DaReNeRF: Direction-aware Representation for Dynamic Scenes

Ange Lou[1,2,*] Benjamin Planche[1] Zhongpai Gao[1] Yamin Li[2]
Tianyu Luan[1,3,*] Hao Ding[1,4,*] Terrence Chen[1] Jack Noble[2] Ziyan Wu[1]
[1]United Imaging Intelligence [2]Vanderbilt University [3]Johns Hopkins University [4]University of Buffalo
{first.last}@uii-ai.com, {first.last}@vanderbilt.edu, tianyulu@buffalo.edu, hding15@jhu.edu

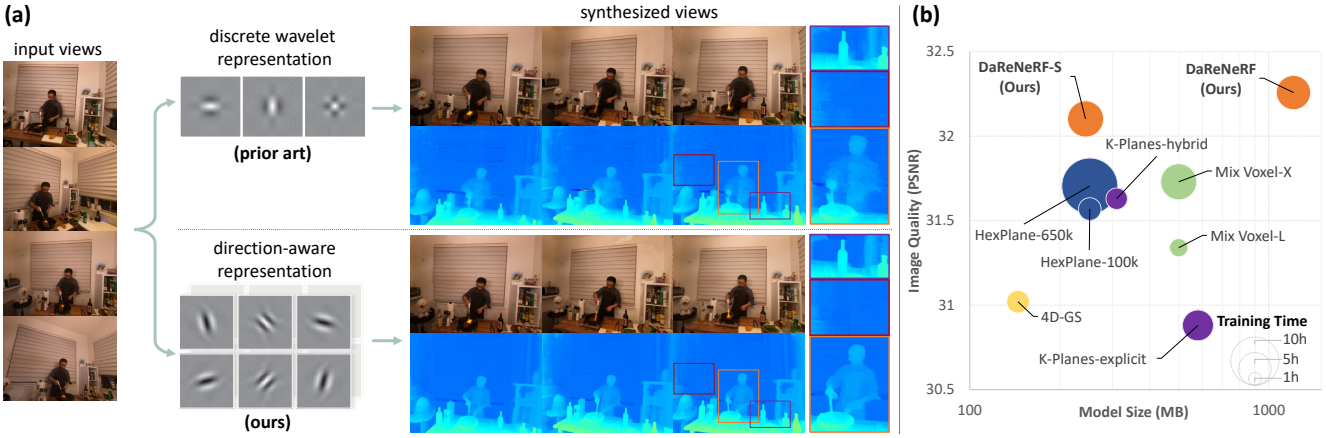

Figure 1. **DaReNeRF performance on dynamic 3D scenes.** Our proposed direction-aware representation excels by capturing features of dynamic scenes from six different directions—a capability beyond the reach of traditional discrete-wavelet representations, *cf*. sub-figure (a). Built upon this advanced representation, our NeRF method outperforms prior work in challenging dynamic scenarios while being competitive in terms of training time and model size, offering the best trade-off overall, *cf*. sub-figure (b).

Abstract

Addressing the intricate challenge of modeling and re-rendering dynamic scenes, most recent approaches have sought to simplify these complexities using plane-based explicit representations, overcoming the slow training time issues associated with methods like Neural Radiance Fields (NeRF) and implicit representations. However, the straightforward decomposition of 4D dynamic scenes into multiple 2D plane-based representations proves insufficient for re-rendering high-fidelity scenes with complex motions. In response, we present a novel direction-aware representation (DaRe) approach that captures scene dynamics from six different directions. This learned representation undergoes an inverse dual-tree complex wavelet transformation (DTCWT) to recover plane-based information. DaReNeRF computes features for each space-time point by fusing vectors from these recovered planes. Combining DaReNeRF with a tiny MLP for color regression and leveraging volume rendering in training yield state-of-the-art performance in novel view synthesis for complex dynamic scenes. Notably, to address redundancy introduced by the six real and six imaginary direction-aware wavelet coefficients, we introduce a trainable masking approach, mitigating storage issues without significant performance decline. Moreover, DaReNeRF maintains a 2× reduction in training time compared to prior art while delivering superior performance.

1. Introduction

The reconstruction and re-rendering of 3D scenes from a set of 2D images pose a fundamental challenge in computer vision, holding substantial implications for a range of AR/VR applications [36, 55, 59]. Despite recent progress in reconstructing static scenes, significant challenges remain. Real-world scenes are inherently dynamic, characterized by intricate motion, further adding to the task complexity.

Recent dynamic scene reconstruction methods build on NeRF's implicit representation. Some utilize a large MLP to process spatial and temporal point positions, generating color outputs [22, 23, 54]. Others aim to disentangle scene motion and appearance [15, 27, 38–40]. However, both ap-

*This work was carried out during the internships of Ange Lou, Tianyu Luan, and Hao Ding at United Imaging Intelligence, Burlington, MA.

proaches face computational challenges, requiring extensive MLP evaluations for novel view rendering. The slow training process, often spanning days or weeks, and the reliance on additional supervision like depth maps [23, 24, 27] limit their widespread adoption for dynamic scene modeling. Several recent studies [7, 14, 46] have proposed decomposition-based methods to address the training time challenge. However, relying solely on decomposition limits NeRF's ability to capture high-fidelity texture details.

Recent studies [42, 52, 58, 64, 66] have explored the possibility of incorporating frequency information into NeRF. These frequency-based representations demonstrate promising performance in static-scene rendering, particularly in recovering detailed information. However, there is limited exploration w.r.t. the ability of these methods to scale from static to dynamic scenes. Additionally, HexPlane [7] has noted a significant degradation in reconstruction performance when using wavelet coefficients as a basis. This limitation is inherent to wavelets themselves, and we delve into a detailed discussion in the following paragraph.

Traditional 2D discrete wavelet transform (DWT) employs low/high-pass real wavelets to decompose a 2D image or grid into approximation and detail wavelet coefficients across different scales. These coefficients offer an efficient representation of both global and local image details. However, there are two significant drawbacks hindering the successful application of 2D DWT-based representations to dynamic scenes. The first is the **shift variance** problem [6], where even a small shift in the input signal significantly disrupts the wavelets' oscillation pattern. In dynamic 3D scenes, shifts are more pronounced than in static scenarios due to factors such as multi-object motion, camera motion, reflections, and variations in illumination. Simple DWT wavelet representations struggle to handle such variability, yielding poor results in dynamic regions. Another critical issue is the **poor direction selectivity** [20] in DWT representations. A 2D DWT produces a checkered pattern that blends representations from $\pm 45°$, lacking directional selectivity, which is less effective for capturing lines and edges in images. Consequently, DWT-based representations fail to adequately model dynamic scenes, leading to results with noticeable ghosting artifacts around moving objects as shown in Figure 1.

This paper addresses these key limitations of the discrete wavelet transform (DWT) by introducing an efficient and robust frequency-based representation designed to overcome the challenges of shift variance and lack of direction selectivity in modeling dynamic scenes. Inspired by the dual-tree complex wavelet transform (DTCWT) [45], we propose a direction-aware representation, aiming to learn features from six distinct orientations without introducing the checkerboard pattern observed in DWT. Leveraging the properties of complex wavelet transforms, our approach ensures shift invariance within the representation. The proposed direction-aware representation proves successful in modeling complex dynamic scenes, achieving state-of-the-art performance.

Furthermore, we observe that our proposed direction-aware representation introduces a 2^d redundancy (with $d = 2$ for plane-based decomposition) compared to the DWT representation, resulting in lower storage efficiency. To address this storage challenge, we leverage a compression pipeline originally designed for static 3D scenes and adapt it for dynamic scenes. This migration of the compression pipeline proves effective in mitigating the storage constraints inherent in the direction-aware representation, making it as memory-efficient as recent state-of-the-art methods.

Additionally, to highlight the generalizability of our proposed method (aimed at 4D scenarios), we extend its application to modelling static 3D scenes. In this context, DaReNeRF demonstrates high-fidelity reconstruction performance and efficient storage capabilities. This versatility underscores the efficacy of our approach not only in dynamic scenes but also in static environments. This affirms its potential as a general representation utility across various scenarios.

In summary, our contributions are as follows:

- We are the first to leverage DTCWT in NeRF optimization, introducing a direction-aware representation to address the shift-variance and direction-ambiguity shortcomings in DWT-based representations. DaReNeRF thereby outperforms prior decomposition-based methods in modeling complex dynamic scenes.
- We implement a trainable mask method for dynamic scene reconstruction, effectively resolving the storage limitations associated with the direction-aware representation. This adaptation ensures that it attains comparable memory efficiency with the current state-of-the-art methods.
- We extend our direction-aware representation to static scene reconstruction, and experiments demonstrate that our proposed method outperforms other state-of-the-art approaches, achieving a superior trade-off between performance and model size.

2. Related Work

Neural Scene Representation. NeRF [33] and its variants [2–4, 29, 34, 37, 53, 65] show impressive results on novel view synthesis and many other application including 3D reconstruction [21, 30, 70, 71], semantic segmentation [26, 35], object detection [17, 61–63], generative model [8, 9, 60], and 3D content creation [11, 31, 56]. Implicit neural representation exhibit remarkable imaging quality but suffer from slow rendering due to the numerous costly MLP evaluations required for each pixel. Numerous spatial decomposition methods [1, 9, 10, 13] have been proposed to address the challenge of training speed in static scenes.

Further applying neural radiance fields to dynamic scenes is a crucial challenge. One straightforward approach involves extending a static NeRF by introducing an addi-

tional time dimension [40] or latent code [16, 24, 27, 50]. While these methods demonstrate strong capabilities in modeling complex real-world dynamic scenes, they face a severely under-constrained problem that necessitates additional supervision—*e.g.*, depth, optical flow, and dense observations—to achieve satisfactory results. The substantial system size and weeks-long training times associated with these approaches hinder their real-world applicability. Another solution involves employing individual MLPs to represent the deformation field and a canonical field [19, 40, 47, 65, 68]. The latter field depicts a static scene, while the former learns coordinate mappings to the canonical space over time. Although this is an improvement over the first approach, it still requires considerable training time.

Scene Decomposition. Recently, decomposition-based methods [7, 14, 46] have emerged for dynamic scenes. These approaches aim to alleviate the lengthy training times associated with dynamic scenes while maintaining the ability to model their complexity. They decompose a 4D scene into plane-based representations and employ a compact MLP to aggregate features for volumetric rendering of resulting images. While these methods significantly reduce training time and memory storage, they still encounter challenges in preserving detailed texture information during rendering.

Wavelet-based representations [42, 44, 64] have garnered significant attention for enhancing NeRF's ability to capture such fine texture details, owing to their capacity for recovering high-fidelity signals. However, there has been limited exploration of the potential of wavelet-based representations for dynamic scene modeling. Applying wavelet-based representations directly to plane-based methods can lead to a significant performance decay, as illustrated in Figure 1. Similar degradation is also reported by HexPlane [7], highlighting the inherent limitations of wavelets, namely, shift variance and direction ambiguity. To overcome these limitations and build a more effective general dynamic NeRF, we propose a direction-aware representation, which preserves the ability to detect detailed textures without requiring additional supervision, achieving state-of-the-art performance in real-world dynamic scene reconstruction.

3. Method

We seek to develop a model for a dynamic scene using a collection of posed images, each timestamped. The objective is to fit a model capable of rendering new images at varying poses and time stamps. Similar to D-NeRF [40], this model assigns color and opacity to points in both space and time. The rendering process involves differentiable volumetric rendering along rays. Training the entire model relies on a photometric loss function, comparing rendered images with ground-truth images to optimize model parameters.

Our primary innovation lies in introducing a novel direction-aware representation for dynamic scenes. This distinctive representation is coupled with the inverse dual-tree complex wavelet transform (IDTCWT) and a compact implicit multi-layer perceptron (MLP) to enable the generation of high-fidelity novel views. Figure 2 shows an overview of the model. Note that for simplicity, we refer to the wavelet representation as wavelet coefficients in this section.

3.1. Dynamic Scene Decomposition

A natural dynamic scene can be represented as a 4D spatio-temporal volume denoted as D. This 4D volume comprises individual static 3D volume for each time step, namely $\{V_1, V_2, ..., V_T\}$. Directly modeling a 4D volume would entail a memory complexity of $\mathcal{O}(N^3TF)$, where N, T, F are spatial resolution, temporal resolution and feature size (with $F = 3$ representing RGB colors). To improve the overall performance, we propose a direction-aware representation applied to baseline plane-based 4D volume decomposition [7]. In such baseline, a representation of the 4D volume can be represented as follows:

$$D = \sum_{r=1}^{R_1} M_r^{XY} \circ M_r^{ZT} \circ v_r^1 + \sum_{r=1}^{R_2} M_r^{XZ} \circ M_r^{YT} \circ v_r^2 + \sum_{r=1}^{R_3} M_r^{YZ} \circ M_r^{XT} \circ v_r^3 \tag{1}$$

where each $M_r^{AB} \in \mathbb{R}^{AB}$ represents a learned 2D plane-based representation with $\{(A, B) \in \{X, Y, Z, T\}^2 \mid A \neq B\}$", and $v_r^i \in \mathbb{R}^F$ are learned vectors along F axes. The parameters R_1, R_2 and R_3 correspond to the number of low rank components. By defining $R = R_1 + R_2 + R_3 \ll N$, the model's memory complexity can be notably reduced from $\mathcal{O}(N^3TF)$ to $\mathcal{O}(RN^2TF)$. This reduction in memory requirements proves advantageous for efficiently modeling dynamic scenes while preserving computational resources.

To compute the density and appearance features of points in space-time, the model multiplies the feature vectors extracted from paired planes (*e.g.*, XY and ZT), concatenates the multiplied results into a single vector, and then multiplies them by V^{RF}, which stacks all v_r^i into a 2D tensor. The point opacities are directly queried from the density features. The RGB color values are regressed by a compact MLP, where the inputs are appearance features and view directions. Finally, images are synthesized via volumetric rendering. To improve the overall performance, we apply our proposed direction-aware representation to this baseline.

3.2. Direction-Aware Representation

Built upon plane-based 4D volume decomposition and drawing inspiration from the dual-tree complex wavelet transform, we introduce the direction-aware representation. This innovative approach enables the modeling of representations from six different directions. In contrast to the prevalent use of 2D discrete wavelet transforms (DWT), the dual tree complex wavelet transform (DTCWT) [45] employs two com-

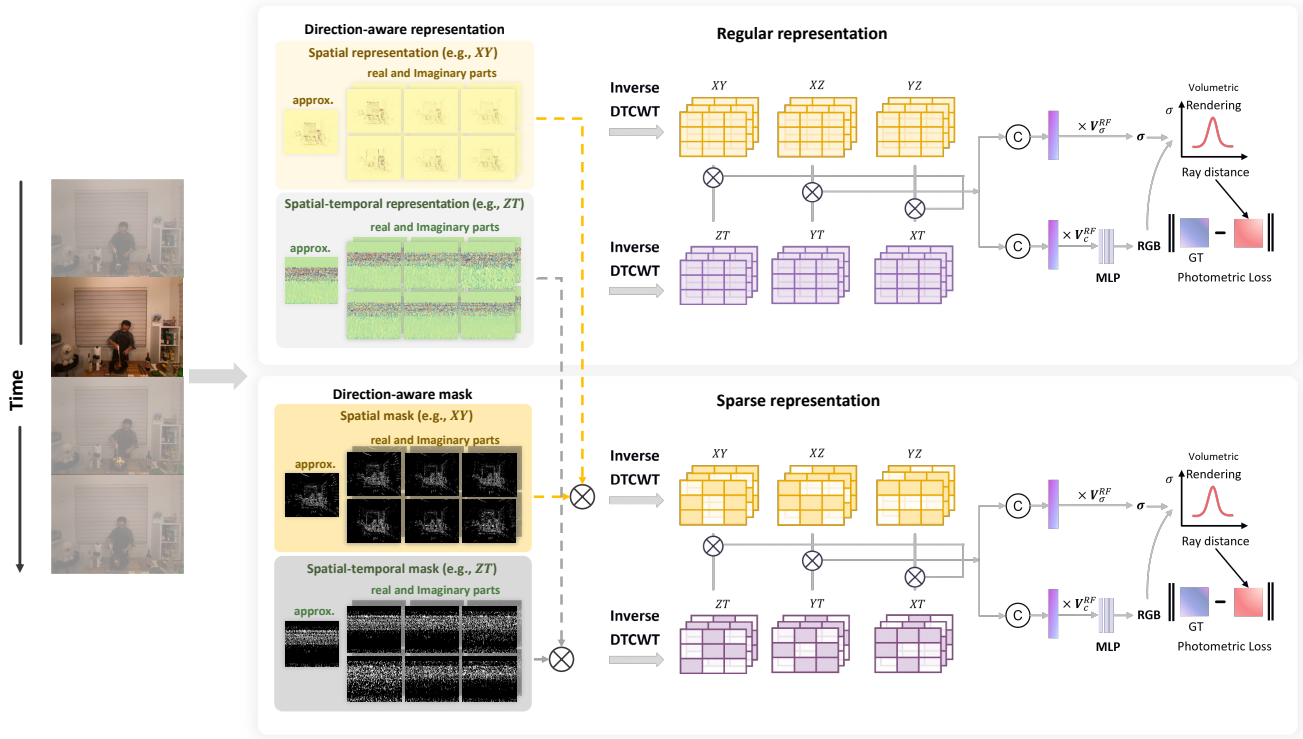

Figure 2. **Method Overview. Top:** The regular DaReNeRF architecture comprises an approximation and 12 direction-aware coefficient maps for both spatial (*e.g.*, XY) and spatial-temporal (*e.g.*, ZT) plane-based representations. To compute features of points in space-time, it multiplies feature vectors extracted from paired planes (*e.g.*, XY and ZT), concatenates the multiplied results into a single vector, and then multiplies them by learned tensor V^{RF} for final results. RGB colors are regressed by a compact MLP, and images are synthesized via volumetric rendering. **Bottom:** The trainable mask is combined with the top architecture to create a sparse DaReNeRF. Each direction-aware representation and the approximation representation are assigned their own sparse masks. The sparse representation undergoes an inverse dual tree complex wavelet transform to obtain plane-based spatial and spatial-temporal representations.

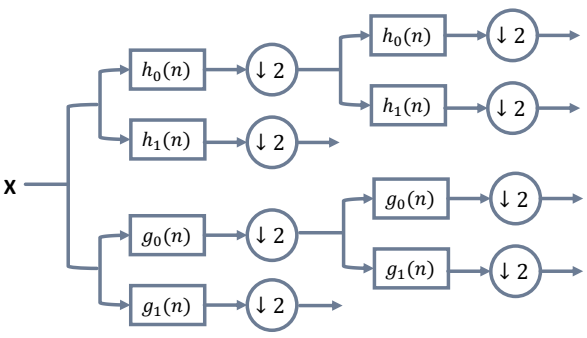

Figure 3. **Analysis Filter Bank**, for the dual tree complex wavelet transform.

plex wavelets as illustrated in Figure 3. Given $h = [h_0, h_1]$ and $g = [g_0, g_1]$ low/high pass filter pairs for upper (real) and lower (imaginary) filter banks, the low-pass and high-pass complex wavelet transforms in DTCWT are denoted as $\phi(x) = \phi_h(x) + j\phi_g(x)$ and $\psi(x) = \psi_h(x) + j\psi_g(x)$. Consequently, applying low- and high-pass complex wavelet transforms to rows and columns of a 2D grid yields wavelet coefficients $\phi(x)\psi(y)$, $\psi(x)\phi(y)$ and $\psi(x)\psi(y)$. Due to filter design, the upper (real) filter and lower filter (imaginary) satisfy the Hilbert transform, denoted as $\psi_g(x) \approx \mathcal{H}(\psi_h(x))$. Finally, three additional wavelet coefficients, $\phi(x)\overline{\psi(y)}$, $\psi(x)\overline{\phi(y)}$ and $\psi(x)\overline{\psi(y)}$, can be obtained, where $\overline{\phi}$ and $\overline{\psi}$ represent the complex conjugate of ϕ and ψ. From these 2D wavelet coefficients, we derive six direction-aware real and imaginary wavelet coefficients, each with the same set of six directions. Compared to 2D DWT, the six wavelet coefficients align along specific directions, eliminating the checkerboard effect, with more results in the supplementary material.

Exploiting the properties of DTCWT, we aim for the plane-based representation $M_r^{AB} \in \mathbb{R}^{m,n}$ of the 4D volume to possess direction-aware capabilities. Here, m and n denote the resolution of the 2D plane-based representation. To imbue each 2D plane-based representation with

direction-aware capabilities, we introduce twelve learned wavelet coefficients—six for the real part and six for the imaginary part—denoted as $\mathbf{R}\{\mathcal{W}_i^{AB}\}_{i=1}^6 \in \mathbb{R}^{m/2^l,n/2^l}$ and $\mathbf{I}\{\mathcal{W}_i^{AB}\}_{i=1}^6 \in \mathbb{R}^{m/2^l,n/2^l}$, respectively. Additionally, a learned approximation coefficient is defined as $\mathcal{W}_a^{AB} \in \mathbb{R}^{m/2^{l-1},n/2^{l-1}}$, with l the DTCWT transformation level. Consequently, a specific plane-based representation can be expressed as:

$$M_r^{AB} = IDTCWT([W_{a,r}^{AB}, \mathbf{R}\{\mathcal{W}_{i,r}^{AB}\}_{i=1}^6, \mathbf{I}\{\mathcal{W}_{i,r}^{AB}\}_{i=1}^6]) \tag{2}$$

Importantly, our representation is not only applicable for modeling dynamic 3D scenes but is also well-suited for static 3D scenes, following a TensorRF-like [10] decomposition:

$$\begin{aligned} D = &\sum_{r=1}^{R_1} M_r^{XY} \circ v_r^Z \circ v_r^1 + \sum_{r=1}^{R_2} M_r^{XZ} \circ v_r^Y \circ v_r^2 \\ &+ \sum_{r=1}^{R_3} M_r^{YZ} \circ v_r^X \circ v_r^3 \end{aligned} \tag{3}$$

In this formulation, a plane-based representation $M_r^{AB} \in \mathbb{R}^{AB}$ and a vector-based representation $v_r^C \in \mathbb{R}^C$ are employed to model a 3D volume. For static scenes, our direction-aware representations also could be applied to represent the plane-based representations.

3.3. Sparse Representation and Model Compression

In contrast to the classical 2D discrete wavelet transform (DWT), our direction-aware representation excels in modeling dynamic 3D scenes. However, it is worth noting that a single-level dual tree complex wavelet transform (DTCWT) necessitates six real direction-aware wavelet coefficients and six imaginary direction-aware wavelet coefficients to impart directional information to the plane-based representation. In contrast, a single-level 2D DWT only has three real wavelet coefficients, albeit with inherent direction ambiguity. To enhance the storage efficiency of our solution, we employ learned masks [42] for each directional wavelet coefficient, selectively masking out less important features.

To address the 2^d redundancies, where $d = 2$ for the 2D DTCWT transform, we employ learned masks $\mathbf{R}\{\mathcal{M}_i^{AB}\}_{i=1}^6 \in \mathbb{R}^{m/2^l,n/2^l}$, $\mathbf{I}\{\mathcal{M}_i^{AB}\}_{i=1}^6 \in \mathbb{R}^{m/2^l,n/2^l}$ and $\mathcal{M}_a^{AB} \in \mathbb{R}^{m/2^{l-1},n/2^{l-1}}$ for the six real wavelet coefficients, six imaginary wavelet coefficients and the approximation coefficients, respectively. The masked wavelet coefficients can be denoted as:

$$\widehat{\mathcal{W}}^{AB} = \mathrm{sg}\left(\left(\mathbf{H}(\mathcal{M}^{AB}) - \mathrm{sigmoid}(\mathcal{M}^{AB})\right) \odot \mathcal{W}^{AB}\right) \tag{4}$$

here $\left\{\mathbf{R}\{\mathcal{M}_i^{AB}\}_{i=1}^6, \mathbf{I}\{\mathcal{M}_i^{AB}\}_{i=1}^6, \mathcal{M}_a^{AB}\right\} \in \mathcal{M}^{AB}$ and $\left\{\mathbf{R}\{\mathcal{W}_i^{AB}\}_{i=1}^6, \mathbf{I}\{\mathcal{W}_i^{AB}\}_{i=1}^6, \mathcal{W}_a^{AB}\right\} \in \mathcal{W}^{AB}$. The functions sg, $\mathbf{H}$ and sigmoid represent the stop-gradient operator, Heaviside step and element-wise sigmoid function, respectively. The masked plane-based representation is obtained from the masked wavelet coefficients through the equation:

$$\widehat{M}_r^{AB} = IDTCWT([\widehat{W}_{a,r}^{AB}, \mathbf{R}\{\widehat{\mathcal{W}}_{i,r}^{AB}\}_{i=1}^6, \mathbf{I}\{\widehat{\mathcal{W}}_{i,r}^{AB}\}_{i=1}^6]) \tag{5}$$

To encourage sparsity in the generated masks, we introduce an additional loss term $\mathcal{L}_m$, defined as the sum of all masks. We employ λ_m as the weight of $\mathcal{L}_m$ to control the sparsity of the representation.

Following the removal of unnecessary representations through masking, we adopt a compression strategy akin to the one employed in masked wavelet NeRF [42], originally designed for static scenes, to compress the sparse representation and masks that identify non-zero elements. The process involves converting the binary mask values to 8-bit unsigned integers and subsequently applying run-length encoding (RLE). Finally, the Huffman encoding algorithm is employed on the RLE-encoded streams to efficiently map values with a high probability to shorter bits.

3.4. Optimization

We leverage our proposed direction-aware representation to effectively represent 3D dynamic scenes. The model is then optimized through a photometric loss function, which measures the difference between rendered images and target images. For a given point (x, y, z, t), its opacity and appearance features are represented by six real and six imaginary direction-aware representation. The final color is regressed through a small multi-layer perceptron (MLP), taking the appearance feature and view direction as inputs. Utilizing the point's opacities and colors, images are obtained through volumetric rendering. The overall loss is expressed as:

$$\mathcal{L} = \frac{1}{|\mathcal{R}|} \sum_{r \in \mathcal{R}} ||\mathbf{C}(r) - \widehat{\mathbf{C}}|(r)||_2^2 + \lambda_{reg}\mathcal{L}_{reg} + \lambda_m \mathcal{L}_m, \tag{6}$$

with $\mathcal{L}_{reg}$, λ_{reg} and $\mathcal{L}_m$, λ_m the regularization loss and mask loss with respective weights, $\mathcal{R}$ the set of rays, and $\mathbf{C}(r)$, $\widehat{\mathbf{C}}(r)$ the rendered and ground-truth ray colors.

Regularization. For the regularization term, we employ the total variational (TV) loss on the direction-aware representation to enforce spatio-temporal continuity.

Training Strategy. We employ the same coarse-to-fine training strategy as in [7, 10, 67], where the resolution of grids progressively increases during training. This strategy not only accelerates the training process but also imparts an implicit regularization on nearby grids.

Emptiness Voxel. We maintain a small 3D voxel representation that indicates the emptiness of specific regions in the scene, allowing us to skip points located in empty regions. Given the typically large number of empty regions, this strategy significantly aids in acceleration. To generate this voxel, we evaluate the opacities of points across different time steps and aggregate them into a single voxel by retaining the maximum opacities. While preserving multiple voxels for distinct

	Model	Steps	PSNR↑	D-SSIM↓	LPIPS↓	Training Time↓	Model Size (MB) ↓
flame-salmon scene	Neural Volumes [28]	-	22.800	0.062	0.295	-	
	LLFF [32]	-	23.239	0.076	0.235	-	-
	NeRF-T [22]	-	28.449	0.023	0.100	-	-
	DyNeRF [22]	650k	29.581	0.020	0.099	1,344h	**28**
	HexPlane [7]	650k	29.470	0.018	**0.078**	12h	252
	HexPlane [7]	100k	29.263	0.020	0.097	**2h**	252
	DaReNeRF-S	100k	30.224	0.015	0.089	5h	244
	DaReNeRF	100k	**30.441**	**0.012**	0.084	4.5h	1,210
all scenes (average)	NeRFPlayer [48]	-	30.690	0.034	0.111	6h	-
	HyperReel [1]	-	31.100	0.036	0.096	9h	-
	HexPlane [7]	650k	31.705	0.014	0.075	12h	252
	HexPlane [7]	100k	31.569	0.016	0.089	2h	252
	K-Planes-explicit [14]	120k	30.880	-	-	3.7h	580
	K-Planes-hybrid	90k	31.630	-	-	1.8h	310
	Mix Voxels-L [51]	25k	31.340	0.019	0.096	**1.3h**	500
	Mix Voxels-X [51]	50k	31.730	0.015	**0.064**	5h	500
	4D-GS [57]	-	31.020	-	0.150	2h	**145**
	DaReNeRF-S	100k	32.102	0.014	0.087	5h	244
	DaReNeRF	100k	**32.258**	**0.012**	0.084	4.5h	1,210

Table 1. **Quantitative Comparison on Plenoptic Video Data.** We present results on synthesis quality and training time (measured in GPU hours). Following prior art, we provide both scene-specific performance (`flame-salmon` scene) and mean performance across all cases from their original papers.

time intervals could potentially enhance efficiency, for the sake of simplicity, we opt to keep only one voxel [7].

4. Experiments

We evaluate the performance of our proposed direction-aware representation on both dynamic and static scenes, conducting a thorough comparison with prior art. Additionally, we delve into the advantages of our direction-aware representation through ablation studies, showcasing its robustness in handling both dynamic and static scenes.

4.1. Novel View Synthesis of Dynamic Scenes

For dynamic scenes, we employ two distinct datasets with varying settings. Each dataset presents its own challenges, effectively addressed by our direction-aware representation.
Plenoptic Video Dataset [22] is a real-world dataset captured by a multi-view camera system using 21 GoPro cameras at a resolution of 2028×2704 and a frame rate of 30 frames per second. Each scene consists of 19 synchronized, 10-second videos, with 18 videos designated for training and one for evaluation. This dataset serves as an ideal testbed to assess the representation ability, featuring complex and challenging dynamic content, including highly specular, translucent, and transparent objects, topology changes, moving self-casting shadows, fire flames, strong view-dependent effects for moving objects, and more.

For a fair and direct comparison, we adhere to the same training and evaluation protocols as DyNeRF [22]. Our model is trained on a single A100 GPU, utilizing a batch size of 4,096. We adopt identical importance sampling strategies and hierarchical training techniques as DyNeRF, employing a spatial grid size of 512 and a temporal grid size of 300. The scene is placed under the normalized device coordinates (NDC) setting, consistent with the approach outlined in [33].

Quantitative compression results with state-of-the-art methods are presented in Table 1. We utilize measurement metrics PSNR, structure dissimilarity index measure (DSSIM) [43], and perception quality measure LPIPS [69] to conduct a comprehensive evaluation. As demonstrated in Table 1, leveraging the proposed direction-aware representation, both regular and sparse DaReNeRF achieve promising results compared to the most recent state-of-the-art , with analogous training time. This more ideal trade-off between performance and computational requirements, compared to prior art, is also illustrated in Figure 1.b, computed over Plenoptic data. Figure 4 presents some novel-view results on the Plenoptic dataset. Four small patches, each containing detailed texture information, are selected for comparison. DaReNeRF, equipped with the proposed direction-aware representation, excels in reconstructing moving objects (*e.g.*, dog and firing gun) and capturing better texture details (*e.g.*, hair and metal rings on the apron).

D-NeRF Dataset [40] is a monocular video dataset with 360° observations of synthetic objects. Dynamic 3D reconstruction from monocular video poses challenges as only one observation is available for each timestamp. State-of-the-art

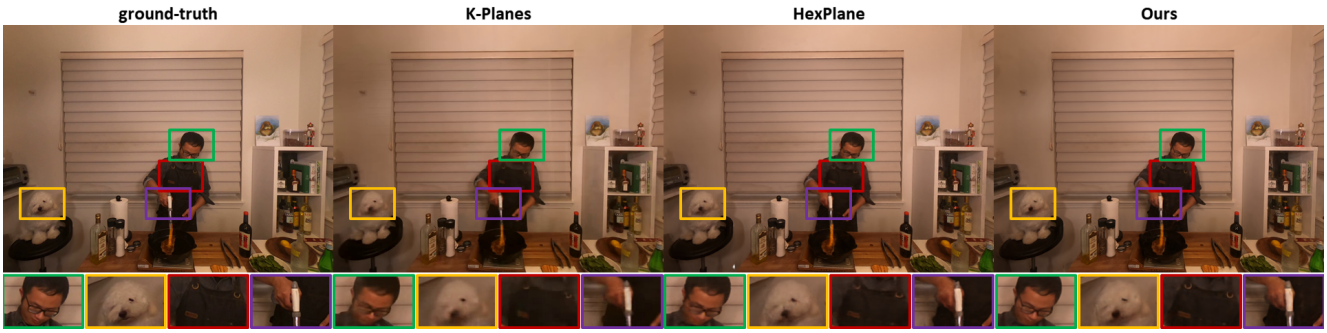

Figure 4. **Visual Comparison on Dynamic Scenes (Plenoptic Data).** K-Planes and HexPlane are concurrent decomposition-based methods. As shown in the four zoomed-in patches, our method better reconstruct fine details and captures motion. Please refer to the supplementary material to see the figure animated.

Table 2. **Quantitative Study on D-NeRF Data.** Without the topological constraints of using deformation fields, DaReNeRF outperforms even some deformation-based methods.

Model	Deform.	PSNR↑	SSIM↑	LPIPS↓
D-NeRF [40]	✓	30.50	0.95	0.07
TiNeuVox-S [12]	✓	30.75	0.96	0.07
TiNeuVox-B [12]	✓	32.67	0.97	0.04
4D-GS [57]	✓	**33.30**	**0.98**	**0.03**
T-NeRF [40]	-	29.51	0.95	0.08
HexPlane [7]	-	31.04	**0.97**	0.04
K-Planes [14]	-	31.05	**0.97**	-
DaReNeRF-S	-	31.82	**0.97**	**0.03**
DaReNeRF	-	**31.95**	**0.97**	**0.03**

methods for monocular video typically incorporate a deformation field. However, the underlying assumption is that the scenes undergo no topological changes, making them less effective for real-world cases (*e.g.*, Plenoptic dataset). Table 2 reports the rendering quality of different methods with and without deformation fields on the D-NeRF data, DaReNeRF outperforms all non-deformation methods, as well as some deformation methods, *e.g.* D-NeRF and TiNeuVox-S [12]. The superiority of our solution on topologically-changing scenes is further highlighted in annex.

4.2. Novel View Synthesis of Static Scenes

For static scenes, we test our proposed direction-aware representation on NeRF synthetic [33], Neural Sparse Voxel Fields (NSVF) [25] and LLFF [32] datasets. We use TensoRF-192 as baseline and apply our proposed representation. We report the performance on these three datasets in Tables 3, 4, and 5 respectively.

Across these three static datasets, our direction-aware representation outperforms most compression-based NeRF models with model sizes ranging from 8 to 14MB. While our method's model size is larger than DWT-based solutions, it achieves comparable sparsity. For instance, with $\lambda_m = 2.5 \times 10^{-11}$, its *sparsity* reaches 94%, closely aligned with the 97% reported in the masked wavelet NeRF [42] paper. Notably, with similar sparsity, our direction-aware method exhibits PSNR improvements of 0.47, 1.57, and 0.60 over DWT-based methods on the three static datasets.

Figure 5 highlights the qualitative differences between DWT-based solutions and our proposed direction-aware method. In static scenes, our solution excels in reconstructing texture details compared to DWT representation, which is less sensitive to lines and edges patterns due to shift variance and direction ambiguity.

Table 3. **Quantitative Comparison on NeRF Synth.**, with models designed for different bit-precisions (* denotes a model quantized post-training; numbers in brackets denote grid resolutions).

Precision	Method	Size (MB)	PSNR ↑
32-bit	KiloNeRF [41]	≤ 100	31.00
32-bit	CCNeRF (CP) [49]	4.4	30.55
8-bit*	NeRF [33]	1.25	31.52
8-bit	cNeRF [5]	**0.70**	30.49
8-bit	PREF [18]	9.88	31.56
8-bit*	VM-192 [10]	17.93	**32.91**
8-bit*	VM-192 (300) + DWT [42]	0.83	31.95
8-bit*	VM-192 (300) + Ours	8.91	32.42

Table 4. **Quantitative Comparison on NSVF** (static scenes).

Bit Precision	Model	Size (MB)	PSNR ↑
32-bit	KiloNeRF [41]	≤ 100	33.37
8-bit*	VM-192 [47]	17.77	36.11
8-bit*	VM-48 [10]	4.53	34.95
8-bit*	CP-384 [10]	**0.72**	33.92
8-bit*	VM-192 (300) + DWT [42]	0.87	34.67
8-bit*	VM-192 (300) + Ours	8.98	**36.24**

Table 5. **Quantitative Comparison on LLFF** (static scenes).

Bit Precision	Model	Size(MB)	PSNR ↑
8-bit	cNeRF [5]	0.96	26.15
8-bit*	PREF [47]	9.34	24.50
8-bit*	VM-96 [10]	44.72	**26.66**
8-bit*	VM-48 [10]	22.40	26.46
8-bit*	CP-384 [10]	**0.64**	25.51
8-bit*	VM-96 (640) + DWT [42]	1.34	25.88
8-bit*	VM-96 (640) + Ours	13.67	26.48

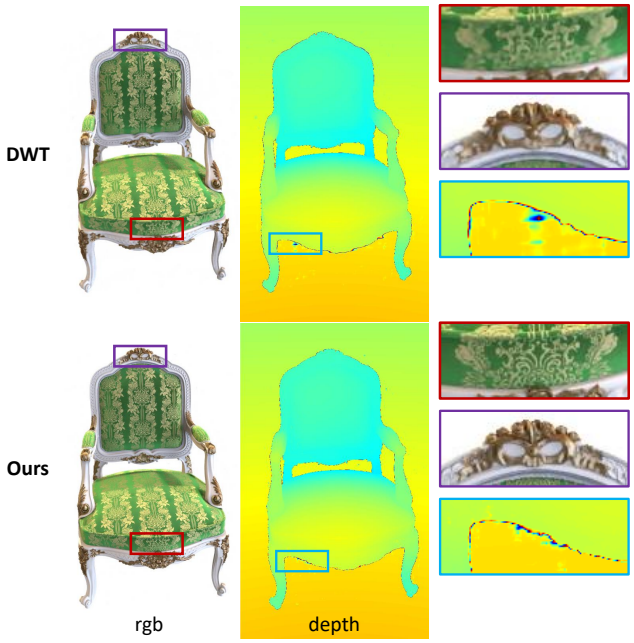

Figure 5. **Visual Comparison of Static Scenes on NSVF Data.** Two representative patches are selected for closer inspection. Our method, free from the DWT limitations of shift variance and direction ambiguity, achieves superior texture reconstruction performance.

4.3. Ablations

Wavelet Function. We analyze the impact of different wavelet functions on reconstruction quality, aiming to facilitate a comparison between our direction-aware representation and DWT wavelet. The evaluation is conducted on NSVF data [25], where several complex wavelet functions with the approximate half-sample delay property—Antonini, LeGall, and two Near Symmetric filter banks (Near Symmetric A and Near Symmetric B)—are selected for comparison. Table 6 reveals that the choice of different wavelets has minimal effect on reconstruction quality. Even the worst-performing wavelet function outperforms the discrete wavelet transform, underscoring the advantages of our direction-aware representation.

Sparsity Analysis. We evaluate the sparsity of our direction-aware representation by varying the sparsity level using different λ_m values on the NSVF dataset. As depicted in Table 7, our direction-aware representation consistently achieves over 99% sparsity. This remarkable sparsity, coupled with a model size of approximately 1MB, demonstrates the efficiency of our method in modeling static scenes while outperforming state-of-the-art sparse representation methods.

Wavelet Levels. We investigated the impact of scene reconstruction performance across different wavelet levels, and the results are presented in supplementary material. We observed that increasing the wavelet level did not lead to significant performance improvements. Conversely, we noted a substantial increase in both training time and model size with the increment of wavelet level. As a result, throughout all experiments, we consistently set the wavelet level to 1.

Table 6. **Impact of Wavelet Transform Type/Function**, on reconstruction performance, evaluated on NSVF data..

Wavelet Type	Wavelet Function	PSNR ↑
DWT	Haar	34.61
	Coiflets 1	34.56
	biorthogonal 4.4	**34.67**
	Daubechies 4	34.44
DTCWT	Antonini	36.10
	LeGall	36.14
	Near Symmetric A	**36.24**
	Near Symmetric B	36.17

Table 7. **Sparsity Analysis of Direction-Aware Representation**, evaluated on NVSF data.

λ_m	Sparsity ↑	Model Size (MB) ↓	PSNR ↑
1.0×10^{-10}	**99.2%**	**1.16 MB**	35.36
5.0×10^{-11}	97.3%	3.18 MB	35.81
2.5×10^{-11}	94.2%	8.98 MB	**36.24**
0	-	135 MB	36.34

5. Discussion

Limitations. Our method is limited for scenarios with extremely sparse observations, as seen in D-NeRF-like datasets. DaReNeRF does not incorporate a deformation field into the model, lacking a robust information-sharing mechanism to learn 3D structures from very sparse views. Another limitation of our proposed direction-aware representation is its lower compactness compared to DWT representation, preventing DaReNeRF from achieving extremely small model sizes, such as less than 1MB on static scene. Exploring more compact methods to construct direction-aware representations would be an interesting direction for future research.

Conclusion. We introduced a novel direction-aware representation capable of effectively capturing information from six different directions. The shift-invariant and direction-selective nature of our proposed representation enables the high-fidelity reconstruction of challenging dynamic scenes without the need for prior knowledge about the scene dynamics. Despite introducing some storage redundancy, we mitigate this by incorporating trainable masks for both static and dynamic scenes, resulting in a model size comparable to recent methods. We believe that this simple yet effective representation has the potential to simplify and streamline dynamic NeRFs, providing a more accessible and efficient solution for complex scene modeling.

References

[1] Benjamin Attal, Jia-Bin Huang, Christian Richardt, Michael Zollhoefer, Johannes Kopf, Matthew O'Toole, and Changil Kim. Hyperreel: High-fidelity 6-dof video with ray-conditioned sampling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16610–16620, 2023. 2, 6

[2] Jonathan T Barron, Ben Mildenhall, Matthew Tancik, Peter Hedman, Ricardo Martin-Brualla, and Pratul P Srinivasan. Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5855–5864, 2021. 2

[3] Jonathan T Barron, Ben Mildenhall, Dor Verbin, Pratul P Srinivasan, and Peter Hedman. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5470–5479, 2022.

[4] Wenjing Bian, Zirui Wang, Kejie Li, Jia-Wang Bian, and Victor Adrian Prisacariu. Nope-nerf: Optimising neural radiance field with no pose prior. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4160–4169, 2023. 2

[5] Thomas Bird, Johannes Ballé, Saurabh Singh, and Philip A Chou. 3d scene compression through entropy penalized neural representation functions. In *2021 Picture Coding Symposium (PCS)*, pages 1–5. IEEE, 2021. 7

[6] Andrew P Bradley. Shift-invariance in the discrete wavelet transform. *Proceedings of VIIth Digital Image Computing: Techniques and Applications. Sydney*, 2003. 2

[7] Ang Cao and Justin Johnson. Hexplane: A fast representation for dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 130–141, 2023. 2, 3, 5, 6, 7

[8] Eric R Chan, Marco Monteiro, Petr Kellnhofer, Jiajun Wu, and Gordon Wetzstein. pi-gan: Periodic implicit generative adversarial networks for 3d-aware image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5799–5809, 2021. 2

[9] Eric R Chan, Connor Z Lin, Matthew A Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas J Guibas, Jonathan Tremblay, Sameh Khamis, et al. Efficient geometry-aware 3d generative adversarial networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16123–16133, 2022. 2

[10] Anpei Chen, Zexiang Xu, Andreas Geiger, Jingyi Yu, and Hao Su. Tensorf: Tensorial radiance fields. In *European Conference on Computer Vision*, pages 333–350. Springer, 2022. 2, 5, 7

[11] Congyue Deng, Chiyu Jiang, Charles R Qi, Xinchen Yan, Yin Zhou, Leonidas Guibas, Dragomir Anguelov, et al. Nerdi: Single-view nerf synthesis with language-guided diffusion as general image priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20637–20647, 2023. 2

[12] Jiemin Fang, Taoran Yi, Xinggang Wang, Lingxi Xie, Xiaopeng Zhang, Wenyu Liu, Matthias Nießner, and Qi Tian. Fast dynamic radiance fields with time-aware neural voxels. In *SIGGRAPH Asia 2022 Conference Papers*, pages 1–9, 2022. 7

[13] Sara Fridovich-Keil, Alex Yu, Matthew Tancik, Qinhong Chen, Benjamin Recht, and Angjoo Kanazawa. Plenoxels: Radiance fields without neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5501–5510, 2022. 2

[14] Sara Fridovich-Keil, Giacomo Meanti, Frederik Rahbæk Warburg, Benjamin Recht, and Angjoo Kanazawa. K-planes: Explicit radiance fields in space, time, and appearance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12479–12488, 2023. 2, 3, 6, 7

[15] Chen Gao, Ayush Saraf, Johannes Kopf, and Jia-Bin Huang. Dynamic view synthesis from dynamic monocular video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5712–5721, 2021. 1

[16] Xiang Guo, Jiadai Sun, Yuchao Dai, Guanying Chen, Xiaoqing Ye, Xiao Tan, Errui Ding, Yumeng Zhang, and Jingdong Wang. Forward flow for novel view synthesis of dynamic scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 16022–16033, 2023. 3

[17] Benran Hu, Junkai Huang, Yichen Liu, Yu-Wing Tai, and Chi-Keung Tang. Nerf-rpn: A general framework for object detection in nerfs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23528–23538, 2023. 2

[18] Binbin Huang, Xinhao Yan, Anpei Chen, Shenghua Gao, and Jingyi Yu. Pref: Phasorial embedding fields for compact neural representations. *arXiv preprint arXiv:2205.13524*, 2022. 7

[19] Erik Johnson, Marc Habermann, Soshi Shimada, Vladislav Golyanik, and Christian Theobalt. Unbiased 4d: Monocular 4d reconstruction with a neural deformation model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6597–6606, 2023. 3

[20] Nick Kingsbury. Image processing with complex wavelets. *Philosophical Transactions of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences*, 357(1760):2543–2560, 1999. 2

[21] Sosuke Kobayashi, Eiichi Matsumoto, and Vincent Sitzmann. Decomposing nerf for editing via feature field distillation. *Advances in Neural Information Processing Systems*, 35:23311–23330, 2022. 2

[22] Tianye Li, Mira Slavcheva, Michael Zollhoefer, Simon Green, Christoph Lassner, Changil Kim, Tanner Schmidt, Steven Lovegrove, Michael Goesele, Richard Newcombe, et al. Neural 3d video synthesis from multi-view video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5521–5531, 2022. 1, 6

[23] Zhengqi Li, Simon Niklaus, Noah Snavely, and Oliver Wang. Neural scene flow fields for space-time view synthesis of dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6498–6508, 2021. 1, 2

[24] Zhengqi Li, Qianqian Wang, Forrester Cole, Richard Tucker, and Noah Snavely. Dynibar: Neural dynamic image-based

rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4273–4284, 2023. 2, 3

[25] Lingjie Liu, Jiatao Gu, Kyaw Zaw Lin, Tat-Seng Chua, and Christian Theobalt. Neural sparse voxel fields. *Advances in Neural Information Processing Systems*, 33:15651–15663, 2020. 7, 8

[26] Yichen Liu, Benran Hu, Junkai Huang, Yu-Wing Tai, and Chi-Keung Tang. Instance neural radiance field. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 787–796, 2023. 2

[27] Yu-Lun Liu, Chen Gao, Andreas Meuleman, Hung-Yu Tseng, Ayush Saraf, Changil Kim, Yung-Yu Chuang, Johannes Kopf, and Jia-Bin Huang. Robust dynamic radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13–23, 2023. 1, 2, 3

[28] Stephen Lombardi, Tomas Simon, Jason Saragih, Gabriel Schwartz, Andreas Lehrmann, and Yaser Sheikh. Neural volumes: Learning dynamic renderable volumes from images. *arXiv preprint arXiv:1906.07751*, 2019. 6

[29] Weng Fei Low and Gim Hee Lee. Robust e-nerf: Nerf from sparse & noisy events under non-uniform motion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 18335–18346, 2023. 2

[30] Ricardo Martin-Brualla, Noha Radwan, Mehdi SM Sajjadi, Jonathan T Barron, Alexey Dosovitskiy, and Daniel Duckworth. Nerf in the wild: Neural radiance fields for unconstrained photo collections. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7210–7219, 2021. 2

[31] Gal Metzer, Elad Richardson, Or Patashnik, Raja Giryes, and Daniel Cohen-Or. Latent-nerf for shape-guided generation of 3d shapes and textures. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12663–12673, 2023. 2

[32] 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 Transactions on Graphics (TOG)*, 38(4):1–14, 2019. 6, 7

[33] 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. *Communications of the ACM*, 65(1):99–106, 2021. 2, 6, 7

[34] Ben Mildenhall, Peter Hedman, Ricardo Martin-Brualla, Pratul P Srinivasan, and Jonathan T Barron. Nerf in the dark: High dynamic range view synthesis from noisy raw images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16190–16199, 2022. 2

[35] Ashkan Mirzaei, Tristan Aumentado-Armstrong, Konstantinos G Derpanis, Jonathan Kelly, Marcus A Brubaker, Igor Gilitschenski, and Alex Levinshtein. Spin-nerf: Multiview segmentation and perceptual inpainting with neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20669–20679, 2023. 2

[36] Arthur Moreau, Nathan Piasco, Dzmitry Tsishkou, Bogdan Stanciulescu, and Arnaud de La Fortelle. Lens: Localization enhanced by nerf synthesis. In *Conference on Robot Learning*, pages 1347–1356. PMLR, 2022. 1

[37] Michael Niemeyer, Jonathan T Barron, Ben Mildenhall, Mehdi SM Sajjadi, Andreas Geiger, and Noha Radwan. Regnerf: Regularizing neural radiance fields for view synthesis from sparse inputs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5480–5490, 2022. 2

[38] Keunhong Park, Utkarsh Sinha, Jonathan T Barron, Sofien Bouaziz, Dan B Goldman, Steven M Seitz, and Ricardo Martin-Brualla. Nerfies: Deformable neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5865–5874, 2021. 1

[39] Keunhong Park, Utkarsh Sinha, Peter Hedman, Jonathan T Barron, Sofien Bouaziz, Dan B Goldman, Ricardo Martin-Brualla, and Steven M Seitz. Hypernerf: A higher-dimensional representation for topologically varying neural radiance fields. *arXiv preprint arXiv:2106.13228*, 2021.

[40] Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. D-nerf: Neural radiance fields for dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10318–10327, 2021. 1, 3, 6, 7

[41] Christian Reiser, Songyou Peng, Yiyi Liao, and Andreas Geiger. Kilonerf: Speeding up neural radiance fields with thousands of tiny mlps. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14335–14345, 2021. 7

[42] Daniel Rho, Byeonghyeon Lee, Seungtae Nam, Joo Chan Lee, Jong Hwan Ko, and Eunbyung Park. Masked wavelet representation for compact neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20680–20690, 2023. 2, 3, 5, 7

[43] Umme Sara, Morium Akter, and Mohammad Shorif Uddin. Image quality assessment through fsim, ssim, mse and psnr—a comparative study. *Journal of Computer and Communications*, 7(3):8–18, 2019. 6

[44] Vishwanath Saragadam, Daniel LeJeune, Jasper Tan, Guha Balakrishnan, Ashok Veeraraghavan, and Richard G Baraniuk. Wire: Wavelet implicit neural representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18507–18516, 2023. 3

[45] Ivan W Selesnick, Richard G Baraniuk, and Nick C Kingsbury. The dual-tree complex wavelet transform. *IEEE signal processing magazine*, 22(6):123–151, 2005. 2, 3

[46] Ruizhi Shao, Zerong Zheng, Hanzhang Tu, Boning Liu, Hongwen Zhang, and Yebin Liu. Tensor4d: Efficient neural 4d decomposition for high-fidelity dynamic reconstruction and rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16632–16642, 2023. 2, 3

[47] Liangchen Song, Xuan Gong, Benjamin Planche, Meng Zheng, David Doermann, Junsong Yuan, Terrence Chen, and Ziyan Wu. Pref: Predictability regularized neural motion fields. In *European Conference on Computer Vision*, pages 664–681. Springer, 2022. 3, 7

[48] Liangchen Song, Anpei Chen, Zhong Li, Zhang Chen, Lele Chen, Junsong Yuan, Yi Xu, and Andreas Geiger. Nerfplayer: A streamable dynamic scene representation with decomposed neural radiance fields. *IEEE Transactions on Visualization and Computer Graphics*, 29(5):2732–2742, 2023. 6

[49] Jiaxiang Tang, Xiaokang Chen, Jingbo Wang, and Gang Zeng. Compressible-composable nerf via rank-residual decomposition. *Advances in Neural Information Processing Systems*, 35:14798–14809, 2022. 7

[50] Chaoyang Wang, Lachlan Ewen MacDonald, Laszlo A Jeni, and Simon Lucey. Flow supervision for deformable nerf. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21128–21137, 2023. 3

[51] Feng Wang, Sinan Tan, Xinghang Li, Zeyue Tian, Yafei Song, and Huaping Liu. Mixed neural voxels for fast multi-view video synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19706–19716, 2023. 6

[52] Liao Wang, Jiakai Zhang, Xinhang Liu, Fuqiang Zhao, Yanshun Zhang, Yingliang Zhang, Minye Wu, Jingyi Yu, and Lan Xu. Fourier plenoctrees for dynamic radiance field rendering in real-time. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13524–13534, 2022. 2

[53] Peng Wang, Yuan Liu, Zhaoxi Chen, Lingjie Liu, Ziwei Liu, Taku Komura, Christian Theobalt, and Wenping Wang. F2-nerf: Fast neural radiance field training with free camera trajectories. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4150–4159, 2023. 2

[54] Qianqian Wang, Zhicheng Wang, Kyle Genova, Pratul P Srinivasan, Howard Zhou, Jonathan T Barron, Ricardo Martin-Brualla, Noah Snavely, and Thomas Funkhouser. Ibrnet: Learning multi-view image-based rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4690–4699, 2021. 1

[55] Yuehao Wang, Yonghao Long, Siu Hin Fan, and Qi Dou. Neural rendering for stereo 3d reconstruction of deformable tissues in robotic surgery. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 431–441. Springer, 2022. 1

[56] Yuxin Wang, Wayne Wu, and Dan Xu. Learning unified decompositional and compositional nerf for editable novel view synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 18247–18256, 2023. 2

[57] Guanjun Wu, Taoran Yi, Jiemin Fang, Lingxi Xie, Xiaopeng Zhang, Wei Wei, Wenyu Liu, Qi Tian, and Xinggang Wang. 4d gaussian splatting for real-time dynamic scene rendering. *arXiv preprint arXiv:2310.08528*, 2023. 6, 7

[58] Zhijie Wu, Yuhe Jin, and Kwang Moo Yi. Neural fourier filter bank. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14153–14163, 2023. 2

[59] Magdalena Wysocki, Mohammad Farid Azampour, Christine Eilers, Benjamin Busam, Mehrdad Salehi, and Nassir Navab. Ultra-nerf: Neural radiance fields for ultrasound imaging. *arXiv preprint arXiv:2301.10520*, 2023. 1

[60] Jiaxin Xie, Hao Ouyang, Jingtan Piao, Chenyang Lei, and Qifeng Chen. High-fidelity 3d gan inversion by pseudo-multi-view optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 321–331, 2023. 2

[61] Yiming Xie, Huaizu Jiang, Georgia Gkioxari, and Julian Straub. Pixel-aligned recurrent queries for multi-view 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 18370–18380, 2023. 2

[62] Chenfeng Xu, Bichen Wu, Ji Hou, Sam Tsai, Ruilong Li, Jialiang Wang, Wei Zhan, Zijian He, Peter Vajda, Kurt Keutzer, et al. Nerf-det: Learning geometry-aware volumetric representation for multi-view 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23320–23330, 2023.

[63] Junkai Xu, Liang Peng, Haoran Cheng, Hao Li, Wei Qian, Ke Li, Wenxiao Wang, and Deng Cai. Mononerd: Nerf-like representations for monocular 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6814–6824, 2023. 2

[64] Muyu Xu, Fangneng Zhan, Jiahui Zhang, Yingchen Yu, Xiaoqin Zhang, Christian Theobalt, Ling Shao, and Shijian Lu. Wavenerf: Wavelet-based generalizable neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 18195–18204, 2023. 2, 3

[65] Zhiwen Yan, Chen Li, and Gim Hee Lee. Nerf-ds: Neural radiance fields for dynamic specular objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8285–8295, 2023. 2, 3

[66] Jiawei Yang, Marco Pavone, and Yue Wang. Freenerf: Improving few-shot neural rendering with free frequency regularization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8254–8263, 2023. 2

[67] Alex Yu, Ruilong Li, Matthew Tancik, Hao Li, Ren Ng, and Angjoo Kanazawa. Plenoctrees for real-time rendering of neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5752–5761, 2021. 5

[68] Junzhe Zhang, Yushi Lan, Shuai Yang, Fangzhou Hong, Quan Wang, Chai Kiat Yeo, Ziwei Liu, and Chen Change Loy. Deformtoon3d: Deformable neural radiance fields for 3d toonification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9144–9154, 2023. 3

[69] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 6

[70] Xiaoshuai Zhang, Sai Bi, Kalyan Sunkavalli, Hao Su, and Zexiang Xu. Nerfusion: Fusing radiance fields for large-scale scene reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5449–5458, 2022. 2

[71] Zihan Zhu, Songyou Peng, Viktor Larsson, Weiwei Xu, Hujun Bao, Zhaopeng Cui, Martin R Oswald, and Marc Pollefeys.

Nice-slam: Neural implicit scalable encoding for slam. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12786–12796, 2022. 2