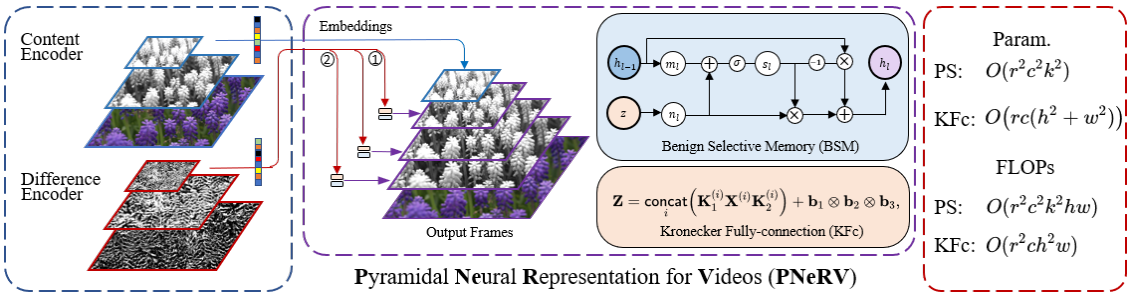

Figure 3. The overall architecture of PNeRV, consists of KFc and BSM. The right part shows the comparison of parameters and FLOPs between PixelShuffle (PS) and KFc, where input feature maps are in $c \times h \times w$, the upscaling rate is r and kernel size in PS is $k \times k$.

Motivation. The primary distinction between previous gated mechanisms and BSM is that BSM learns features (referred to as "Knowledge" and "Memory") separately before merging them. This disentangled fashion aids PNeRV in adaptively merging features from different levels. The ablation studies in Tab. 7 show the superiority of BSM.

3.3. Overall Structure

Therefore, the proposed PNeRV consists of three parts, as follows (where X is the input embedding, $\hat{H}_l$ are featured in the mainstream l^{th} layer, Z_l are features upsampled by shortcuts, and H_l are the features after fusion):

1. A mainstream comprises cascaded upsampling layers (containing Conv, PixelShuffle, and GELU) to provide high-resolution reconstruction, $\hat{H}_l = Block(H_{l-1}), 1 \le l \le L, L = 6, H_0 = X$.
2. Various shortcuts upsample the high-level embeddings X into Z_l before merging into the mainstream, forming a multi-level hierarchical architecture, $Z_l = Shortcut(X), 2 \le l \le L_0, L_0 = 5$.
3. A feature fusion mechanism is employed to merge Z_l with $\hat{H}_l$ adaptively for the final output, $H_l = Fusion(Z_l, \hat{H}_l)$.

In implementation, we conducted two versions, namely PNeRV-M and PNeRV-L. PNeRV-M has only a single stream which takes content embeddings [8] $\mathbf{X}^C$ in $16\times2\times4$ as input. For PNeRV-L, temporal embeddings [55] $\mathbf{X}^T$ in $2 \times 40 \times 80$ are involved. $\mathbf{X}^C$ is delivered to the mainstream and $\mathbf{X}^T$ is upscaled in shortcuts via KFc and merged into each mainstream layers through BSM. We choose PNeRV-L as the final version. All kernels are 3×3 except for the first and final output layer. For the input video $\mathbf{V}$ and reconstructions $\tilde{\mathbf{V}}$, the key equations of the entire model in the l-th layer ($1 < l \le L$) are presented as follows:

$$
\begin{aligned}
\textit{Encoder}: \ & \mathbf{X}^C, \mathbf{X}^T = \mathcal{E}(\mathbf{V}), \\
\textit{Decoder}: \ & \hat{\mathbf{H}}_l = \textsc{Block}_l(\mathbf{H}_{l-1}) \\
& \quad = \textsc{Block}_l \circ \cdots \circ \textsc{Block}_1(\mathbf{X}^C), \\
& \mathbf{Z}_l = \textsc{Shortcut}_l(\mathbf{X}^T), \\
& \mathbf{H}_l = \textsc{BSM}_l(\hat{\mathbf{H}}_l, \mathbf{Z}_l),
\end{aligned}
$$

where $\mathbf{H}_0 = \mathbf{X}^C$. The final output will be passed through an output layer, $\tilde{\mathbf{V}} = \textsc{Conv}_{1\times1}(\mathbf{H}_{l=L})$.

4. Universal Approximation Theory on NeRV

First, we will clarify some concepts in NeRV within the UAT framework. A NeRV-based neural video coding pipeline is defined in Sec. 4.2. We describe the limitations of existing NeRV models in Sec. 4.3, discuss the significance of shortcuts and the multi-scale structure in the proposed PNeRV in Sec. 4.4.

4.1. Basic Definitions and Notations

One of the main issues for the UAT analysis of a finite length L feed-forward network is to find out the minimal width $w^* := \min\max d_i, 1 \le i \le L$ where d_i is the width of the i-th layer so that neural nets with width w^* and length L can approximate any scalar continuous function arbitrarily well [15, 16, 37]. Following the statement in [16], a deep affine net is defined as follows.

Definition 1. (Deep Affine Net). A deep affine net of L layers is given as follows:

$$\mathcal{N} := A_L \circ \sigma \circ A_{L-1} \circ \cdots \circ \sigma \circ A_1. \tag{2}$$

where the i^{th} layer is an affine transformations $A_i := \mathbb{R}^{d_i} \to \mathbb{R}^{d_{i+1}}$, $d_1 = d_{in}$, $d_L = d_{out}$ with σ as activation.

In existing NeRV research, NeRV [7] and HNeRV [8] meet this definition.

4.2. Implicit Neural Video Coding

Recently, INR-based video coding has received increasing attention, and it uses a lightweight model to fit a video clip. We formulate this coding pipeline as *Implicit Neural Video Coding* (INVC), and the decoder with its embeddings together is known as the *NeRV system* [7, 8, 17, 24, 55].

Definition 2. (NeRV System). Each frame V_t in an RGB video clip $V = \{V_t\}_{t=1}^T \in \mathbb{R}^{T\times3\times H\times W}$ is represented by an implicit unknown continuous function $\mathcal{F} : [0,1]^{d_{in}} \to \mathbb{R}^{d_{out}}$ with the embedding $\mathcal{E}(t)$ obtained by encoder $\mathcal{E} : \mathbb{N} \to [0,1]^{d_{in}}$ on the t time stamp,

$$V_t = \mathcal{F} \circ \mathcal{E}(t),$$

where $\mathcal{F}$ can be approximated by a learnable neural network $\mathcal{D}$ of finite length $L_{\mathcal{D}}$, width $w_{\mathcal{D}}$ and activation σ. The reconstruction $\tilde{V}_t$ via $\mathcal{D}$ and $\mathcal{E}$ is given as follows:

$$\tilde{V}_t = \mathcal{D} \circ \mathcal{E}(t),$$

where the decoder $\mathcal{D}$ and embedding $\mathcal{E}(t)$ together are known as NeRV system, $\{\mathcal{D}, \mathcal{E}(t)\}_{t=1}^{T}$.

For the index-based models [7] and [24], the encoder $\mathcal{E}$ is Positional Encoding [42]. In content-based models [8, 55], $\mathcal{E}$ is learnable and provides content embeddings. When $\mathcal{D}$ is a deep affine net, it is named as a *serial cascaded NeRV system*, such as NeRV [7] and HNeRV [8], and $\mathcal{D}$ is formulated as follows, where B_l is the l-th upsampling layer.

$$\mathcal{D} := B_L \circ \sigma \circ B_{L-1} \circ \cdots \circ \sigma \circ B_1. \tag{3}$$

We present the proposed **I**mplicit **N**eural **V**ideo **C**oding **P**roblem (**INVCP**) as follows. More discussions between INVCP and existing pipelines are given in Sec. A.

Problem 1. *(INVCP). The goal of INVC is to obtain the minimal parameter quantity under a certain approximation error ϵ between input V and reconstruction $\tilde{V}$,*

$$\arg\min_{\mathcal{D},\mathcal{E}} \ \mathsf{Param}(\mathcal{D}) + \sum_{t=1}^{T} d_{in}^{t},$$
$$\text{s.t. } L_{\mathcal{D}}, w_{\mathcal{D}} \in [1, \infty), \ \sup \sum \|\tilde{V}_t - V_t\| \leq \epsilon, \ t \in [1, T].$$

where d_{in}^{t} is the dimension of embedding $\mathcal{E}(t)$ w.r.t. the t-th frame, $L_{\mathcal{D}}$ and $w_{\mathcal{D}}$ are the length and width.

4.3. UAT Analysis of Cascaded NeRV Model

For video INRs, the model strives to capture the implicit function that efficiently encodes a video. Within the UAT framework, a keen focus is on the smoothness properties of this implicit function, as it also encapsulates the video's inherent dynamics.

We name these properties as *rate of dynamics*, referring to the differences and transitions between consecutive frames within the video. We introduce $\omega_{\mathcal{V}}^{-1}$ to informally represent the rate of dynamics for video $\mathcal{V}$, inspired by the mathematical techniques used in UAT analysis [16].

Definition 3. The dual modulus of continuity ω_f^{-1} w.r.t. a continuous f defined on Ω is set as

$$\omega_f^{-1}(\epsilon) := \sup\{\delta : \omega_f(\delta) \leq \epsilon\},$$

where ω_f represents the modulus of continuity of f

$$\omega_f(\delta) := \sup_{x,y \in \Omega} \{\|f(x) - f(y)\| : d(x, y) \leq \delta\}.$$

Remark 1. *Using a function $\mathcal{F} : \mathbb{N} \to \mathbb{R}^{d_V}$ to roughly represent a video V, when the variation of frames (video dynamics) $\|\mathcal{F}(t_i) - \mathcal{F}(t_j)\|$ is at a certain level ϵ for two time stamps t_i and t_j, then the longer the duration sustains, the larger $\omega_{\mathcal{F}}^{-1}$ gets. Smoother video has larger $\omega_{\mathcal{F}}^{-1}$.*

Notably, the explicit calculation of ω_f^{-1} is hard to obtain, and it is more like an empirical judgment, such as camera movement, subject speed, noise, and others. We present the estimation of the upper bound of the minimal parameter quantity of the cascaded NeRV model as Theorem 1. The proof of Theorem 1 can be found in Sec. A.3.

Theorem 1. *For a cascaded NeRV system to ϵ-approximate a video V which is implicitly characterized by a certain unknown L-Lipschitz continuous function $\mathcal{F} : K \to \mathbb{R}^{d_{out}}$ where $K \subseteq \mathbb{R}^{d_{in}}$ is a compact set, then the upper bound of the minimal parameter quantity $\mathsf{Param}(\mathcal{D})$ is given as*

$$\mathsf{Param}_{\min}(\mathcal{D}) \leq d_{out}^2 \left(\frac{O\left(diam\left(K\right)\right)}{\omega_{\mathcal{F}}^{-1}\left(\epsilon\right)} \right)^{d_{in}+1}.$$

From Theorem 1, it can be seen that for a video, the fitting performance of the cascaded NeRV model depends on the rate of dynamics $\omega_{\mathcal{F}}^{-1}$ and the dimension of the video, d_{out}. The smoother and lower the dimension of the video to be modeled, the less difficult it is to approximate.

Remark 2. *The rate of dynamics for a given video will determine the performance of the NeRV system.*

4.4. UAT Analysis of PNeRV

According to Theorem 1, the upper bound of parameters of cascaded NeRV required for model fitting only depends on the properties of the target video. It demonstrates that, although different models can exhibit diverse architectures, their fitting behavior on the same video tends to be similar, indicating a limitation in the model's ability. However, according to observations in UAT research [12, 25], the model with shortcuts will reduce the maximum width to 1, indicating that the model size can be greatly reduced while maintaining the performance. Therefore, the involvement of **shortcuts** is the key to enhancing model capability.

Besides, we believe the implicit function representing a video can be decomposed into diverse sub-functions from a pattern-disentangled perspective. If we treat each stream in $\mathcal{D}$ as a sub net, the whole $\mathcal{D}$ is an ensemble,

$$\mathcal{D} := \sum A_L^{(i)} \circ \rho_{L-1}^{(i)} \circ A_{L-1}^{(i)} \circ \cdots \circ \rho_1^{(i)} \circ A_1^{(i)}. \tag{4}$$

Different shortcut pathways can fit various patterns, as a single shortcut has the universal approximation ability. For example, in Fig. 3, ① may capture the low-frequency motions. Whereas ②, directed towards fine-grained layers, signifies spatial details. This hypothesis aligns with the empirical evidence observed in other vision areas, which shows that the pyramid structure, a widely adopted hierarchical topology, can improve dense prediction tasks. That is why PNeRV outperforms others and achieves less semantic uncertainty and better perceptual quality.

Remark 3. *As the ensemble of sub-nets, the Pyramidal structure will enhance the perceptual quality of NeRV systems.*

5. Experiment

Settings. We perform video regression on 2 datasets, and all videos are center cropped to a 1×2 ratio. UVG [33] has 7 videos with a size of 960×1920 in 300 or 600 frames at 120 FPS. DAVIS [38] is a large dataset of 47 videos in 960×1920, containing large motions and complex spatial details. We choose 9 videos[1] from DAVIS as a subset, containing different types of spatiotemporal features.

Metrics. We use PSNR and MS-SSIM to evaluate pixel-wise errors. For spatial consistency, we choose the Learned Perceptual Image Patch Similarity (LPIPS) [54] and Frechet Video Distance (FVD) [46] as perceptual metrics, where LPIPS is based on AlexNet and FVD is based on the I3D model. The difference between PNeRV (P) and the baseline (B) is calculated as $(B - P)/B$ to show the improvement.

Training. We adopt Adam as the optimizer, where beta is (0.9, 0.999) and weight decay is 0. The learning rate is 5e-4 with a cosine annealing schedule. The loss function is L2, and the batch size is 1. All experiments are conducted using PyTorch 1.8.1 on NVIDIA GPU RTX2080ti, training for 300 epochs. We choose NeRV [7], E-NeRV [24], HNeRV [8], DivNeRV [17] and DiffNeRV [55] as baseline models. All models are trained with a similar 3M size, and we follow the setting of embedding size as the baseline method.

5.1. Video Regression on UVG

Pixel-wise error. PSNR comparison on UVG is reported in Tab. 1, where bold font is the best result and underline is the second best. PNeRV-L surpasses other models (+0.42 dB against DiffNeRV and +4.25 dB against NeRV). PNeRV-M achieves the best result against other single-stream models (+1.96 dB against HNeRV and +3.02 dB against NeRV). The proposed pyramidal architecture shows its effectiveness when combined with various encoders.

Perceptual quality. The perceptual results are given in Tab. 3 (LPIPS) and Tab. 4 (FVD), and the results of PNeRV show a significant improvement, especially for "Bospho" and "ShakeN". The FVD results in Tab. 4 indicate that PNeRV provides better spatiotemporal consistency compared to other baseline models (+231% against NeRV [7] and +64.5% against DiffNeRV [55]).

Case study. The visualized comparison on UVG is exhibited in the bottom three rows of Fig. 4. For dynamic objects with indistinct boundaries or noisy backgrounds, such as the horse in "ReadyS" and the tail in "ShakeN," PNeRV demonstrates superior visual quality without requiring additional semantic information.

Compared with the SOTA. As shown in Tab. 1, PNeRV obtained competitive PSNR results on dynamic and smooth videos. [17] is less effective for videos with fewer motions but complicated contextual spatial correlation. Also, [55] makes it hard to reconstruct the videos filled with high-frequency details. By comparison, PNeRV achieves comparable performance on all videos.

[1] Bmx-bumps, Camel, Dance-jump, Dog, Drift-chicane, Elephant, Parkour, Scooter-gray, Soapbox.

5.2. Video Regression on DAVIS

Pixel-wise error. In Tab. 2, we present the PSNR and SSIM comparison on the DAVIS dataset. PNeRV gains a +0.88 dB PSNR increase compared to DiffNeRV and +3.28 dB compared to vanilla NeRV. Despite the challenges posed by complex spatiotemporal features, PNeRV exhibits significant improvements (refer to "Parkour", which is the most difficult one, or "Drift-chicane", where the racing car undergoes intense motion amidst smoke-induced noise).

Perceptual quality. The LPIPS results on DAVIS are reported in Tab. 3, where PNeRV achieved a 32.0% increase compared to NeRV and 12.6% against the second-best DiffNeRV. In Tab. 5, PNeRV gains a 634% FVD increase over NeRV and 128% against DiffNeRV. For the worst case, "Dog", although PNeRV obtained a poor FVD result owing to the severe global blurring caused by camera motion, the PSNR is only slightly lower than the best (-0.24 db).

Case study. Visualizations are shown in Fig. 4. PNeRV reduced spatial inconsistency, particularly in "Dance Jump" and "Elephant," which are filled with irrelevant high-frequency details obscuring semantic clarity.

5.3. Ablation Studies

The ablation of the effectiveness of the proposed pyramidal architecture is in Tab. 6, and the contributions of two proposed modules are validated in Tab. 7, where the parameters of different models remain the same for a fair comparison.

Overall structure. We validate the design of the multi-level structure on the most dynamic and smooth videos ("Parkour" and "HoneyB"). In Tab. 6, the "serial" in the first row represents HNeRV [8]. "Pyram.+Concat." incorporates solely shortcuts without fusion modules. The main difference between DiffNeRV and PNeRV-L is the quantity of shortcuts (2 vs 5), and PNeRV-L performs better.

Modules contribution. We compare KFC with two upscaling layers, Deconv [53] and Bilinear (the combination of bilinear upsampling and Conv2D). KFC performs better due to the global receptive field, as shown in Tab. 7.

Also, we compare BSM with Concat, GRU [10] and LSTM [19]. The results suggest that, disentangled feature fusion significantly enhances performance. Detailed results for each video are listed in Tab. C.6 in the appendix.

5.4. Validation of Theoretical Analysis

The results in Tab. 1 and Tab. 6 validate the Remark 2. For those smooth videos with larger ω_f^{-1} and a smaller upper bound, models may obtain better performance; vice versa. The results of PNeRV in Fig. 4, which exhibit less noise and

PSNR ↑	D.P.	E.S.	Beauty	Bospho	HoneyB	Jockey	ReadyS	ShakeN	YachtR	Avg. M.
Avg. V.	N/A	N/A	36.06	35.32	39.48	33.27	27.53	35.27	30.03	N/A
NeRV [7]	3M	160	33.25	33.22	37.26	31.74	24.84	33.08	28.03	31.63
NeRV* [7]	3.2M	160	32.71	33.36	36.74	32.16	26.93	32.69	28.48	31.87
E-NeRV [24]	3M	160	33.17	33.69	37.63	31.63	25.24	34.39	28.42	32.02
HNeRV [8]	3M	128	33.58	34.73	38.96	32.04	25.74	34.57	29.26	32.69
DiffNeRV [55]	3.4M	6528	**40.00**	36.67	41.92	35.75	28.67	36.53	31.10	35.80
DivNeRV* [17]	3.2M	N/A	33.77	**38.66**	37.97	35.51	**33.93**	35.04	**33.73**	35.52
PNeRV-M	1.5M	128	37.51	33.80	41.76	29.96	24.15	36.18	28.92	33.18
	3M	128	39.08	35.56	42.59	31.51	25.94	37.61	30.27	34.65
PNeRV-L	1.5M	6528	37.98	35.18	41.78	34.43	27.28	36.65	28.29	34.51
	3.3M	6528	39.46	36.68	**42.73**	**35.81**	28.97	**38.25**	30.92	**36.12**

Table 1. PSNR comparison on UVG: the larger, the better. $*$ indicates methods that fit videos in a shared model while others fit each video in a single model. D.P. is the parameter quantity of the decoder, and E.S. is the corresponding embedding size per frame. Avg. V is the average PSNR across all models for the same video. Avg. M is the average PSNR for a single model on the entire dataset.

PSNR / SSIM ↑	Bmx-B	Camel	Dance-J	Dog	Drift-C	Elephant	Parkour	Scoo-gray	Soapbox	Avg.
NeRV [7]	29.42/0.864	24.81/0.781	27.33/0.794	28.17/0.795	36.12/0.969	26.51/0.826	25.15/0.794	28.16/0.892	27.68/0.848	27.99/0.840
E-NeRV [24]	28.90/0.851	25.85/0.844	29.52/0.855	30.40/0.882	39.26/0.983	28.11/0.871	25.31/0.845	29.49/0.907	28.98/0.867	29.62/0.878
HNeRV [8]	29.98/0.872	25.94/0.851	29.60/0.850	30.96/0.898	39.27/0.985	28.25/0.876	26.56/0.851	31.64/0.939	29.81/0.881	30.22/0.889
DiffNeRV [55]	30.58/0.890	27.38/0.887	29.09/0.837	**31.32/0.905**	**40.29**/0.987	27.30/0.848	25.75/0.827	30.35/0.923	**31.47/0.912**	30.39/0.890
PNeRV-L (ours)	**31.05/0.896**	**27.89/0.892**	**30.45/0.873**	31.08/0.898	40.23/**0.987**	**29.72/0.903**	**27.53/0.878**	**32.68/0.950**	30.85/0.902	**31.27/0.908**

Table 2. PSNR and MS-SSIM comparison on DAVIS.

LPIPS ↓	Beauty	Bospho	HoneyB	Jockey	ReadyS	ShakeN	YachtR	Avg.
NeRV [7]	0.229	0.203	0.043	0.251	0.326	0.189	0.276	0.216
ENeRV [24]	0.224	0.179	0.039	0.279	0.318	0.168	0.363	0.224
HNeRV [8]	0.218	0.172	0.042	0.270	0.348	0.191	0.253	0.213
DiffNeRV [55]	**0.205**	0.164	0.042	0.196	**0.206**	0.181	0.241	0.176
PNeRV (ours)	0.210	**0.132**	**0.037**	**0.177**	0.211	**0.146**	**0.230**	**0.163**

LPIPS ↓	Bmx-B	Camel	Dance	Dog	Drift	Eleph	Parko	Scoo-g	Soapb	Avg.
NeRV [7]	0.374	0.476	0.517	0.573	0.136	0.490	0.481	0.308	0.424	0.419
ENeRV [24]	0.386	0.357	0.426	0.404	0.061	0.419	0.429	0.282	0.380	0.349
HNeRV [8]	0.315	0.331	0.392	0.405	0.058	0.387	0.414	0.226	0.357	0.321
DiffNeRV [55]	0.320	**0.278**	0.423	0.394	**0.053**	0.431	0.478	0.268	**0.297**	0.326
PNeRV (ours)	**0.308**	0.284	**0.363**	**0.387**	0.054	**0.343**	**0.314**	**0.188**	0.324	**0.285**

Table 3. LPIPS comparison on UVG (left) and DAVIS (right) dataset.

FVD↓ *Gap*↑	Beauty		Bospho		HoneyB		Jockey		ReadyS		ShakeN		YachtR		Avg. ↑
NeRV [7]	3.76e-5	281%	1.00e-4	253%	1.45e-5	193%	5.81e-4	499%	1.98e-3	122%	3.27e-5	178%	4.07e-4	92.8%	231%
ENeRV [24]	2.66e-5	169%	7.86e-5	176%	5.88e-6	186%	1.00e-3	936%	1.46e-3	64.2%	2.12e-5	80.7%	1.00e-3	376%	284%
HNeRV [8]	3.29e-5	233%	6.74e-5	137%	1.50e-5	203%	9.46e-4	874%	2.07e-3	132%	5.06e-5	331%	3.56e-4	68.8%	282%
DiffNeRV [55]	1.29e-5	30.7%	4.28e-5	50.3%	6.50e-6	31.1%	1.55e-4	60.1%	**6.58e-4**	-26.3%	4.69e-5	300%	2.23e-4	5.9%	64.5%
PNeRV (ours)	**9.88e-6**	-	**2.85e-5**	-	**4.96e-6**	-	**9.70e-5**	-	8.94e-4	-	**1.17e-5**	-	**2.11e-4**	-	-

Table 4. FVD comparison on UVG.

FVD↓ *Gap*↑	Bmx-B		Camel		Dance-Jump		Dog		Drift-C		Elephant		Parkour		Scoo-gray		Soapbox		Avg. ↑
NeRV [7]	8.99e-5	146%	2.70e-4	404%	6.66e-5	1273%	3.02e-5	336%	3.85e-6	2830%	2.470e-5	95.8%	1.35e-4	309%	3.815e-5	197%	9.39e-5	115%	634%
ENeRV [24]	1.20e-4	229%	1.08e-4	102%	6.05e-6	24.8%	4.04e-6	-41.5%	5.41e-7	311%	2.647e-5	110%	7.09e-5	114%	3.961e-5	208%	7.01e-5	61.1%	124%
HNeRV [8]	4.97e-5	36.2%	1.04e-4	94.1%	9.58e-6	97.5%	4.51e-6	-34.6%	1.21e-6	821%	4.439e-5	252%	7.81e-5	135%	2.256e-5	75.8%	7.36e-5	69.3%	171%
DiffNeRV [55]	**3.11e-5**	-14.8%	**3.85e-5**	-28.1%	1.19e-5	146%	**3.61e-6**	-47.6%	6.48e-7	392%	6.408e-5	408%	1.45e-4	339%	1.614e-5	25.7%	**1.64e-5**	-62.2%	128%
PNeRV (ours)	3.65e-5	-	5.36e-5	-	**4.85e-6**	-	6.91e-6	-	**1.31e-7**	-	**1.261e-5**	-	**3.31e-5**	-	**1.283e-5**	-	4.35e-5	-	-

Table 5. FVD comparison on DAVIS.

blurring, validate Remark. 3. Hierarchy structure reduces ambiguity and artifacts caused by semantic uncertainty.

5.5. Additional Experiment Results

Additional results are provided in the appendix. Video interpolation on UVG is discussed in Sec. C.1 where PNeRV achieves the second-best PSNR (31.18 dB), exceeding the vanilla NeRV (26.54 dB). Video compression is shown in Sec. C.2, where competitive results are achieved over different coding pipelines. Video inpainting on the DAVIS subset is provided in Sec. C.3, where an average PSNR of 25.54 dB is achieved, outperforming NeRV (22.71 dB) and DNeRV (25.20 dB). More visual examples are shown in Sec. C.4, and visualization of feature maps in Sec. D.1. More detailed ablations are presented in Sec. D.2. More video examples with the link are listed in Sec. C.6.

6. Conclusion

To resolve the spatiotemporal inconsistency issue, we propose Pyramidal NeRV realizing multi-level information interaction by a low-cost KFC and a fusion module BSM. Further, we use UAT to provide some explanations and insights for NeRV. Competitive results on various tasks and metrics validate the superiority of PNeRV.

Limitation and future work. Hierarchical structure brings higher computational complexity. We will optimize redun-

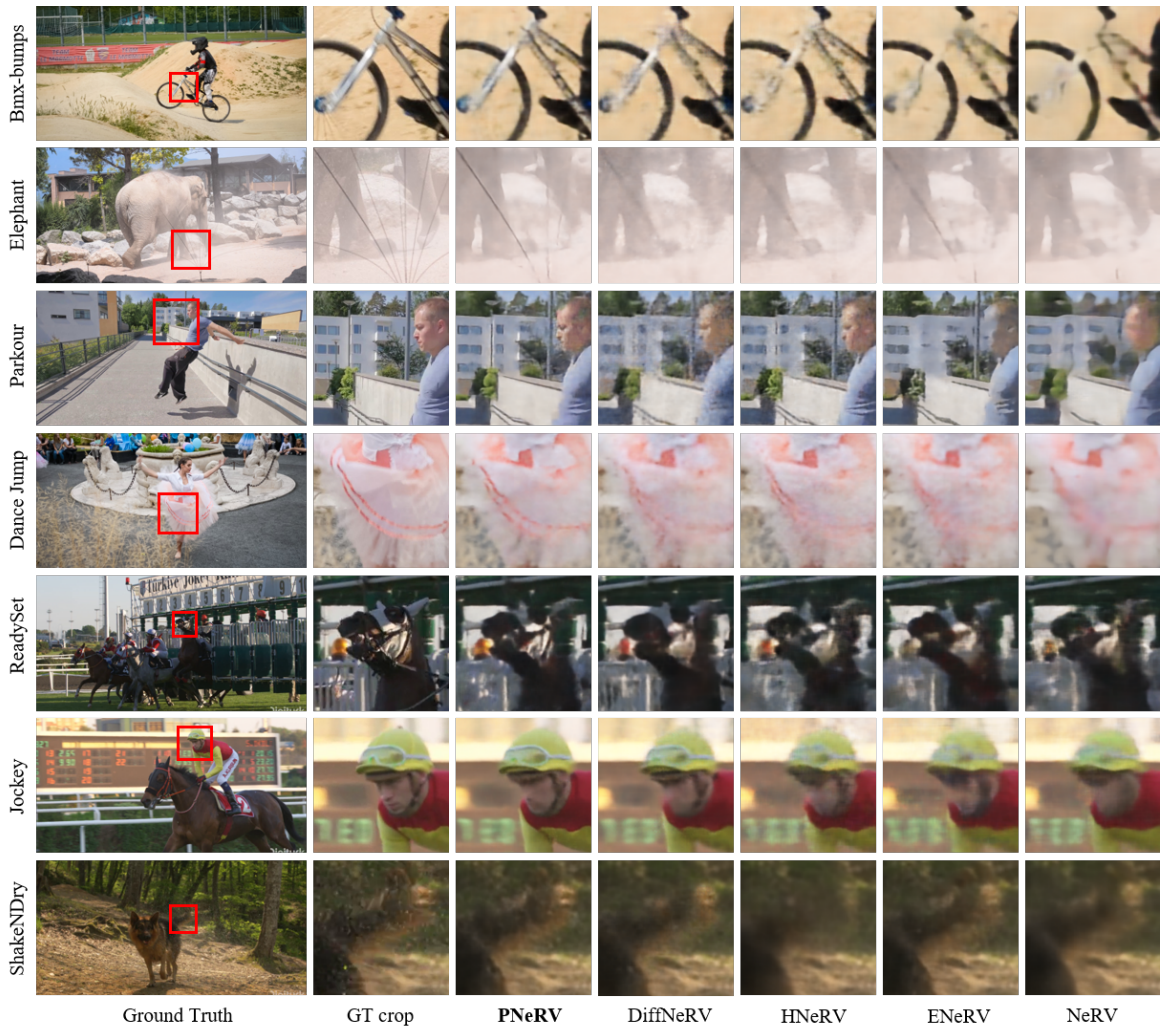

Figure 4. Visual comparison on various videos. "Bmx" has larger motion, "Elephant" has massive droplets blurring, "Parkour" involves both camera rotation and extreme dynamics, "Dance" contains large motion under high-frequency reed leaves. "Jockey", "ReadyS", and "ShakeN" are videos with complex spatiotemporal correlation in UVG. Zoom in for a detailed comparison.

	Parkour (*Dynamic*)				HoneyB (*Smooth*)			
Models Size	1.5M	3M	5M	Avg.	0.75M	1.5M	3M	Avg.
Serial (HNeRV [8])	25.07	26.56	24.34	25.32	36.65	36.72	38.96	37.44
Pyram. + Concat.	24.20	25.45	25.83	25.16	40.07	41.58	42.34	41.33
Pyram. + BSM. (**PNeRV-M**)	24.81	26.02	27.13	25.99	40.34	41.36	42.59	41.43
Serial + Diff. (DiffNeRV [55])	25.49	25.75	25.71	25.65	**40.52**	41.52	41.92	41.32
Pyram. + Diff. + BSM. (**PNeRV-L**)	**25.62**	**27.08**	**27.21**	**26.67**	39.81	**41.85**	**42.73**	**41.46**

Table 6. Ablation studies for model size and overall architecture on "HoneyB" and "Parkour".

PSNR↑ SSIM↑ (A.P.G.)↑	Concat	GRU	LSTM	**BSM**
Bilinear	27.16/0.816(-4.14)	28.39/0.847(-2.91)	28.07/0.834(-3.23)	29.08/0.862(-2.22)
Deconv	27.37/0.803(-3.93)	29.00/0.845(-2.30)	28.91/0.850(-2.39)	29.96/0.881(-1.34)
KFc	28.68/0.848(-2.62)	29.31/0.868(-1.99)	29.04/0.866(-2.26)	**31.30/0.904**(+0)

Table 7. Contribution ablations for KFc and BSM, reported as average results on 7 DAVIS videos. A.P.G. indicates the average PNSR gap compared with the final version of PNeRV (KFc + BSM); the larger the better. Detailed results for each video are given in Sec. D.2.

dant modules of the model for acceleration in the future.

Acknowledgements. The work was supported in part by National Key Research and Development Project of China (2022YFF0902402) and U.S. National Science Foundation award (CCF-2046293).

References

[1] Armen Aghajanyan, Sonal Gupta, and Luke Zettlemoyer. Intrinsic dimensionality explains the effectiveness of language model fine-tuning. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*. 3

[2] Nasir Ahmed, T. Raj Natarajan, and K. R. Rao. Discrete cosine transform. *IEEE Transactions on Computers*, 2019. 2

[3] Yochai Blau and Tomer Michaeli. The perception-distortion tradeoff. In *Computer Vision and Pattern Recognition*, 2017. 2

[4] Yochai Blau and Tomer Michaeli. Rethinking lossy compression: The rate-distortion-perception tradeoff. *Proceedings of the 36th International Conference on Machine Learning*, 2019. 2

[5] Peter J. Burt and Edward H. Adelson. The laplacian pyramid as a compact image code. *IEEE Trans. Commun.*, 1983. 2

[6] Eric Chan and Connor Z. Lin et al. Efficient geometry-aware 3d generative adversarial networks. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 1

[7] Hao Chen, Bo He, Hanyu Wang, Yixuan Ren, Ser-Nam Lim, and Abhinav Shrivastava. Nerv: Neural representations for videos. In *NeurIPS*, 2021. 1, 2, 4, 5, 6, 7

[8] Hao Chen, M. Gwilliam, Ser Nam Lim, and Abhinav Shrivastava. Hnerv: A hybrid neural representation for videos. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2, 4, 5, 6, 7, 8

[9] Zhiqin Chen, Thomas A. Funkhouser, Peter Hedman, and Andrea Tagliasacchi. Mobilenerf: Exploiting the polygon rasterization pipeline for efficient neural field rendering on mobile architectures. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, 2023. 2

[10] Junyoung Chung, Caglar Gülcehre, Kyunghyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *ArXiv*, 2014. 3, 6

[11] Emilien Dupont, Adam Goli'nski, Milad Alizadeh, Yee Whye Teh, and A. Doucet. Coin: Compression with implicit neural representations. *ArXiv*, 2021. 2

[12] Fenglei Fan, Dayang Wang, and Ge Wang. Universal approximation by a slim network with sparse shortcut connections. *ArXiv*, abs/1811.09003, 2018. 5

[13] Sara Fridovich-Keil, Alex Yu, Matthew Tancik, Qinhong Chen, Benjamin Recht, and Angjoo Kanazawa. Plenoxels: Radiance fields without neural networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, 2022. 2

[14] Didier J. Le Gall. Mpeg: a video compression standard for multimedia applications. *Commun. ACM*, 1991. 2

[15] Boris Hanin. Universal function approximation by deep neural nets with bounded width and relu activations. *ArXiv*, abs/1708.02691, 2017. 4

[16] Boris Hanin and Mark Sellke. Approximating continuous functions by relu nets of minimal width. *ArXiv*, abs/1710.11278, 2017. 2, 4, 5

[17] Bo He and Xitong Yang et al. Towards scalable neural representation for diverse videos. *ArXiv*, abs/2303.14124, 2023. 4, 6, 7

[18] Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. Spatial pyramid pooling in deep convolutional networks for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2014. 2

[19] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 1997. 3, 6

[20] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. 3

[21] Zhihao Hu, Guo Lu, and Dong Xu. Fvc: A new framework towards deep video compression in feature space. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2

[22] Jiaxin Li, Zijian Feng, Qi She, Henghui Ding, Changhu Wang, and Gim Hee Lee. MINE: towards continuous depth MPI with nerf for novel view synthesis. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*, 2021. 2

[23] Jiahao Li, Bin Li, and Yan Lu. Deep contextual video compression. *NeurIPS*, 2021. 2

[24] Zizhang Li, Mengmeng Wang, Huaijin Pi, Kechun Xu, Jianbiao Mei, and Yong Liu. E-nerv: Expedite neural video representation with disentangled spatial-temporal context. *arXiv:2207.08132*, 2022. 2, 4, 5, 6, 7

[25] Hongzhou Lin and Stefanie Jegelka. Resnet with one-neuron hidden layers is a universal approximator. *ArXiv*, abs/1806.10909, 2018. 3, 5

[26] Jianping Lin, Dong Liu, Houqiang Li, and Feng Wu. M-lvc: Multiple frames prediction for learned video compression. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[27] Tsung-Yi Lin, Piotr Dollár, Ross B. Girshick, Kaiming He, Bharath Hariharan, and Serge J. Belongie. Feature pyramid networks for object detection. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 2

[28] Shu Liu, Lu Qi, Haifang Qin, Jianping Shi, and Jiaya Jia. Path aggregation network for instance segmentation. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018. 2

[29] Songtao Liu, Di Huang, and Yunhong Wang. Learning spatial fusion for single-shot object detection. *ArXiv*, abs/1911.09516, 2019. 2

[30] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022*. 2

[31] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. *2021*

IEEE/CVF International Conference on Computer Vision (ICCV), 2021. 2, 3

[32] Guo Lu, Wanli Ouyang, Dong Xu, Xiaoyun Zhang, Chunlei Cai, and Zhiyong Gao. Dvc: An end-to-end deep video compression framework. 2019. 2

[33] 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, MMSys 2020*. 6

[34] Ben Mildenhall, Pratul P. Srinivasan, Rodrigo Ortiz Cayon, Nima Khademi Kalantari, Ravi Ramamoorthi, Ren Ng, and Abhishek Kar. Local light field fusion: practical view synthesis with prescriptive sampling guidelines. *ACM Trans. Graph.*, 2019. 2

[35] 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 *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. 1, 2

[36] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Trans. Graph.*, 2022. 2

[37] Sejun Park, Chulhee Yun, Jaeho Lee, and Jinwoo Shin. Minimum width for universal approximation. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. 2, 4

[38] Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus H. Gross, and Alexander Sorkine-Hornung. A benchmark dataset and evaluation methodology for video object segmentation. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016*. 6

[39] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. *ArXiv*, abs/1505.04597, 2015. 2

[40] Katja Schwarz, Yiyi Liao, Michael Niemeyer, and Andreas Geiger. Graf: Generative radiance fields for 3d-aware image synthesis. *ArXiv*, abs/2007.02442, 2020. 1

[41] Wenzhe Shi and Jose Caballero et al. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2, 3

[42] 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, 5

[43] Gary J. Sullivan and Thomas Wiegand. Rate-distortion optimization for video compression. *IEEE Signal Process. Mag.*, 1998. 2

[44] Gary J. Sullivan, Jens-Rainer Ohm, Woojin Han, and Thomas Wiegand. Overview of the high efficiency video coding (HEVC) standard. *IEEE Trans. Circuits Syst. Video Technol.*, 2012. 2

[45] Towaki Takikawa and Joey Litalien et al. Neural geometric level of detail: Real-time rendering with implicit 3d shapes. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 1

[46] Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphaël Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *ArXiv*, abs/1812.01717, 2018. 6

[47] Ashish Vaswani, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, 2017. 3

[48] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 548–558, 2021. 2

[49] X. Wang, Ross Girshick, Abhinav Kumar Gupta, and Kaiming He. Non-local neural networks. *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018. 3

[50] Thomas Wiegand, Gary J. Sullivan, Gisle Bjøntegaard, and Ajay Luthra. Overview of the H.264/AVC video coding standard. *IEEE Trans. Circuits Syst. Video Technol.*, 2003. 2

[51] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelnerf: Neural radiance fields from one or few images. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 1

[52] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. *CoRR*, abs/1511.07122, 2015. 3

[53] Matthew D. Zeiler, Dilip Krishnan, Graham W. Taylor, and Rob Fergus. Deconvolutional networks. *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2010. 2, 3, 6

[54] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6

[55] Qi Zhao, M. Salman Asif, and Zhan Ma. Dnerv: Modeling inherent dynamics via difference neural representation for videos. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2, 4, 5, 6, 7, 8

[56] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: A nested u-net architecture for medical image segmentation. *MICCAI 2018*. 2