Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

FANeRV: frequency separation and augmentation based neural representation for video

Li Yu [a,b,*], Zhihui Li [a], Chao Yao [c], Jimin Xiao [d], Moncef Gabbouj [e]

[a] School of Computer Science, Nanjing University of Information Science and Technology, 210044, Nanjing, China
[b] Jiangsu Collaborative Innovation Center of Atmospheric Environment and Equipment Technology, Nanjing University of Information Science and Technology, 210044, Nanjing, China
[c] School of Computer and Communication Engineering, University of Science and Technology Beijing, 100083, Beijing, China
[d] Department of Intelligent Science, School of Advanced Technology, Xi'an Jiaotong-Liverpool University, 215000, Suzhou, China
[e] Faculty of Information Technology and Communication Sciences, Tampere University, 33101, Tampere, Finland

ARTICLE INFO

Keywords:
Video compression
Implicit neural representation
Neural representation for video
Wavelet transform

ABSTRACT

Neural Representations for Video (NeRV) have emerged as a powerful paradigm for video representation, gaining considerable attention for their strong performance across various video tasks like video compression. However, NeRV models are fundamentally limited by the inherent spectral bias of neural networks, struggling to learn high-frequency details, which leads to pervasively blurred reconstructions and loss of texture. While existing works attempt to address this, they often either fail to treat high- and low-frequency components differently according to their distinct statistical properties, or rely on upsampling operators for multi-scale fusion that are prone to introducing visual artifacts such as blurring and checkerboarding, thus hindering further improvements in reconstruction quality. To precisely tackle these challenges, this paper proposes a Frequency Separation and Augmentation based Neural Representation for video (FANeRV). Our method utilizes the Discrete Wavelet Transform (DWT) to explicitly decompose features into high- and low-frequency components and, for the first time, introduces an asymmetric architecture for specialized processing. Furthermore, we discard traditional upsampling operators and instead leverage the DWT's inherent multi-resolution properties to design a novel multi-scale fusion mechanism. This mechanism directly fuses features from lower-resolution stages with corresponding frequency subbands at higher resolutions, optimizing information propagation. Finally, a residual enhancement block is integrated into the network's later stages to further bolster the restoration of high-frequency details. Extensive experiments demonstrate that FANeRV achieves superior reconstruction quality and outperforms existing NeRV methods across multiple tasks, including video compression, interpolation, and inpainting.

1. Introduction

Implicit Neural Representations (INRs) have emerged as a powerful paradigm for efficiently representing and processing complex multimedia signals, such as images, video, and 3D scenes (Dupont et al., 2021; Mildenhall et al., 2021; Sitzmann et al., 2020, 2019). Unlike traditional explicit representations, such as pixel grids for images or vertex meshes for 3D models, INRs utilize a neural network to parameterize the signal as a continuous function that maps input coordinates (e.g., spatial coordinates) to corresponding signal attributes (e.g., RGB color). This approach yields inherently compact, continuous, and potentially resolution-agnostic representations, offering notable advantages for diverse applications including data inpainting (Li et al., 2023; Saragadam et al., 2023), signal compression (Strümpler et al., 2022; Zhang et al., 2021), and generative modeling (Skorokhodov et al., 2021, 2022).

Neural Representations for Videos (NeRV) (Chen et al., 2021) represents a significant application of INRs in the video processing domain. NeRV employs Convolutional Neural Networks (CNNs) to represent a video as a continuous function mapping frame indices to the corresponding video frames. Using a single network with shared parameters for the entire sequence, the model learns to represent both the invariant content and the varying details across video frames. This process naturally captures spatio-temporal correlations, thereby exploiting redundancy for efficient compression. In this paradigm, the video encoding process is converted into optimizing the network parameters to fit the video data

[*] Corresponding author.
E-mail addresses: li.yu@nuist.edu.cn (L. Yu), 202212490325@nuist.edu.cn (Z. Li), yaochao@ustb.edu.cn (C. Yao), jimin.xiao@xjtlu.edu.cn (J. Xiao), moncef.gabbouj@tuni.fi (M. Gabbouj).

https://doi.org/10.1016/j.eswa.2025.129935
Received 15 May 2025; Received in revised form 10 September 2025; Accepted 2 October 2025
Available online 6 October 2025

while decoding process becomes a forward inference pass through the trained network. Compared to traditional video compression methods reliant on complex procedures like motion estimation, motion compensation, and transform coding, the NeRV approach significantly simplifies the codec pipeline and achieves high decoding speeds, emerging as a promising alternative for video compression. Consequently, subsequent research (Chen et al., 2023; Li et al., 2022; Zhang et al., 2024) has advanced NeRV-based models through enhanced network architectures and optimized training strategies, yielding notable improvements in reconstruction fidelity and compression efficiency, achieving competitive rate-distortion performance against established coding standards like H.264/AVC (Wiegand et al., 2003) and H.265/HEVC (Sullivan et al., 2012). Furthermore, the inherent continuous nature of this representation facilitates various downstream tasks (Chen et al., 2022; Jung et al., 2023; Lu et al., 2023; Rho et al., 2022),including video frame interpolation, video inpainting, and video super-resolution, showcasing its broad application prospects.

Although the universal approximation theorem for neural networks suggests that INRs can achieve high-fidelity video reconstruction with sufficient network capacity and training duration, practical implementations remain constrained by the inherent spectral bias of neural networks (Rahaman et al., 2019; Xu et al., 2019). This indicates that neural networks preferentially learn low-frequency information during training and exhibit a weaker capability in fitting high-frequency details. Consequently, INR-based video reconstructions often suffer from a loss of image texture and sharpness, which directly impacts the overall visual quality.

To address the challenge of spectral bias, several approaches have been proposed from a frequency-domain perspective. Some works attempt to facilitate high-frequency recovery by directly incorporating priors into the network (Hayami et al., 2025; Wu et al., 2024a; Xu et al., 2024; Yu et al., 2025), though this often increases model size and computational cost. More recent methods leverage the Discrete Wavelet Transform (DWT) for explicit frequency decomposition. For instance, SNeRV (Kim et al., 2025) decomposes the input video for staged recovery, while HFS-HNeRV (Zhao et al., 2024) uses wavelets to build frequency-aware attention mechanisms. While these methods validate the potential of frequency-domain processing, they often treat the decomposed components uniformly. A key limitation of this approach is its uniform treatment of low-frequency (global structure) and high-frequency (local details) components, which possess distinct statistical properties and thus demand specialized, asymmetric processing. Hu et al. (2024), Lu et al. (2022), Yang et al. (2024). Other works aim to improve detail preservation by fusing features across different resolutions during decoding, as seen in models like NeRV++ (Chen et al., 2025; Ghorbel et al., 2024). However, the effectiveness of these architectures is often constrained by the upsampling operators they rely on. Common operators, such as transposed convolution and bilinear interpolation, are prone to introducing visual artifacts like checkerboarding or blurring. These artifacts can degrade the very high-frequency details the fusion is intended to preserve, thereby hindering the overall reconstruction quality.

Building on these insights, this paper proposes a Frequency Separation and Augmentation based Neural Representation for Video (FANeRV). FANeRV utilizes the Discrete Wavelet Transform (DWT) to explicitly decompose intermediate features within the reconstruction process into high-frequency and low-frequency components. Subsequently, lightweight enhancement modules tailored to the characteristics of each component are applied. These enhanced components are then fused and optimized using the Inverse Discrete Wavelet Transform (IDWT) and a gating feed-froward network. Crucially, the multi-resolution property inherent to the DWT provides a parameter-free, and frequency-specific mechanism for feature alignment across different stages, obviating the need for artifact-prone upsampling operations like transposed convolution. We leverage this by fusing low-resolution features from a preceding decoding stage directly with the low-frequency subband to enhance information propagation and guide the network towards high-frequency details. Finally, to further bolster high-frequency detail recovery and refine the learned representations, a residual enhancement block is integrated into the network's deeper stages. Experimental results on the Bunny, UVG, and DAVIS datasets demonstrate that FANeRV significantly improves video reconstruction quality and achieves superior performance across multiple tasks, including video compression, video inpainting, and video interpolation, outperforming existing NeRV methods while maintaining the comparable model capacity. The main contributions of this paper are summarized as follows:

- We propose FANeRV, a novel video representation framework that mitigates spectral bias by employing the Discrete Wavelet Transform (DWT) for explicit frequency separation and applying frequency-specific enhancement strategies.
- We propose a Frequency-Specific Multi-Resolution Fusion, which leverages the inherent multi-resolution properties of the DWT to facilitate multi-resolution feature fusion. This parameter-free approach avoids the blurring and checkerboard artifacts.
- We conduct comprehensive experiments, demonstrating that FANeRV achieves significantly improved reconstruction performance and achieves superior results across multiple tasks.

2. Related work

2.1. Neural representations for video

Implicit Neural Representations (INRs) have emerged as a powerful paradigm for modeling diverse signals via a function F, which maps input coordinates $\theta \in \mathbb{R}^n$ to corresponding signal values $y = F(\theta)$, where $y \in \mathbb{R}^m$. Owing to their ability to represent signals accurately using neural networks with parameter counts significantly smaller than the raw data, INRs have been increasingly applied to video representation.

Early video INRs (Zhang et al., 2021) often directly adapted methods from image INRs (Dupont et al., 2021), employing Multilayer Perceptrons (MLPs) to learn a mapping function from individual pixel coordinates to corresponding pixel values. However, this pixel-wise representation strategy incurs substantial computational costs, as the neural network must be evaluated independently for every pixel coordinate, leading to slow decoding speeds and suboptimal reconstruction quality. To address this limitation, NeRV (Chen et al., 2021) introduces a frame-wise approach that utilizes Convolutional Neural Networks (CNNs) to directly learn a mapping from a frame index t to the entire corresponding RGB image, significantly enhancing both coding efficiency and reconstruction fidelity. Building upon this foundation, E-nerv (Li et al., 2022) proposed decomposing the neural representation input into separate spatial and temporal contexts, thereby reducing model parameters while maintaining representational power. Addressing the lack of visual priors in earlier methods, HNeRV (Chen et al., 2023) introduced a hybrid representation scheme. This approach replaces the simple time coordinate t input with compact, content-related feature embeddings, markedly improving reconstruction quality and convergence speed. Subsequently, Zhang et al. (2024) tackled the issue of inadequate temporal alignment of intermediate features during decoding in NeRV-based models by introducing a temporal modulation mechanism. They further enhanced both reconstruction quality and compression performance by employing improved entropy minimization techniques for model compression.

Despite these advancements, overcoming the inherent spectral bias of NeRV-based models remains a critical bottleneck. Research to address this issue has largely followed three directions. The first involves injecting high-frequency priors, for instance, through vector quantization (Wu et al., 2024a; Xu et al., 2024; Yu et al., 2025), though this often increases model complexity. A second, more common strategy is multi–resolution feature fusion, as seen in models like LNeRV (Chen et al., 2025) and NeRV++ (Ghorbel et al., 2024). These methods typically employ residual learning principles and use upsampling operators like transposed convolution to combine features from different decoder stages. More recently, a third direction has emerged that leverages the Discrete Wavelet Transform (DWT). For example, SNeRV (Kim et al., 2025) applies DWT to decompose the input video into different frequency sub-bands, while HFS-HNeRV (Zhao et al., 2024) uses wavelets to construct frequency-aware attention mechanisms. While these works validate the promise of frequency-domain processing, our approach is fundamentally different. FANeRV applies frequency separation not to the input, but to the intermediate features within the decoder for frequency-specific enhancement. Critically, we also leverage the DWT's inherent multi-resolution properties as a principled, parameter-free mechanism for multi-resolution feature fusion, which completely replaces the problematic upsampling operators used in prior art. This integrated design allows FANeRV to simultaneously address two key challenges: spectral bias is mitigated through targeted frequency enhancement, while upsampling artifacts are eliminated by the alias-free nature of the DWT-based fusion.

2.2. Frequency domain learning

Frequency domain analysis is a well-established technique in traditional low-level computer vision. Following the advancements in deep learning, it has increasingly been leveraged to enhance performance across various computer vision applications. Among frequency analysis methods, the Discrete Wavelet Transform (DWT) is particularly notable for its inherent multi-resolution analysis capabilities. The DWT decomposes a signal, such as an image, into multiple subbands at different scales, where each subband represents information within a specific frequency range. This multi-resolution decomposition allows the DWT to effectively represent both fine-grained local details (e.g., edges, textures) and coarse-grained global structures within the signal simultaneously.

Researchers have successfully integrated this technique with Convolutional Neural Networks (CNNs), yielding significant advancements in diverse tasks (Liu et al., 2018; Wu et al., 2024b; Xu et al., 2023). For instance, Liu et al. (2018) proposed a novel multi-level wavelet convolutional neural network designed to enlarge the receptive field, thereby achieving a favorable trade-off between efficiency and reconstruction performance. Xu et al. (2023) designed a downsampling module based on the Haar wavelet transform to reduce image resolution while alleviating the loss of information crucial for semantic segmentation tasks. Wu et al. (2024b) utilized DWT to decompose features into low-frequency and high-frequency components, subsequently computing attention based on the low-frequency components, which significantly mitigated the detrimental effects of noise on model performance. Drawing inspiration from these successes, we identify the specific properties that make DWT uniquely suited for our method. Unlike the spatially-agnostic Fast Fourier Transform (FFT), it preserves crucial spatial locality for convolutional processing. In contrast to learnable operators like transposed convolution, it is a parameter-free transformation that fundamentally avoids training artifacts such as checkerboarding. We leverage these properties to tackle the dual challenges in our work: using its frequency decomposition for targeted enhancement against spectral bias, and its multi-resolution nature for an artifact-free feature fusion to solve the feature alignment problem.

3. Proposed method

3.1. Overview

To directly combat the inherent spectral bias of neural networks, which hinders the reconstruction of high-frequency details, our core design philosophy is to divide and conquer in the frequency domain. Our approach is built on three key strategies: explicitly separating features into distinct frequency components, applying specialized enhancements to each, and leveraging a novel, artifact-free mechanism for multi-resolution fusion. This philosophy is embodied in our proposed Frequency Separation and Augmentation based Neural Representation for Video (FANeRV) architecture, whose workflow we detail next. As shown in Fig. 1, for a given video frame I_t, a ConvNeXt-based encoder (Liu et al., 2022) first produces a compact, content-related embedding that encapsulates its essential spatial information. Concurrently, the frame index t is converted into a high-dimensional temporal embedding t_{emb} via a positional encoding function and a small MLP. These two embeddings serve as the primary inputs to decoder.The decoder is a five-stage network designed to progressively synthesize the final video frame. The intermediate stages (2-4) pair a standard NeRV block with the proposed Wavelet Frequency Upgrade (WFU) block. The initial stage omits the WFU block, as its low feature resolution is incompatible with the dyadic decomposition required by the Discrete Wavelet Transform, while the final stage is augmented with a Residual Enhancement Block (REB) to enhance the final reconstruction. The WFU block utilizes the DWT for frequency separation, subsequently enhancing the resulting components via specialized branches and integrating information from the previous stage. The REB is to further boost the model's capacity for restoring challenging fine-grained details, ensuring a faithful and sharp final output $\hat{I}_t$. Throughout this process, the

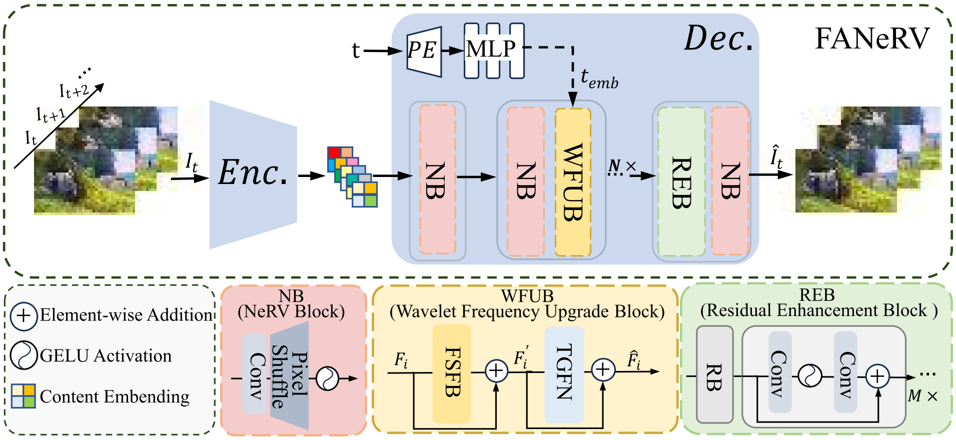

Fig. 1. Architecture of the proposed FANeRV. The model consists of an encoder-decoder structure. The encoder maps an input frame I_t to a content embedding, while the frame index t is processed into a temporal embedding t_{emb}. The decoder then synthesizes the output frame $\hat{I}_t$ using these embeddings. The bottom panel details the decoder's three key building blocks: the NeRV Block (NB) for spatial upsampling, our Wavelet Frequency Upgrade Block (WFUB) for frequency-specific enhancement under temporal modulation, and the Residual Enhancement Block (REB) for final feature refinement.

temporal embedding t_{emb} provides frame-specific conditioning at each stage. It does this by generating parameters that dynamically modulate the features, ensuring the decoder synthesizes the correct frame corresponding to index t. After training is complete, the video is represented by the combination of explicit embeddings and implicit network weights, enabling high-speed frame reconstruction through a simple forward pass.

3.2. Wavelet frequency upgrade block

A key limitation of existing methods is their uniform treatment of all frequency components, which is suboptimal for recovering both global structures and fine details simultaneously. To address this, we propose an asymmetric processing strategy where low- and high-frequency components are handled by specialized modules. The rationale for this approach stems from their significant differences in information content and processing requirements. Low-frequency components primarily carry global structural information, such as shape contours, background, and overall object forms. These features often exhibit long-range dependencies across regions, making the exploration of non-local feature interactions crucial for their effective recovery. Conversely, high-frequency components mainly contain local details like textures, edges, and fine structural information. These details often vary significantly across different local regions; therefore, the precise recovery of high-frequency information demands fine-grained modeling capable of capturing these localized characteristics. To implement this strategy, this paper proposes a Wavelet Frequency Upgrade (WFU) module. This module utilizes the Discrete Wavelet Transform (DWT) to explicitly separate high-frequency and low-frequency components of the image features and applies distinct enhancement processes to each. Furthermore, the module leverages the DWT's inherent multi-resolution property for our Frequency-Specific Multi-Resolution Fusion, guiding the network to focus on the recovery of high-frequency details while providing an artifact-free fusion path. As illustrated in Fig. 1, The WFU module comprises a Frequency Separation Feature Boosting (FSFB) module and a Time-Modulated Gated Feed-Forward Network (TGFN). The FSFB is responsible for feature enhancement on the separated frequency components, while the TGFN module dynamically fuses the enhanced components under temporal modulation.

Specifically, given an input feature map $F_i \in \mathbb{R}^{H\times W\times C}$, we first employ the Haar wavelet transform (Daubechies, 1990) to decompose it. This transform is fundamentally constructed from a pair of 1D filters: a low-pass filter L for capturing approximations and a high-pass filter H for capturing details. For the Haar wavelet, they are defined as:

$$L = \frac{1}{\sqrt{2}}[1,1], \quad H = \frac{1}{\sqrt{2}}[-1,1] \tag{1}$$

By applying these filters sequentially along the feature map's spatial axes—first horizontally, then vertically—the transform decomposes the feature map from the current stage, (F_i), into the following four subbands:

$$A^i_{LL}, H^i_{LH}, V^i_{HL}, D^i_{HH} = \mathrm{DWT}(F_i). \tag{2}$$

where A^i_{LL} represents the low-frequency subband containing global structural information , while H^i_{LH}, V^i_{HL}, and D^i_{HH} represent the high-frequency subbands capturing horizontal, vertical, and diagonal texture details, respectively. Each subband has dimensions $\mathbb{R}^{(H/2)\times(W/2)\times C}$.

Since the spatial upsampling factor at each decoding stage is 2, the subbands generated by the DWT at the current stage (i) naturally possess the same spatial dimensions as the low-resolution feature map $\hat{F}_{i-1}$ from the previous stage $(i-1)$. Due to low-resolution features predominantly contain low-frequency information, we directly fuse $\hat{F}_{i-1}$ with the current low-frequency subband A^i_{LL}. This provides a strong, relevant prior for the current stage, effectively guiding the network to leverage existing low-frequency information and dedicate more resources to learning the challenging high-frequency details. After obtaining the high-frequency components and the enhanced low-frequency component, 1×1 convolutions are applied to transform them into preliminary high-frequency features F^h_i and low-frequency features F^l_i. Subsequently, the low-frequency and high-frequency branches within the FSFB module operate distinctly to capture global and local characteristics, respectively, facilitating the precise recovery of both low- and high-frequency information.This process can be formulated as:

$$\begin{bmatrix}\hat{F}^l_i,\\ \hat{F}^h_i\end{bmatrix} = \mathrm{FSFB}\left(\begin{matrix}\mathrm{Conv}_{1\times1}(A^i_{LL} \otimes \hat{F}_{i-1}),\\ \mathrm{Conv}_{1\times1}(H^i_{LH}, V^i_{HL}, D^i_{HH})\end{matrix}\right). \tag{3}$$

where $\hat{F}^l_i \in \mathbb{R}^{(H/2)\times(W/2)\times C}$ and $\hat{F}^h_i \in \mathbb{R}^{(H/2)\times(W/2)\times 3C}$ are the refined low-frequency and high-frequency features, respectively, $\otimes$ denotes the concatenation operation along the channel dimension. Finally, the Inverse Discrete Wavelet Transform (IDWT) is applied to $\hat{F}^l_i$ and $\hat{F}^h_i$ to reconstruct a preliminary fused feature F'_i, incorporating a residual connection with the original input F_i:

$$F'_i = \mathrm{IDWT}(\hat{F}^l_i, \hat{F}^h_i) + F_i \tag{4}$$

To further refine the fused features and integrate temporal information at each stage, we first generate a temporal embedding from the frame index. Given the integer frame index t, it is first converted into a positional encoding (PE) and then passed through a small multi-layer perceptron (MLP) to produce the high-dimensional temporal embedding t_{emb}:

$$t_{\mathrm{emb}} = \mathrm{MLP}(\mathrm{PE}(t)). \tag{5}$$

This temporal embedding t_{emb} is then used by the Time-Modulated Gated Feed-Forward Network (TGFN) to modulate the intermediate feature F'_i, and generate the final output feature of the WFU block $\hat{F}_i$.

$$\hat{F}_i = \mathrm{TGFN}(F'_i | t_{\mathrm{emb}}) + F'_i. \tag{6}$$

3.2.1. Frequency Separation Feature Boosting module

The design of the FSFB module is driven by two core principles: (1) Asymmetric Processing for different frequency components and (2) Frequency-Specific Multi-Resolution Fusion, as described in the previous section. We recognize that low-frequency (global structure) and high-frequency (local details) components have different modeling needs. Therefore, FSFB uses two specialized, parallel branches to process them asymmetrically, which is more effective than a single, uniform structure (Fig. 2).

Specifically, the low-frequency branch is designed to model the long-range dependencies characteristic of global structures. While Transformer-based self-attention mechanisms excel at this, their quadratic complexity can be a bottleneck. Inspired by recent works showing that large kernel convolutions (Ding et al., 2022; Guo et al., 2023) can achieve similar long-range modeling capabilities with higher efficiency, we design a multi-resolution deep feature modulation branch. As illustrated in Fig. 1, this branch employs parallel Atrous Separable Convolution (ASC) with varying kernel sizes and dilation rates to construct multi-scale feature maps f^j_i, capturing low-frequency characteristics at different scales. Given an input low-frequency feature $F^l_i \in \mathbb{R}^{H\times W\times C}$, this process is represented as:

$$f^j_i = \mathrm{ASC}_{k_j,d_j}(F^l_i), \quad 0 \le j \le 3. \tag{7}$$

where k_j and d_j denote the kernel size and dilation rate, respectively, specifically set as $k_j \in \{5,7,9,11\}$ and $d_j \in \{1,2,3,4\}$ for $j = 0,1,2,3$.

To refine the low-frequency features F^l_i, the branch employs spatial-adaptive modulation, dynamically weighting spatial locations based on aggregated multi-scale representations. Specifically, the multi-scale features are first concatenated along the channel dimension. Then, a 1×1 convolutional layer followed by a GELU activation function (σ) is applied to produce an attention map S. This attention map S is then applied via element-wise multiplication $(\odot)$ to the original input feature F^l_i. Finally, another 1×1 convolutional layer models inter-channel relationships, yielding the output $\hat{F}^l_i$:

$$S = \sigma(\mathrm{Conv}_{1\times1}(\mathrm{Concat}([f^0_i, f^1_i, f^2_i, f^3_i]))). \tag{8}$$

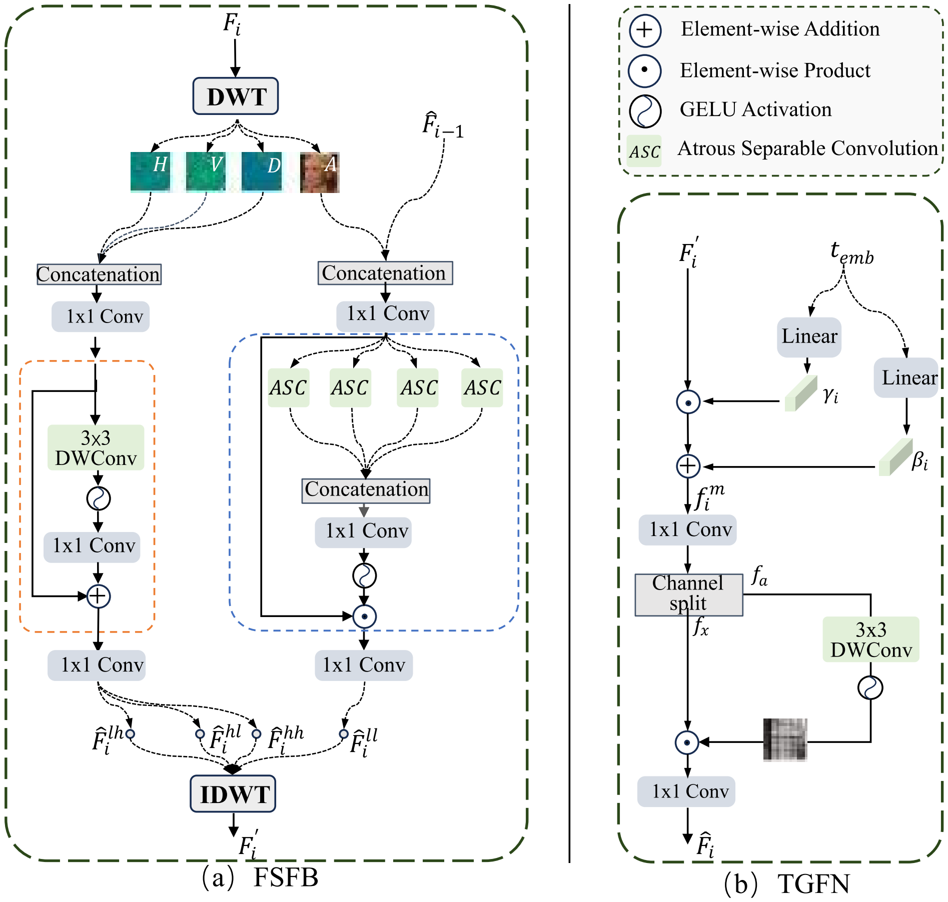

Fig. 2. Detailed architecture of the (a) Frequency Separation Feature Boosting (FSFB) module and the (b) Time-Modulated Gated Feed-Forward Network (TGFN) module. In FSFB, an input feature map F_i is decomposed by DWT into low-frequency (LF) and high-frequency (HF) sub-bands. The LF branch fuses its sub-band with features from the previous stage ($\hat{F}_{i-1}$) and processes them with parallel Atrous Separable Convolutions (ASC) in a spatial attention mechanism to capture multi-scale global context. The HF branch uses a simpler residual block to enhance local details. The enhanced sub-bands are then reconstructed via IDWT. In TGFN, the module injects temporal information and refines features. It first uses the time embedding t_{emb} to generate scale (γ_i) and shift (β_i) parameters for an affine transformation on the input feature F_i'. Subsequently, a gating mechanism with a channel split adaptively modulates the feature pathways to select and refine salient information.

$$\hat{F}_i^l = \mathrm{Conv}_{1\times1}(S \odot F_i^l). \tag{9}$$

The high-frequency Branch is designed to restore fine-grained local details. It takes the high-frequency feature F_i^h as input and employs a lightweight residual block to process it, as detailed in Eqs. (10) and (11). This branch uses a 3×3 DWConv to efficiently extract texture information, while a crucial residual connection ensures the stable propagation of high-frequency signals without degradation.

$$f_h = \mathrm{Conv}_{1\times1}(\mathrm{DWConv}_{3\times3}(F_i^h)) \tag{10}$$

$$\hat{F}_i^h = \mathrm{Conv}_{1\times1}(\sigma(f_h)) + F_i^h \tag{11}$$

3.2.2. Time-Modulated Gated Feed-Forward Network

The Time-Modulated Gated Feed-Forward Network (TGFN) is the final component of the WFU block. Its primary input is the feature map F_i', which represents the preliminary result of our frequency-specific processing, obtained by reconstructing the enhanced LF and HF components via IDWT(as defined in Eq. (4)). The TGFN is designed to perform two critical roles on this feature: (1) Dynamic Feature Refinement, using a gating mechanism to adaptively select and integrate features while suppressing redundancy, and (2) Temporal Conditioning, injecting the frame index to ensure the features are correctly aligned with the target frame. As shown in Fig. 1, the input feature F_i' first undergoes an affine transformation using modulation parameters γ_i and β_i, which are derived from the time embedding t_{emb}:

$$\gamma_i, \beta_i = \mathrm{MLP}_i(t_{\mathrm{emb}}). \tag{12}$$

$$f_i^m = F_i' \odot (1 + \gamma_i) + \beta_i. \tag{13}$$

where $\odot$ denotes element-wise multiplication, and γ_i and β_i control the scaling and shifting of the features, respectively. This transformation allows the network to learn a frame-specific adjustment for the features. By dynamically scaling and shifting the feature channels based on the frame index t, the network can precisely control the representation, ensuring that the synthesized details are appropriate for the target frame. Subsequently, a 1×1 convolution is applied to the modulated feature f_i^m for cross-channel interaction, generating an expanded hidden representation. This output is then split into two components, f_a and f_x. The component f_a passes through a 3×3 depthwise convolution to capture local spatial patterns, followed by a GELU non-linear activation function (σ) to generate gating weights. These weights are used to adaptively modulate f_x. Finally, a 1×1 convolution mixes the features and reduces the channel dimension to match the input feature dimension:

$$f_a, f_x = \mathrm{Split}(\mathrm{Conv}_{1\times1}(f_i^m)), \tag{14}$$

$$\hat{F}_i = \mathrm{Conv}_{1\times1}(f_x \odot \sigma(\mathrm{DWConv}_{3\times3}(f_a))). \tag{15}$$

3.3. Residual enhancement block

Existing video INR methods typically reduce channel dimensions during progressive spatial upsampling to maintain model compactness. However, this common strategy can limit representational capacity in

Table 1
Video Regression Performance on the UVG Dataset. In each cell, the top value is PSNR and the bottom value is MS-SSIM. Best results are in bold, second best are underlined.

Video	Resolution: 1920 × 960					Resolution: 960 × 480				
	NeRV	E-NeRV	HNeRV	Boost	Ours	NeRV	E-NeRV	HNeRV	Boost	Ours
Beauty	33.25 0.8886	33.53 0.8958	33.58 0.8941	33.93 0.9006	**34.11 0.9032**	32.38 0.9346	32.59 0.9399	32.81 0.9341	33.06 0.9437	**33.48 0.9462**
Honey.	37.26 0.9794	39.04 0.9845	38.96 0.9844	39.62 0.9854	**39.69 0.9854**	36.64 0.9912	38.47 0.9936	38.52 0.9936	38.96 0.9942	**39.17 0.9958**
Bosph.	33.22 0.9305	33.81 0.9442	34.73 0.9451	36.00 0.9652	**37.09 0.9729**	32.95 0.9577	33.72 0.9705	34.58 0.9703	36.41 0.9843	**36.64 0.9858**
Yacht.	28.03 0.8726	27.74 0.8951	29.26 0.8907	29.69 0.9079	**30.58 0.9230**	28.07 0.9183	27.86 0.9393	29.24 0.9354	30.10 0.9512	**30.40 0.9557**
Ready.	24.84 0.8310	24.09 0.8515	25.74 0.8420	28.33 0.9173	**29.68 0.9367**	24.55 0.8884	24.05 0.9069	25.73 0.9112	28.80 0.9603	**29.28 0.9693**
Jockey	31.74 0.8874	29.35 0.8805	32.04 0.8802	34.51 0.9326	**35.36 0.9438**	31.33 0.9154	28.98 0.9084	32.04 0.9151	34.29 0.9570	**34.83 0.9647**
Shake.	33.08 0.9325	34.54 0.9467	34.57 0.9450	35.89 0.9581	**35.97 0.9589**	32.74 0.9603	34.06 0.9715	34.34 0.9698	35.25 0.9768	**35.81 0.9793**
Avg.	31.63 0.9031	31.73 0.9140	32.69 0.9116	34.00 0.9382	**34.64 0.9463**	31.24 0.9380	31.39 0.9472	32.47 0.9470	33.84 0.9668	**34.23 0.9710**

later stages where the network's focus shifts primarily towards reconstructing fine, high-frequency details. To mitigate this limitation, FANeRV retains channel reduction in early stages for efficiency but strategically incorporates a Residual Enhancement Block (REB) in the final high-resolution stage. Composed of multiple stacked residual blocks, REB selectively increases network depth and enhances local feature extraction capabilities precisely. thereby further improving overall video reconstruction quality.

3.4. Loss function

To improve frame detail and structural accuracy, we use a hybrid loss function combining L1 loss and Multi-Scale Structural Similarity Index Measure (MS-SSIM). To further enhance the retention of high-frequency details, we employ frequency constraints to regularize network training, the loss function as:

$$L_{\text{spa}} = \alpha \| \hat{I}_t - I_t \|_1 + (1-\alpha)(1 - \text{MS-SSIM}(\hat{I}_t, I_t)), \tag{16}$$

$$L_{\text{fft}} = \| \mathcal{F}(\hat{I}_t) - \mathcal{F}(I_t) \|_1, \tag{17}$$

$$L_{\text{total}} = L_{\text{spa}} + \mu L_{\text{fft}}. \tag{18}$$

Here, $\hat{I}_t$ and I_t represent the reconstructed and original frames, respectively. The symbol $\| \cdot \|_1$ denotes the L_1-norm, and $\mathcal{F}$ represents the Fast Fourier Transform (FFT). Additionally, α and μ are weight parameters, empirically set to 0.7 and 70, respectively.

4. Experimental results and discussion

4.1. DataSets and implementations

4.1.1. Datasets

We evaluated our approach using the Bunny (Roosendaal, 2008), UVG (Mercat et al., 2020), and DAVIS (Perazzi et al., 2016) datasets. The Bunny consists of 132 frames at 720 × 1280 resolution. The UVG dataset consists of seven video sequences, each with 300 or 600 frames at a resolution of 1080 × 1920. For the DAVIS dataset, we used the validation set, which includes 20 videos, each with a resolution of 1080 × 1920. Following the setup of prior works, we center-cropped the videos for our experiments: inputs treated as 1280 × 720 resolution were cropped to 1280 × 640, and inputs at 1920 × 1080 resolution were cropped to 1920 × 960.

4.1.2. Implements details

The decoder spatial upsampling factor was set to [5, 2, 2, 2, 2] for the Bunny dataset and [5, 2, 2, 2, 2, 2] for UVG and DAVIS. For evaluation, we used Peak Signal-to-Noise Ratio (PSNR) and Multi-Scale Structural Similarity Index Measure (MS-SSIM) to assess distortion and measured the video compression bit rate in bits per pixel (bpp). Unless specified otherwise for a particular task, all reported quantitative metrics represent the arithmetic mean calculated over all frames of the respective video sequences. Training was performed using the Adan optimizer (Xie et al., 2022) with a cosine learning rate decay, starting from an initial learning rate of 3×10^{-3}. The batch size was set to 1. All experiments were implemented in PyTorch and executed on a single NVIDIA RTX 4090 GPU, with a model size of 3M and 300 training epochs unless otherwise specified.

4.2. Results

We conduct a comprehensive evaluation of FANeRV against four leading NeRV-based models: NeRV (Chen et al., 2021), E-NeRV (Li et al., 2022), HNeRV (Chen et al., 2023), and Boosting NeRV (Boost) (Zhang et al., 2024). To ensure a fair comparison centered on architectural advantages, all models are configured to a similar parameter count of approximately 3M. For all baseline methods, we utilize their official public implementations and train them with the optimal configurations recommended by the authors, facilitating a "best-vs-best" analysis. Furthermore, we demonstrate our model's broad effectiveness by assessing its performance across a diverse set of applications, including video regression, compression, interpolation, and inpainting. The following sections present detailed quantitative and qualitative results.

4.2.1. Video regression

Table 1 shows the video regression results on the UVG dataset. From the first list of results, it can be seen that the proposed model achieves the optimal performance in both PSNR and MS-SSIM metrics, with an average improvement of 1.88 % and 0.86 %, respectively, compared with the suboptimal method. In addition, additional experiments are conducted using the 960 × 480 downsampled version of the UVG dataset to validate the robustness of the proposed method, and it can be seen that the proposed method also achieves the optimal results, with an average improvement of 1.15 % and 0.43 % compared with the suboptimal method. The performance gains of FANeRV are most significant in texture-rich sequences such as 'Bosph.' and 'Jockey', demonstrating its enhanced capability in handling complex details. This quantitative improvement is visually corroborated in Fig. 3, where our model reconstructs markedly sharper text and facial features. A corresponding frequency spectrum analysis reveals that FANeRV robustly preserves high-frequency energy, in stark contrast to competing models. This directly validates the efficacy of our Frequency Separation Feature Boosting (FSFB) module in mitigating the spectral bias of standard NeRV architectures by explicitly enhancing high-frequency components.

Furthermore, since the training process of neural representation is essentially a function fitting process, the number of parameters of the representation network and the number of training rounds directly affect the video reconstruction effect. Therefore, we also investigate the effects of different model sizes and different iteration times on the reconstruction performance. The experimental results in Table 2 show that under different model sizes and training epochs, the proposed method shows the best performance, which proves the robustness of the method.

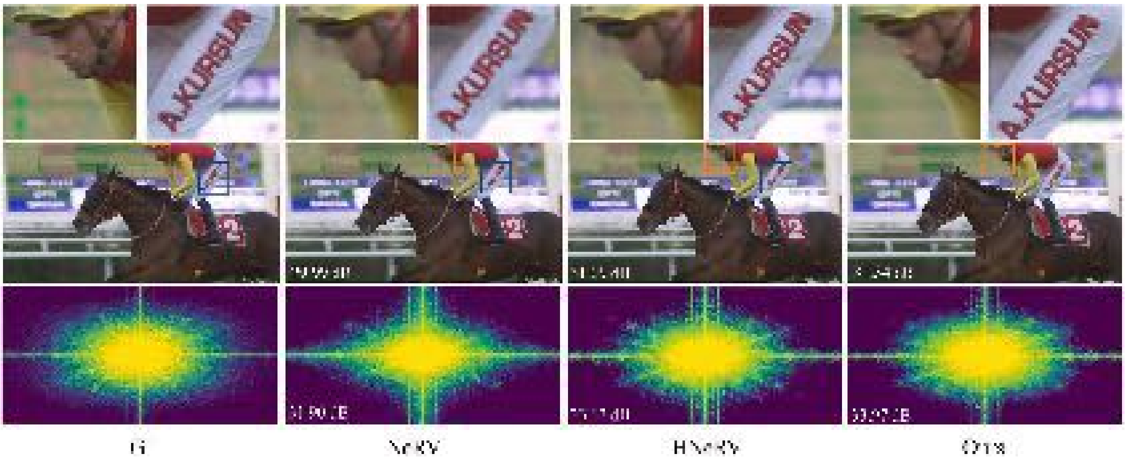

Fig. 3. Visual comparisons of reconstructed frames from different methods on the “Jockey” sequence. The first two rows show the reconstructed frames of each method and magnified details of key regions. The second row displays their corresponding frequency spectrums (frequency increasing outwards). PSNR values are shown on the bottom-left. The results shows that FANeRV achieves superior preservation of fine details and sharpness (e.g., on text and faces) than NeRV and HNeRV. The frequency spectrum of FANeRV is more similar to Ground Truth (GT) than the other two methods, with retention of more high-frequency energy.

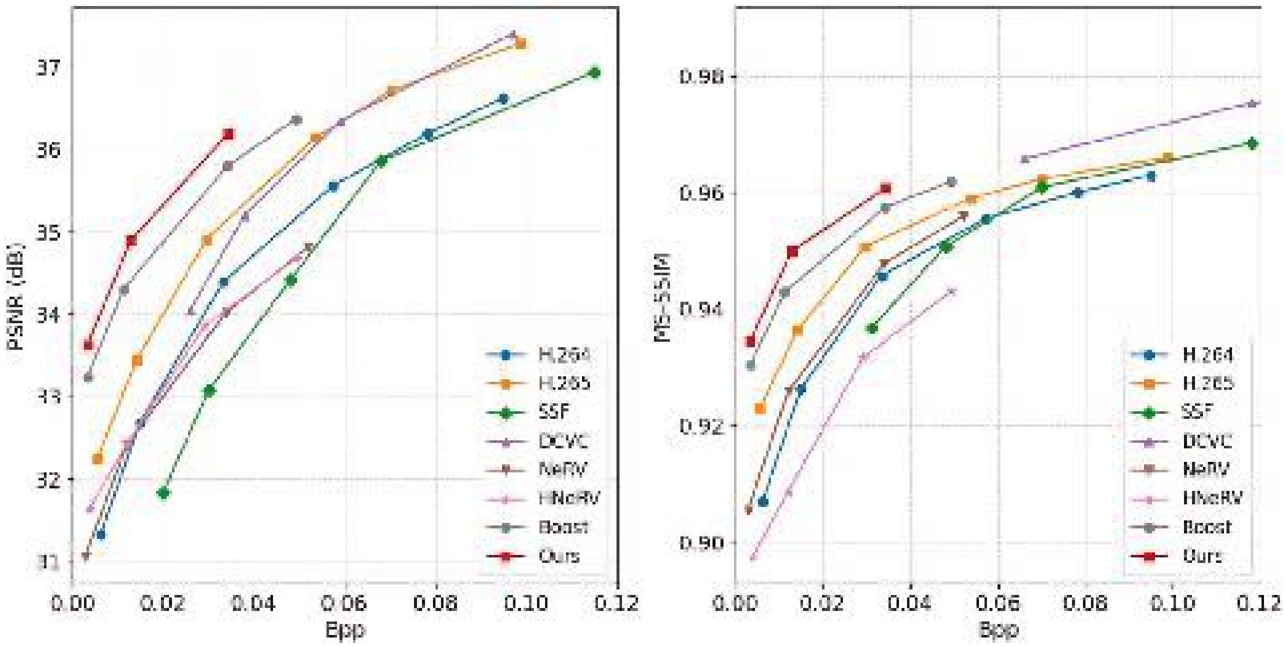

Fig. 4. The rate-distortion curve on UVG dataset in terms of PSNR and MS-SSIM.

Table 2
PSNR with varying model size and epochs on Bunny.

Method	Size			Epoch		
	0.75M	1.5M	3M	300	600	1200
NeRV	28.46	30.87	33.21	33.21	34.47	35.07
E-NeRV	30.95	32.09	36.72	36.72	38.20	39.48
HNeRV	32.18	35.19	37.43	37.43	39.36	40.02
Boost	35.53	38.95	41.50	41.50	42.03	42.34
Ours	**35.82**	**39.10**	**41.82**	**41.82**	**42.40**	**42.73**

4.2.2. Video compression

We employed the consistent entropy minimization method (Zhang et al., 2024) for model compression training. Each model was fine-tuned for 100 iterations using an initial learning rate of 5×10^{-4} with a cosine decay schedule. Our approach was benchmarked against other NeRV methods as well as traditional codecs H.264 (Wiegand et al., 2003) and H.265 (Sullivan et al., 2012) and end-to-end neural video compression methods SSF (Agustsson et al., 2020) and DCVC (Li et al., 2021). Rate-distortion (R-D) curves, evaluated using PSNR and MS-SSIM metrics on the average of the UVG dataset, are shown in Fig. 4. The visualization of these graphs provides clear evidence of our model’s superiority. FANeRV’s R-D curves are consistently positioned to the top-left relative to competing methods. This signifies that for any given bitrate (x-axis), our method achieves higher reconstruction quality (y-axis), and conversely, requires a lower bitrate to reach a specific quality level. The results indicate that, the proposed method outperforms other similar methods in both PSNR and MS-SSIM metrics, as well as traditional codecs and other end-to-end neuro-video compression methods, and this result fully proves the potential of the proposed method in terms of compression efficiency and reconstruction quality.

4.2.3. video interpolation

In order to validate the performance of the proposed method in the video frame interpolation task, the odd frames of the UVG dataset are

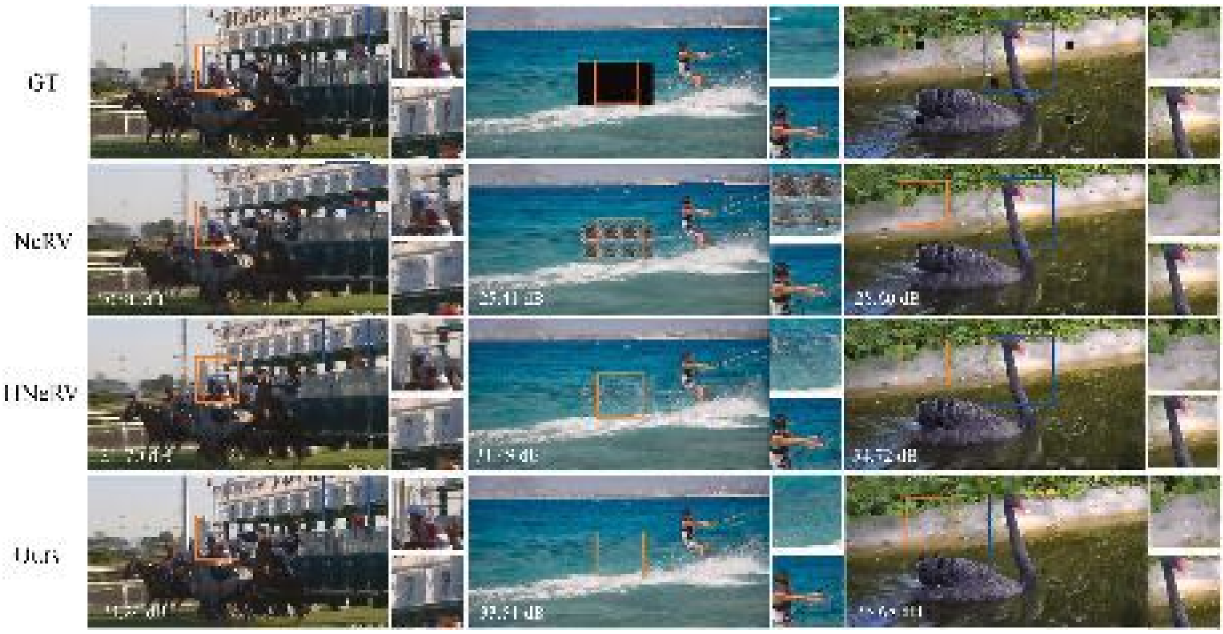

Fig. 5. The visualization comparison results are arranged in a left-to-right format, highlighting video interpolation, central inpainting (Mask-C), and dispersed inpainting (Mask-S) tasks on the UVG datasets and DAVIS dataset. PSNR values are shown on the bottom-left.

Table 3
Video interpolation results on UVG dataset in PSNR.

Method	Beauty	Bosph.	Honey.	Jockey	Ready.	Yacht.	Shake.	Avg.
NeRV	31.26	32.21	36.84	22.24	20.05	26.09	32.09	28.68
E-NeRV	31.25	33.36	38.62	22.35	20.08	26.74	<u>32.82</u>	29.32
Hnerv	31.42	34.00	39.07	<u>23.02</u>	20.71	26.74	32.58	29.65
Boost	<u>31.59</u>	<u>35.92</u>	**39.32**	22.95	<u>21.34</u>	**27.98**	32.65	<u>30.25</u>
Ours	**31.67**	**36.48**	<u>39.22</u>	**23.56**	**22.43**	<u>27.87</u>	**32.92**	**30.59**

selected as inputs for network training, while the even frames are entirely withheld and are never seen by the model during training. We adopt this odd-even frame split from standard prior works (Chen et al., 2021; Li et al., 2022; Zhang et al., 2024). This protocol provides a fair and representative evaluation by testing the model's ability to learn continuous temporal dynamics while preventing data leakage through a strict train-test separation. Next, in the inference phase, we generate the intermediate, unseen frames. This is achieved by feeding the corresponding even-numbered frame indices into the trained network, leveraging the continuous function it has learned to predict the appearance at these intermediate temporal points. Finally, for evaluation, we calculate performance metrics by comparing these generated even frames $\hat{I}_{2k}$ against the ground-truth even frames I_{2k} that were withheld. This process ensures that the reported scores in Table 3 purely reflect the model's ability to interpolate missing information rather than simply reconstruct seen data. The experimental results show that our method excels in this task, outperforming existing approaches. This is further corroborated by the visualization results in Fig. 5 (second row), where our method produces fewer artifacts and better preserves object structure in the interpolated frames, demonstrating superior performance in capturing spatio-temporal continuity.

4.2.4. *Video inpainting*

We evaluated the performance of our method on video inpainting tasks using the DAVIS validation dataset. We conducted both disperse and central masking experiments. In the disperse masking setup, five 50×50 regions were masked in each video frame during training. For the central masking experiment, a region covering one-quarter of the video's width and height was masked. The training goal was to reconstruct the complete video frame. After training, the masked region is reconstructed by the network and the reconstruction quality is evaluated in comparison with the original video. The quantitative experimental results in Table 4 show that the proposed method outperforms other methods on most of the datasets. The qualitative reconstruction results in the third and fourth rows of Fig. 5 further validate the effectiveness of the proposed method. As demonstrated in the central masking experiment, our method plausibly reconstructs the complex underlying textures with realistic details. Competing methods, however, tend to fill the masked region with overly smooth or repetitive patterns. This proves that our model, by avoiding common artifacts, can better maintain the continuity of video content and visual consistency.

4.3. *Complexity analysis*

To provide a comprehensive assessment of our model's efficiency, we evaluate its complexity from two complementary dimensions: inference complexity and storage complexity. We empirically measure inference complexity by the decoding speed, reported in Frames Per Second (FPS), where a higher value indicates lower computational cost. For storage complexity, we measure the number of model parameters (Params (M)), which reflects the intrinsic size of the model. To ensure a fair and reproducible evaluation, all speed benchmarks were conducted on the UVG dataset. The hardware platform was specified for each method: all learning-based models (our FANeRV, Boost, DCVC) were executed on a single NVIDIA RTX 4090 GPU, while the traditional H.265 codec was benchmarked on a single-threaded CPU using its FFmpeg implementation. As reported in Table 5, our proposed method demonstrates an excellent balance between these two aspects. Leveraging a streamlined network structure, FANeRV achieves decoding speeds comparable to the highly optimized H.265 implementation while maintaining a compact model size, demonstrating its computational efficiency. Crucially, FANeRV achieves its significant performance gains without additional computational or memory overhead over the strong baseline, an efficiency enabled by deliberate design choices such as the lightweight Haar DWT and depthwise separable convolutions.

Table 4
Detailed video inpainting results (PSNR) on the DAVIS dataset.

Video	Mask-S					Mask-C				
	NeRV	E-NeRV	HNeRV	Boost	Ours	NeRV	E-NeRV	HNeRV	Boost	Ours
Black.	27.06	29.53	30.20	34.10	**35.30**	24.11	26.38	26.45	29.18	**29.52**
Bmx-t.	26.77	27.75	29.05	32.99	**33.43**	22.43	**23.79**	22.28	22.28	22.81
Break.	25.48	26.97	26.34	**33.10**	32.78	20.16	**22.15**	20.23	20.24	20.61
Camel	23.70	25.70	26.13	31.08	**31.87**	21.21	**22.62**	17.74	19.81	21.30
Car-r.	23.92	26.32	28.64	31.90	**32.26**	21.24	**22.73**	21.71	22.36	22.41
Car-s.	26.58	30.63	31.01	35.85	**36.58**	23.07	23.21	21.05	**23.69**	23.65
Cows	22.17	23.92	24.68	28.30	**28.50**	20.48	21.88	21.82	24.14	**24.35**
Dnc-t.	25.29	27.42	28.74	30.79	**31.77**	21.17	22.40	21.06	21.77	**22.55**
Dog	29.29	31.72	28.80	33.87	**35.69**	25.37	**27.07**	24.16	24.66	25.83
Drf-c.	34.09	39.26	38.52	43.32	**43.89**	27.52	**29.81**	23.40	27.44	28.51
Drf-s.	26.78	29.53	30.81	36.16	**36.93**	22.76	**24.69**	18.88	21.49	23.81
Goat	24.04	25.34	26.91	30.59	**31.33**	22.03	23.43	23.06	25.10	**25.22**
Hrs-j.	25.74	29.27	29.31	30.86	**32.84**	21.54	23.06	20.72	23.16	**23.31**
Kite.	29.34	32.87	33.49	37.08	**37.26**	23.92	26.71	24.73	27.49	**27.71**
Libby	29.81	31.39	28.66	37.35	**38.27**	25.71	26.91	23.39	26.96	**27.73**
Mtx-j.	29.82	34.15	28.27	**36.42**	**36.42**	26.19	**28.75**	22.36	26.25	27.56
Para.	29.03	30.62	30.99	33.64	**34.54**	25.95	26.65	26.00	28.07	**28.68**
Prk.	24.74	25.62	26.34	28.79	**29.29**	22.32	**22.99**	19.06	20.55	20.74
Sct-b.	23.35	26.46	28.41	30.42	**31.76**	19.24	20.99	18.94	19.86	**21.88**
Soap.	27.20	28.83	30.30	32.95	**33.81**	22.29	**23.82**	17.98	19.20	22.59
Avg.	27.71	29.17	29.28	33.48	**34.23**	22.94	24.50	21.75	23.68	**24.54**

Table 5
Complexity test results on UVG dataset.

Method	Params (M) ↓	Speed (FPS) ↑	Time (ms) ↓
H.265	–	39.2	26
DCVC	35.2	1.98	505
Boost	3.07	26.5	38
Ours	3.10	33.0	30

Table 6
Ablation study of FANeRV.“✓” indicates the component is included.

Variant	FSFB	TGFN	REB	PSNR	SSIM
Ours (Full Model)	✓	✓	✓	**41.82**	**0.9942**
w/o FSFB		✓	✓	41.44	0.9936
w/o TGFN	✓		✓	41.58	0.9937
w/o REB	✓	✓		41.51	0.9939

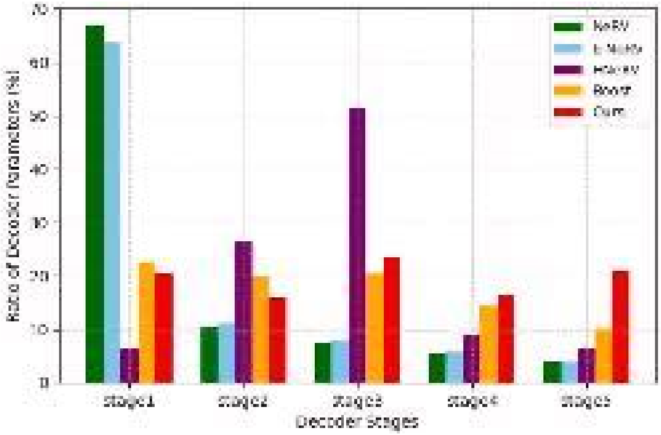

Fig. 6. Parameter distribution across decoder blocks in various models.

Table 7
Ablation studies on the modeling strategy for the low-frequency branch and the multi-resolution fusion strategy.

Variant	PSNR	SSIM
Ours	**41.82**	**0.9942**
w/ Restormer Block	41.65	0.9935
w/ VSS Block	41.57	0.9938
w/ Transposed Conv	41.57	0.9938
w/ Bilinear Interpolation	40.89	0.9929

4.4. Ablation study

To validate the effectiveness of each component in FANeRV and the rationale behind our key design choices, we conduct a series of ablation studies on the Bunny dataset. To ensure a fair comparison, all experiments in this section were conducted under identical controlled conditions. This includes using the exact same training strategies and ensuring that all architectural alternatives have a closely matched parameter count (approx. 3.1M).

First, we analyze the contribution of our three main architectural modules: the Frequency Separation Feature Boosting (FSFB) module, the Time-Modulated Gated Feed-Forward Network (TGFN), and the final Residual Enhancement Block (REB). As shown in Table 6, removing any of these components leads to a noticeable drop in performance, confirming their complementary and critical roles. The REB's importance is further contextualized in Fig. 6, which shows how it helps allocate more parameters to the network's later stages, enhancing the capacity for fine detail recovery.

Next, we validate our specific design choices against strong alternatives, with results summarized in Table 7. To evaluate our Frequency-Specific Multi-Resolution Fusion strategy, we replace it with conventional upsampling operators used in prior works (Chen et al., 2025; Kim et al., 2025), namely bilinear interpolation and transposed convolution. The quantitative results clearly show our method's superiority. This performance gain is visually substantiated in Fig. 7, which illustrates how our fusion mechanism completely circumvents the blurring and checkerboard artifacts that plague these traditional operators, thereby preserving feature integrity during upsampling. To validate our low-frequency branch design, we benchmark it against popular methods for capturing long-range dependencies, such as efficient self-attention (Restormer Block) (Zamir et al., 2022) and the Mamba block (VSS Block), a state-space model that has recently shown great promise in image restoration (Guo et al., 2024). The results show that while modern general-purpose blocks are effective, our designed lightweight multi-resolution deep feature modulation branch effectively models long-range

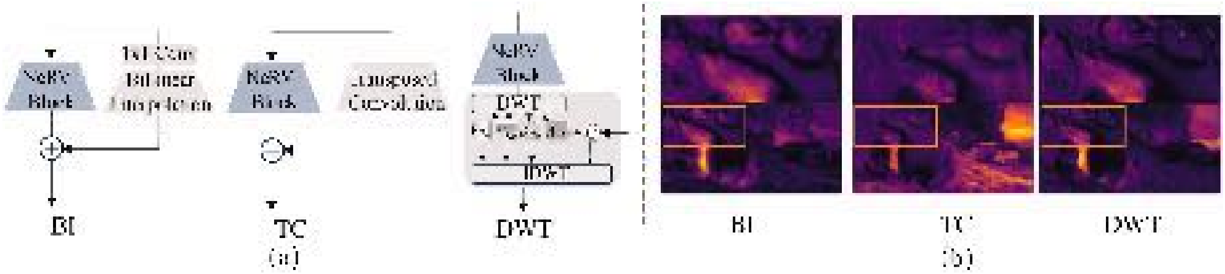

Fig. 7. Comparative analysis of multi-resolution fusion strategies. (a) Schematic diagrams of Bilinear Interpolation Fusion (BI), Transposed Convolution Fusion (TC), and the proposed Frequency-Specific Multi-Resolution Fusion (DWT). (b) Corresponding intermediate feature maps, visually demonstrating that BI leads to blurred features and TC introduces checkerboard artifacts. In contrast, the proposed fusion method effectively mitigates these upsampling-induced issues, achieving superior artifact reduction and detail preservation.

dependencies while maintaining a compact set of parameters, thus achieving a better trade-off between performance and efficiency.

Collectively, these studies verify that each component of FANeRV is essential and that our core design principles are superior to both conventional and other modern alternatives for this task.

5. Conclusion

This paper introduced FANeRV, a novel video neural representation framework designed to overcome the spectral bias of neural networks. Its architecture, which leverages an Asymmetric Discrete Wavelet Transform (DWT) and Frequency-Specific Multi-Resolution Fusion, delivers significant improvements in video regression, outperforming the strongest baseline by 0.64 dB in PSNR on the UVG dataset. The framework's effectiveness and flexibility are also proven by its strong performance in video compression, interpolation, and inpainting. Crucially, FANeRV achieves these results while remaining efficient, with a competitive model size and fast decoding speed. Future work will aim to reduce implementation complexity and enhance performance on dynamic scenes by integrating explicit motion priors, such as optical flow.

CRediT authorship contribution statement

Li Yu: Conceptualization, Methodology, Software, Writing – original draft, Writing – review & editing. **Zhihui Li:** Methodology, Writing – original draft, Validation, Writing – review & editing. **Chao Yao:** Supervision, Project administration. **Jimin Xiao:** Methodology, Supervision. **Moncef Gabbouj:** Supervision, Project administration.

Data availability

Data will be made available on request.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China under grant 62002172; and in part by The Startup Foundation for Introducing Talent of NUIST under grant 2023r131.

References

Agustsson, E., Minnen, D., Johnston, N., Balle, J., Hwang, S. J., & Toderici, G. (2020). Scale-space flow for end-to-end optimized video compression. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 8503–8512).

Chen, H., Gwilliam, M., Lim, S.-N., & Shrivastava, A. (2023). HNeRV: A hybrid neural representation for videos. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 10270–10279).

Chen, H., He, B., Wang, H., Ren, Y., Lim, S. N., & Shrivastava, A. (2021). NeRV: Neural representations for videos. *Advances in Neural Information Processing Systems*, *34*, 21557–21568.

Chen, J., Liu, X., Chen, B., An, B., Tao, D., & Xia, S.-T. (2025). LNeRV: Learnable hierarchical encoding improve neural representation video codec. In *ICASSP 2025-2025 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 1–5). IEEE.

Chen, Z., Chen, Y., Liu, J., Xu, X., Goel, V., Wang, Z., Shi, H., & Wang, X. (2022). Videoinr: Learning video implicit neural representation for continuous space-time super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 2047–2057).

Daubechies, I. (1990). The wavelet transform, time-frequency localization and signal analysis. *IEEE Transactions on Information Theory*, *36*(5), 961–1005.

Ding, X., Zhang, X., Han, J., & Ding, G. (2022). Scaling up your kernels to 31×31: Revisiting large kernel design in CNNS. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 11963–11975).

Dupont, E., Goliński, A., Alizadeh, M., Teh, Y. W., & Doucet, A. (2021). Coin: Compression with implicit neural representations. arXiv:2103.03123.

Ghorbel, A., Hamidouche, W., & Morin, L. (2024). NeRV + + : An enhanced implicit neural video representation. In *2024 IEEE international conference on visual communications and image processing (VCIP)* (pp. 1–5). IEEE.

Guo, H., Li, J., Dai, T., Ouyang, Z., Ren, X., & Xia, S.-T. (2024). Mambair: A simple baseline for image restoration with state-space model. In *European conference on computer vision* (pp. 222–241). Springer.

Guo, M.-H., Lu, C.-Z., Liu, Z.-N., Cheng, M.-M., & Hu, S.-M. (2023). Visual attention network. *Computational Visual Media*, *9*(4), 733–752.

Hayami, T., Koizumi, K., & Watanabe, H. (2025). Sr-nerv: Improving embedding efficiency of neural video representation via super-resolution. arXiv:2505.00046.

Hu, K., Liu, Y., Xu, F., Liu, R., Wang, H., & Song, S. (2024). Asymmetric neural image compression with high-preserving information. In *2024 IEEE international symposium on circuits and systems (ISCAS)* (pp. 1–5). IEEE.

Jung, H., Hui, Z., Luo, L., Yang, H., Liu, F., Yoo, S., Ranjan, R., & Demandolx, D. (2023). Anyflow: Arbitrary scale optical flow with implicit neural representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 5455–5465).

Kim, J., Lee, J., & Kang, J.-W. (2025). SNeRV: Spectra-preserving neural representation for video. In *European conference on computer vision* (pp. 332–348). Springer.

Li, J., Li, B., & Lu, Y. (2021). Deep contextual video compression. *Advances in Neural Information Processing Systems*, *34*, 18114–18125.

Li, Z., Wang, H., & Meng, D. (2023). Regularize implicit neural representation by itself. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 10280–10288).

Li, Z., Wang, M., Pi, H., Xu, K., Mei, J., & Liu, Y. (2022). E-NeRV: Expedite neural video representation with disentangled spatial-temporal context. In *European conference on computer vision* (pp. 267–284). Springer.

Liu, P., Zhang, H., Zhang, K., Lin, L., & Zuo, W. (2018). Multi-level wavelet-CNN for image restoration. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops* (pp. 773–782).

Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., & Xie, S. (2022). A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 11976–11986).

Lu, Y., Wang, Z., Liu, M., Wang, H., & Wang, L. (2023). Learning spatial-temporal implicit neural representations for event-guided video super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 1557–1567).

Lu, Z., Li, J., Liu, H., Huang, C., Zhang, L., & Zeng, T. (2022). Transformer for single image super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 457–466).

Mercat, A., Viitanen, M., & Vanne, J. (2020). UVG dataset: 50/120fps 4k sequences for video codec analysis and development. In *Proceedings of the 11th ACM multimedia systems conference* (pp. 297–302).

Mildenhall, B., Srinivasan, P. P., Tancik, M., Barron, J. T., Ramamoorthi, R., & Ng, R. (2021). Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, *65*(1), 99–106.

Perazzi, F., Pont-Tuset, J., McWilliams, B., Van Gool, L., Gross, M., & Sorkine-Hornung, A. (2016). A benchmark dataset and evaluation methodology for video object segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 724–732).

Rahaman, N., Baratin, A., Arpit, D., Draxler, F., Lin, M., Hamprecht, F., Bengio, Y., & Courville, A. (2019). On the spectral bias of neural networks. In *International conference on machine learning* (pp. 5301–5310). PMLR.

Rho, D., Cho, J., Ko, J. H., & Park, E. (2022). Neural residual flow fields for efficient video representations. In *Proceedings of the Asian conference on computer vision* (pp. 3447–3463).

Roosendaal, T. (2008). Big buck bunny. In *ACM SIGGRAPH Asia 2008 computer animation festival* (pp. 62).

Saragadam, V., LeJeune, D., Tan, J., Balakrishnan, G., Veeraraghavan, A., & Baraniuk, R. G. (2023). Wire: Wavelet implicit neural representations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 18507–18516).

Sitzmann, V., Martel, J., Bergman, A., Lindell, D., & Wetzstein, G. (2020). Implicit neural representations with periodic activation functions. *Advances in Neural Information Processing Systems*, *33*, 7462–7473.

Sitzmann, V., Zollhöfer, M., & Wetzstein, G. (2019). Scene representation networks: Continuous 3d-structure-aware neural scene representations. *Advances in Neural Information Processing Systems*, *32*.

Skorokhodov, I., Ignatyev, S., & Elhoseiny, M. (2021). Adversarial generation of continuous images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 10753–10764).

Skorokhodov, I., Tulyakov, S., & Elhoseiny, M. (2022). Stylegan-V: A continuous video generator with the price, image quality and perks of stylegan2. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 3626–3636).

Strümpler, Y., Postels, J., Yang, R., Gool, L. V., & Tombari, F. (2022). Implicit neural representations for image compression. In *European conference on computer vision* (pp. 74–91). Springer.

Sullivan, G. J., Ohm, J.-R., Han, W.-J., & Wiegand, T. (2012). Overview of the high efficiency video coding (HEVC) standard. *IEEE Transactions on Circuits and Systems for Video Technology*, *22*(12), 1649–1668.

Wiegand, T., Sullivan, G. J., Bjontegaard, G., & Luthra, A. (2003). Overview of the h. 264/AVC video coding standard. *IEEE Transactions on Circuits and Systems for Video Technology*, *13*(7), 560–576.

Wu, C., Quan, G., He, G., Lai, X.-Q., Li, Y., Yu, W., Lin, X., & Yang, C. (2024a). Qs-NeRV: Real-time quality-scalable decoding with neural representation for videos. In *Proceedings of the 32nd ACM international conference on multimedia* (pp. 2584–2592).

Wu, Z., Sun, C., Xuan, H., Liu, G., & Yan, Y. (2024b). Waveformer: wavelet transformer for noise-robust video inpainting. In *Proceedings of the AAAI conference on artificial intelligence* (pp. 6180–6188). (*38*).

Xie, X., Zhou, P., Li, H., Lin, Z., & Yan, S. (2022). Adan: Adaptive nesterov momentum algorithm for faster optimizing deep models. arXiv:2208.06677.

Xu, G., Liao, W., Zhang, X., Li, C., He, X., & Wu, X. (2023). Haar wavelet downsampling: A simple but effective downsampling module for semantic segmentation. *Pattern Recognition*, *143*, 109819.

Xu, Y., Feng, X., Qin, F., Ge, R., Peng, Y., & Wang, C. (2024). Vq-nerv: A vector quantized neural representation for videos. arXiv:2403.12401. unpublished.

Xu, Z.-Q. J., Zhang, Y., & Xiao, Y. (2019). Training behavior of deep neural network in frequency domain. In *Neural information processing: 26th international conference, ICONIP 2019, Sydney, NSW, Australia, December 12–15, 2019, proceedings, Part I 26* (pp. 264–274). Springer.

Yang, K., Hu, T., Dai, K., Chen, G., Cao, Y., Dong, W., Wu, P., Zhang, Y., & Yan, Q. (2024). Crnet: A detail-preserving network for unified image restoration and enhancement task. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 6086–6096).

Yu, L., Li, Z., Xiao, J., & Gabbouj, M. (2025). High-frequency enhanced hybrid neural representation for video compression. *Expert Systems with Applications*, (p. 127552).

Zamir, S. W., Arora, A., Khan, S., Hayat, M., Khan, F. S., & Yang, M.-H. (2022). Restormer: Efficient transformer for high-resolution image restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 5728–5739).

Zhang, X., Yang, R., He, D., Ge, X., Xu, T., Wang, Y., Qin, H., & Zhang, J. (2024). Boosting neural representations for videos with a conditional decoder. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)* (pp. 2556–2566).

Zhang, Y., Van Rozendaal, T., Brehmer, J., Nagel, M., & Cohen, T. (2021). Implicit neural video compression. arXiv:2112.11312.

Zhao, J., Li, X. J., & Chong, P. H. J. (2024). HFS-HNeRV: High-frequency spectrum hybrid neural representation for videos. In *Proceedings of the 6th ACM international conference on multimedia in Asia* (pp. 1–7).