DMiT: Deformable Mipmapped Tri-Plane Representation for Dynamic Scenes

Jing-Wen Yang[1,2], Jia-Mu Sun[1,2], Yong-Liang Yang[3], Jie Yang[1], Ying Shan[4], Yan-Pei Cao[5,4], and Lin Gao (✉)[1,2]

[1] Beijing Key Laboratory of Mobile Computing and Pervasive Device, Institute of Computing Technology, Chinese Academy of Sciences, China
[2] University of Chinese Academy of Sciences, China
[3] Department of Computer Science, University of Bath, United Kingdom
[4] Tencent AI Lab, China [5] VAST, China

Abstract. Neural Radiance Fields (NeRF) have achieved remarkable progress on dynamic scenes with deformable objects. Nonetheless, most previous works required multi-view inputs or long training time (several hours), making it hard to apply them for real-world scenarios. Recent works dedicated to addressing blurry artifacts may fail to predict stable and accurate deformation while keeping high-frequency details when rendering at various resolutions. To this end, we introduce a novel framework **DMiT** (**D**eformable **Mi**pmapped **T**ri-Plane) that adopts the mipmaps to render dynamic scenes at various resolutions from novel views. With the help of hierarchical mipmapped tri-planes, we incorporate an MLP to effectively predict a mapping between the observation space and the canonical space, enabling not only high-fidelity dynamic scene rendering but also high-performance training and inference. Moreover, a training scheme for joint geometry and deformation refinement is designed for canonical regularization to reconstruct high-quality geometries. Extensive experiments on both synthetic and real dynamic scenes demonstrate the efficacy and efficiency of our method.

Keywords: Neural radiance fields · Mipmapping · Dynamic scene reconstruction

1 Introduction

High-quality 3D reconstruction and novel view synthesis from 2D images play a significant role in computer vision and computer graphics [11,46]. While a growing body of research works focuses on static scenes, dynamic scenes with objects under different types of movements are also ubiquitous in practice. However, it is extremely challenging to accurately reconstruct dynamic scenes from images, since the observed complex dynamics can easily introduce uncertainty and ambiguity in motion. To overcome such difficulties, previous approaches often involve capture systems equipped with multiple calibrated cameras [21] or depth sensors [3] to provide more priors and regularizations.

✉ Corresponding author is Lin Gao(gaolin@ict.ac.cn).

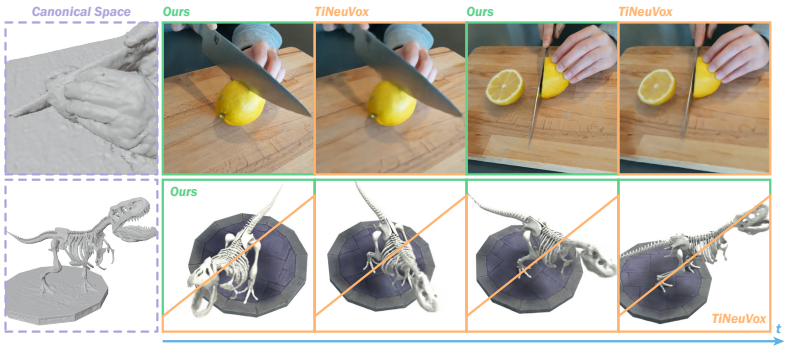


Fig. 1: We propose a novel framework **DMiT** that can generate high-fidelity results for time interpolation (top row) and novel view synthesis (bottom row) tasks on dynamic scenes from both synthetic and real-world datasets. We also show the restored canonical geometry with faithful details on the left as well as the comparison with TiNeuVox [15] to demonstrate the effectiveness of our work.

In recent years, the rapid development of 3D reconstruction has been witnessed due to the advance of neural rendering, particularly Neural Radiance Fields (NeRF) [38]. It has been proved that neural implicit representation can largely benefit high-fidelity reconstruction of static scenes and 3D generation with geometry and appearance details [6, 50, 54, 68]. For dynamic scenes, D-NeRF [45] incorporates a deformation field to map dynamic points sampled from an observation space to their static counterparts in the canonical space. Another approach simply formulates dynamics using a NeRF conditioned on time embedding [31, 61]. Nevertheless, both approaches suffer from slow convergence and loss of geometry/appearance details especially under varying viewing distances. This can be attributed to either the deformation network being prone to overfitting or the time-conditioned NeRF having limited correspondence between frames. Recently, 3D Gaussian Splatting [25] utilized explicit Gaussian kernels in 3D space along with differentiable rasterization, achieving real-time rendering speed while preserving high-quality rendering results. Based on this prominent work, [23,33,58,63,64] excelled at common dynamic datasets in speed against time-consuming queries of MLPs in NeRF-based methods. Nonetheless, 3DGS and other NeRF-based works combined with explicit architecture [8, 16] also suffer from aliasing artifacts, which remain unsolved in the dynamic setting.

To address the above challenges, we introduce a novel NeRF-based framework, called **DMiT**, to model dynamic scenes based on a **D**eformable **Mi**pmapped **T**ri-Plane (see Fig. 1) which exploits dynamic information across scales and distances. First, we employ the 'canonical+deformation' paradigm to represent dynamic scenes, benefiting from the ease of dynamics modeled with a formal canonical geometry. Moreover, to achieve efficient anti-aliased rendering, we introduce mipmaps [6] into compact tri-plane representation [9]. Specifically, we build a hierarchy of tri-planes with different resolutions via pre-filtering oper-

ations. Thanks to the compactness and efficiency of the Tri-Mip encoding, our method achieves much faster speed in training and inference. Additionally, the mipmap techniques and area sampling strategy play an important role in enhancing the details of rendered dynamic scenes. To demonstrate the efficiency and effectiveness of our developed method, extensive experiments are conducted on both synthetic and real-world datasets. The results have shown the competitiveness of the proposed DMiT against previous SOTA methods quantitatively and qualitatively. Apart from commonly used single-scale datasets, experiments on multi-scale dynamic datasets also proved the effectiveness of our method.

To summarize, our main contributions are: 1) We propose a novel Deformable Mipmapped Tri-Plane representation, called **DMiT**, to reduce the blurriness and aliasing effects in dynamic scenes, which organizes the tri-planes with different resolutions in a hierarchical fashion. 2) We propose a joint geometry and deformation refinement procedure to improve the fidelity of deformable canonical space that incorporates our new representation, achieving a better decomposition of geometry and motion. 3) Our method has achieved competitive reconstruction and rendering quality compared to SOTA methods both quantitatively and qualitatively on extensive datasets of both synthetic and real-world scenes.

2 Related Work

Dynamic Neural Radiance Fields. NeRF and its variants [6, 52, 65] have demonstrated the efficacy of modeling 3D static scenes from 2D photographs. A number of studies [41, 42, 45, 51] have sought to extend NeRF for dynamic scenes. The topic that has garnered the most interest is the monocular capture and reconstruction of dynamic scenes. This often encounters challenges in capturing complete scene motion due to occlusion and limited capture time, leading to inaccurate reconstruction and noticeable artifacts in novel view synthesis. Consequently, this often results in inaccurate reconstruction and obvious artifacts for the novel view synthesis task. To enhance the accuracy of dynamic reconstruction with finer details, numerous studies have incorporated additional prior knowledge like previous image-based 3D modeling techniques [67], including depth information [34, 61], optical flow estimation [5, 14], and 2D convolutional neural network (CNN) priors [26, 44]. Two distinct formulations of dynamics are proposed to accurately reconstruct dynamic scenes or non-rigid deformation scenes without requiring additional information. One is to decouple time from the original (stationary) NeRF and incorporate dynamics by using a common canonical space across all frames, as initially introduced by D-NeRF [45]. Several recent works [41, 42, 51], have proposed extensions to address the issue of deformation by introducing regulations and conditions, or by elevating the dimensional space to capture topological changes. Other attempts modeled dynamics through explicit representations, such as voxel [19, 32] and space-time tri-planes [48]. Nevertheless, deformation-based approaches typically employ two separate networks for geometry and deformation, leading to overfitting artifacts on dynamic scenes, *e.g.*, high-frequency floaters or fused canonical geometry with

deformed objects at different times (see Fig. 3). An alternative approach avoids treating time as a distinct dimension and instead includes temporal information directly into the vanilla NeRF. Some works [8, 15, 16] combined explicit representation with implicit NeRF, while others focused on efficient sampling strategy [4] or the decomposition of spatial-temporal fields [49]. The limitations of time-conditioned NeRFs lie in their limited capacity to capture high-frequency details, as they rely on a single neural network which often misses correlations across consecutive frames.

Recently, 3D Gaussian Splatting [25] has gained attention for its real-time rendering speed and high-fidelity rendering quality [60]. Several works have extended this technique to model dynamic scenes [23, 33, 58, 63, 64]. However, the point sample rendering technique suffers from holes and aliasing artifacts in rendering and fails to reconstruct high-quality geometry compared to NeRF-based methods, due to discontinuities in its representation.

Our pipeline is inspired by the previous 'canonical space + deformation' formulation. Unlike prior studies that employ MLPs for both canonical space and deformation, our approach leverages mipmapped tri-planes to facilitate efficient and impactful learning of the canonical space. To mitigate the presence of artifacts in the canonical space, we incorporate a joint refinement procedure.

Anti-aliasing in Neural Rendering. Anti-aliasing has been a substantial research subject in computer graphics, fundamental for various imaging and rendering applications. Existing approaches fall into two categories: super-sampling anti-aliasing and pre-filtering. One approach involves enhancing the sample rate to accurately reconstruct signals with higher frequencies [1, 13, 18, 20, 35, 57]. On the other hand, the latter approach achieves anti-aliasing by employing a filtering technique to attenuate high-frequency components. Consequently, the filtered signals can be effectively reconstructed without altering the sampling rate [2, 24, 27, 29, 40, 59]. Pre-filtering techniques are commonly favored in the domain of NeRF, mostly because of their relatively higher rendering speed in comparison to the more computationally intensive super-sampling approach. The rendering quality was greatly improved by the integrated positional encoding(IPE) method introduced by MipNeRF [6]. Subsequently, Zip-NeRF [7] extended MipNeRF with hash encoding and super-sampling technique to furthur mitigate aliasing effects. However, these methods faces challenges in achieving real-time inference speed. In order to enhance the efficiency of the anti-aliasing process without compromising the fidelity of the reconstruction, the Tri-MipRF [22] was proposed to optimize the rendering speed. The Tri-MipRF algorithm utilized the nvdiffrast [28] to achieve fast mipmapping, leading to convergence within a few minutes and enabling real-time rendering on consumer-level devices.

The above NeRF-based approaches mostly focused on static scenes but not dynamic scenes, where anti-aliasing plays a crucial role due to the increased visibility of aliasing artifacts during object motion. As the first method extending pre-filtering to dynamic NeRF representation, our DMiT can efficiently achieve anti-aliasing on dynamic scenes with reasonable training and rendering speed.

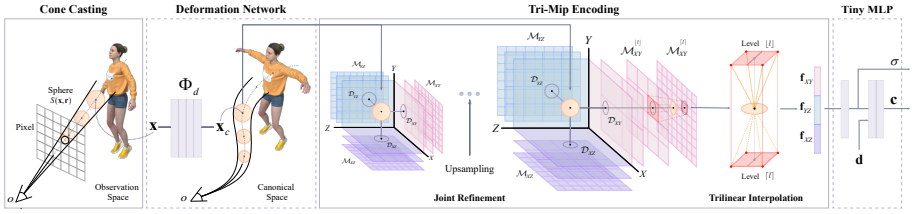


Fig. 2: Overview of our framework. Given images from dynamic scene (monocular or multi-view), we employ a deformed cone casting in observation space and sample multiple spheres which are then transformed to a common canonical space modeled with tri-miplanes. The interpolated tri-miplane features are fed into a tiny MLP for volume rendering. For clarity, we only show limited mipmap feature planes, while the full setup is provided in Sec. 4.1.

3 Methodology

Given a collection of images depicting dynamic scenes accompanied by precise timestamps and calibrated camera parameters obtained from structure-from-motion (SfM) [47] or Blender [12] generation, our work can render photorealistic novel views and interpolate between input frames as in Fig. 2.

In the remaining section, the preliminaries directly associated with our research will be introduced in Sec. 3.1. Next, the formulation of our deformable mipmapped tri-plane representation will be presented in Sec. 3.2. Our innovative joint deformation and geometry refinement will be detailed in Sec. 3.3. Finally, the optimization and regularization objectives will be elaborated in Sec. 3.4.

3.1 Preliminaries

Tri-Mip Radiance Field aaa For each pixel in one of the captured images of a static scene, there exists a ray denoted as $\mathbf{a}(t) = \mathbf{o} + t(\mathbf{p} - \mathbf{o}) = \mathbf{o} + t\mathbf{d}$, where $\mathbf{o}$, $\mathbf{p}$ and $\mathbf{d}$ represent the camera optical center, the pixel center and the direction of the ray respectively. In contrast to the vanilla NeRF approach, which simplifies one pixel as a point, Mip-NeRF [6] introduces a novel cone-casting strategy. This strategy involves emitting a cone, denoted as $\mathcal{C}$, with its central axis defined as $\mathbf{r}(t)$. As a result, Mip-NeRF is able to incorporate additional area information when formulating an image pixel. Building upon the work of EG3D [9], the Tri-MipRF [22] further relieves the burden of complex IPE of Mip-NeRF by integrating tri-planes with pre-filtering techniques, resulting in efficient training and superior rendering quality with reduced aliasing effects.

Specifically, a collection of inscribed spheres denoted as $\mathcal{S}$ is sampled within the cast cone. Each sphere $\mathcal{S}$ undergoes orthogonal projection onto the tri-planes $\mathcal{M}$, which serve as the fundamental mipmap features for the purpose of efficient pre-filtering. The computation of the adaptive mipmap level l is performed to determine the appropriate mipmap features for interpolation, based on the sample distance t along the cone axis $\mathbf{r}(t)$. The features $\mathbf{f}$ concatenated with the

view direction $\mathbf{d}$ that has undergone original positional encoding proposed by NeRF [38], can be mapped to density and color using a tiny MLP parameterized by Θ_c: $\Phi_c(\mathbf{f}, \gamma(\mathbf{d})) = (\sigma, \mathbf{c})$.

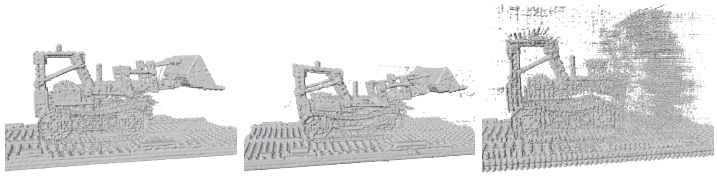
Dynamic NeRFs As previously stated in Sec. 2, there exists a category of studies that explicitly incorporate timestamps into NeRF, resulting in the formulation $f(\mathbf{x}, \mathbf{d}, t) = (\sigma, \mathbf{c})$. The approach employed by other researchers involves utilizing a mix of a deformation field $\mathbf{\Psi}_t$ and a canonical 3D representation as an extension of NeRF to the dynamic setting, which was initially introduced by D-NeRF [45]. It proposed an extension of NeRF to the dynamic setting by introducing a deformation network $\mathbf{\Psi}_t$, which is responsible for mapping the observation space at timestamp t to the unified canonical space.

3.2 Deformable Mipmapped Tri-Planes (DMiT)

Methods mentioned in Sec. 2 effectively enables anti-aliasing during rendering with high-quality visual effects, which are also applied to neural rendering for static scenes [6, 7, 22]. Despite their successful application to reconstruct high-quality images, it is important to note that their applicability is limited to static scenes. On the other hand, the standard D-NeRF framework encounters challenges in accurately capturing high-frequency details in dynamic scenes, not to mention its long training time. In this work, we introduce a novel anti-aliasing framework (DMiT) for dynamic NeRF by using mipmaps for tri-planes representation to enhance the overall quality of dynamic scene rendering.

Inspired by [22,45], mipmapped tri-planes (Tri-Mip) are utilized as the representation for the shared 3D static canonical space to learn effect and accurate deformation. However, it is non-trivial to directly incorporate the efficient Tri-Mip encoding with MLP-based deformation, since these two modules bear different convergence speeds. The straightforward pipeline is prone to causing severe visual artifacts, such as the blending of canonical space with high-frequency signals from other frames (see Fig. 3c), as a result of the overfitting of tri-planes. Moreover, when the deformation gains faster convergence, the canonical space may end up being compressed and stretched as shown in Fig. 3b, thus being unable to restore a reasonable geometry. Both categories of artifacts have a significant impact on the accuracy of reconstructing the canonical space which makes it almost impossible to separate shape from motion in dynamic scenes. To alleviate these artifacts, a joint deformation and geometry refinement is well-designed and will be thoroughly discussed in Sec. 3.3. Fig. 3a shows the artifact-free canonical geometry decomposed from our proposed refinement.

Cone Casting in Observation Space Specifically, we adopt the cone-casting strategy from [6], which emits a cone $\mathcal{C}$ with central axis $\mathbf{a}(t)$. Unlike the point intersection approach used in NeRF [38], this cone-casting strategy results in a planar (disk) intersection with the image plane. The intersection area is set to be the same with the corresponding pixel of width Δx and Δy. Therefore, the



(a) Ours **(b)** compression **(c)** fused

Fig. 3: Canonical geometry visualization comparison between our method and trivial baselines with different artifacts. Please zoom in for more details.

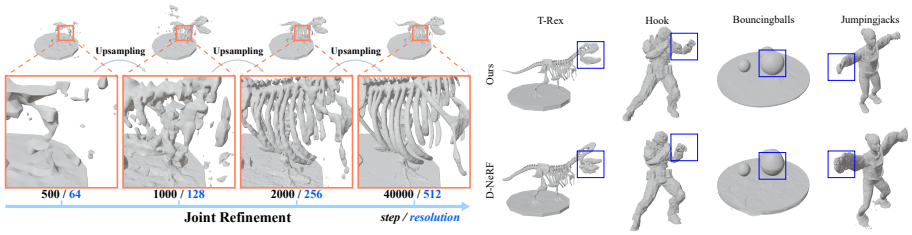
radius of the intersection disk can be estimated as $\dot{r} = \sqrt{\frac{\Delta x \cdot \Delta y}{\pi}}$. We proceed by selecting numerous isotropic spheres $\mathcal{S}(\mathbf{x}, r)$, that are inscribed within $\mathcal{C}$ with their radius r calculated as [22].

Deformation Formulation Consequently, the sampled spheres are deformed from observation space at time t to the unified canonical space through the deformation network parameterized by Θ_d. The deformation network is capable of making predictions regarding the spatial displacement from vector $\mathbf{x}$ to its corresponding position $\mathbf{x}_c = \mathbf{x} + \Delta\mathbf{x}$ within the canonical space. In a manner akin to the widely utilized ray casting method, the central axis $\mathbf{a}(t)$ of the emitted cone undergoes bending, while the radius r of the sampled sphere $\mathcal{S}(\mathbf{x}, r)$ remains constant, even if the whole cone has been distorted. Then the formulation of dynamics can be denoted as follows:

$$\Phi_d(\gamma(\mathbf{x}), \gamma(t)) = \Delta\mathbf{x} \tag{1}$$

where $\gamma(\cdot)$ is the frequency encoding. In contrast to [45], where the initial frame is designated as the canonical frame, our method does not explicitly specify a canonical frame. This design promotes the integration of dynamic information across all frames within a unified canonical space, which then serves as a template for the optimization of deformation. The comparison of canonical geometry against D-NeRF can be referred to Fig. 4b and a more detailed analysis of the canonical space will be provided in the supplementary.

Mipmapped Tri-Plane in Canonical Space To make use of the mipmapped tri-planes, which are known for the capacity to accurately record high-frequency appearances and reconstruct geometry with high fidelity, the deformed spheres $\mathcal{S}_c(\mathbf{x}_c, r)$ in the canonical space are further featured with an efficient Tri-Mip encoding. Specifically, each sphere is orthogonally projected onto tri-planes $\mathcal{M} = \{\mathcal{M}_{XY}, \mathcal{M}_{XZ}, \mathcal{M}_{YZ}\}$, forming three discs $\mathcal{D} = \{\mathcal{D}_{XY}, \mathcal{D}_{XZ}, \mathcal{D}_{YZ}\}$ of radius r, respectively. For pre-filtering, three compact feature planes serve as the base level mipmap $\mathcal{M}^{L_0}$ representing the decomposed 3D space, while the other mipmaps at different levels L_i are downscaled (by a factor of two) from the base mipmap. To query the corresponding features from the hierarchical mipmapped tri-planes, we can determine the mipmap level l that is suitable for the disk radius r. This



(a) Joint Refinement **(b)** Canonical Geometry Visualization

Fig. 4: (a) The diagram of our proposed joint geometry and deformation refinement procedure along with canonical geometry visualization results at each pre-set upsampling step. (b) Qualitative comparison of canonical geometry visualization results between our method and D-NeRF [45]. In the highlighted areas, the canonical geometry of D-NeRF exhibits obvious fusing artifacts, while our method does not.

can be achieved by utilizing the Axis Aligned Bounding Box, which is defined by two corners $\mathcal{B}_{min}$ and $\mathcal{B}_{max}$, of the specific 3D space of interest. Additionally, the mipmap resolution $(H,\ W)$ is also required.

$$\ddot{r} = \sqrt{\frac{(\mathcal{B}_{max} - \mathcal{B}_{min})_X \cdot (\mathcal{B}_{max} - \mathcal{B}_{min})_Y}{HW \cdot \pi}}, \quad l = \log_2(\frac{r}{\ddot{r}}) \tag{2}$$

The query coordinate of the pre-defined mipmap stack comprises the mipmap level l and the three disk centers (*e.g.*, $\mathbf{x}_{XY}$) and is then used to perform trilinear interpolation on the encoded feature (*e.g.*, $\mathbf{f}_{XY}$). The concatenated Tri-Mip feature $\mathbf{f} = \{\mathbf{f}_{XY}, \mathbf{f}_{XZ}, \mathbf{f}_{YZ}\}$ and the encoded ray direction $\gamma(\mathbf{d})$ are regressed to density σ and color $\mathbf{c}$ by a tiny MLP. Volume rendering [36, 38] is employed to generate the final color value for each pixel along the axis $\mathbf{a}(t)$ of cone $\mathcal{C}$.

3.3 Joint Geometry and Deformation Refinement

The canonical tri-planes of high resolution from Tri-MipRF [22] show a tendency of modeling high-frequency signals with a relatively fast learning speed, which has advantages in the case of static scenes. However, when reconstructing dynamic scenes, we rely on a deformation MLP to predict the motion. The fast convergence speed of the canonical Tri-MipRF branch may cause overfitting, producing fused canonical space of different time frames with discontinuous geometry or compressed shape as in Fig. 3. For instance, different parts from the moving shovel of the Lego truck at different times may be mixed. Hence the deformation network has difficulty in converging, leading to unsatisfactory rendering results.

To alleviate this issue, we incorporate two techniques to introduce a balance between the canonical Tri-MipRF and the deformation MLP. First, we apply a warm-up strategy on the resolution of the tri-planes with implicit regulation on high-frequency signals. We start with a low-resolution tri-plane, and gradually

upsample it during training. The evolution of the canonical geometry and tri-plane resolution over training is shown in Fig. 4a, demonstrating the regulation ability for floaters as well as the preservation of fine-grained geometry. Secondly, We also introduce a warm-up phase for more effective and stable motion decomposition, starting with a low learning rate for the deformation network that progressively increases until the tri-plane upsampling process is complete. After the warm-up phase for both motion and geometry, an annealing training strategy [64] is applied for efficient convergence.

Meanwhile, the original frequency encoding is no longer suitable for the refinement, as the encoded high-frequency signals conflict with the coarse canonical space initialization. Thus we introduce the annealing frequency encoding [41], which progressively increases the encoding dimension from zero to maximum:

$$\gamma_\alpha(\mathbf{x}) = \left(w_k(\alpha)\sin(2^k\pi\mathbf{x}), w_k(\alpha)\cos(2^k\pi\mathbf{x})\right)_{k=0}^{L-1}, \tag{3}$$

where $w_k(\alpha) = \frac{1}{2}(1-\cos(\pi\text{clamp}(\alpha-k,0,1)))$ is the corresponding weight, $\alpha(t) = \frac{Lt}{N}$ is the annealing coefficient with t denoting the training step and N representing the end step of the frequency annealing.

3.4 Optimization

Following previous works, we train our model using the photometric loss between rendered images and the ground truth. Let $\mathcal{R}$ represent the set of all rays and $\hat{C}(\mathbf{r})$ denotes the target pixel color of ray $\mathbf{r}$, the photometric loss is:

$$\mathcal{L}_{pho} = \frac{1}{|\mathcal{R}|}\sum_{\mathbf{r}}\left\|C(\mathbf{r}) - \hat{C}(\mathbf{r})\right\|_2^2 \tag{4}$$

To encourage the sparsity and compactness of the feature planes, we adopt spatial total variation loss as in [8, 10, 16, 17, 48],:

$$\mathcal{L}_{TV} = \sum_{h,w}\left(\left\|\mathcal{M}_{h,w}^{L_0} - \mathcal{M}_{h-1,w}^{L_0}\right\|_2^2 + \left\|\mathcal{M}_{h,w}^{L_0} - \mathcal{M}_{h,w-1}^{L_0}\right\|_2^2\right) \tag{5}$$

To summarize, the final optimization objective is $\mathcal{L} = \mathcal{L}_{pho} + \lambda_{TV}\mathcal{L}_{TV}$, where λ_{TV} is a hyperparameter.

4 Experiment

In this section, we qualitatively and quantitatively compare our work with several SOTA methods for validation. The datasets cover synthetic scenes as well as real-world scenes (including both monocular and multi-view setting). Due to the page limitation of the main paper, please refer to the supplementary material for per-scene metrics and additional qualitative comparisons. We highly recommend readers to watch the supplemental video to better evaluate the anti-aliased high-fidelity rendering results of our proposed method. All experiments are done with a single NVIDIA GeForce RTX 3090 GPU with 24GB memory.

Table 1: Quantitative comparison between our method and SOTA methods on the original D-NeRF dataset.

Methods	PSNR ↑	SSIM ↑	LPIPS$_v$ ↓	LPIPS$_a$ ↓
D-NeRF [45]	30.39	0.953	0.076	0.030
TiNeuVox [15]	31.63	0.963	0.058	0.039
HexPlane [8]	30.20	0.964	0.050	0.036
V4D [19]	<u>32.37</u>	0.973	0.038	0.024
K-Planes-H [16]	30.32	0.967	0.050	0.035
K-Planes-E [16]	29.63	0.962	0.056	0.041
Tensor4D [48]	25.71	0.940	0.091	0.095
4DGS [58]	32.30	<u>0.974</u>	<u>0.036</u>	<u>0.021</u>
Ours	**34.22**	**0.982**	**0.024**	**0.014**

Table 2: Quantitative comparison between our method and SOTA methods on the PlenopticVideo dataset.

Methods	PSNR ↑	MS-SSIM ↑	LPIPS$_a$ ↓
K-Planes [16]	31.30	0.963	-
LLFF [37]	23.24	0.848	0.235
DyNeRF [30]	29.58	0.961	**0.083**
MixVoxels-L [53]	31.09	0.963	0.099
HexPlane [8]	31.57	<u>0.969</u>	0.090
NeRFPlayer [49]	30.53	0.927*	0.116
HyperReel [4]	<u>31.54</u>	0.965	0.107
4DGS [58]	29.71	0.957	0.117
Ours	**31.73**	**0.974**	<u>0.086</u>

4.1 Implementation Details

Our implementation is based on Pytorch [43] with acceleration provided by tiny-cuda-nn [39]. Following [22], the Tri-Mip encoding takes advantage of mature mipmapping techniques included in the nvdiffrast library [28] for efficiency. For coarse tri-planes, We set the shape of the base mipmap $\mathcal{M}^{L_0}$ to $H_c = 64$, $W_c = 64$, $C = 16$ with maximum mipmap level $L_c = 7$. For the fine tri-planes after refinement, the shape of the base mipmap $\mathcal{M}^{L_0}$ is $H_f = 512$, $W_f = 512$, $C = 16$ with maximum mipmap level $L_f = 10$. The deformation network is implemented using an MLP with $D = 8$, $W = 128$. We train our model for 40K iterations for synthetic scenes and 100K iterations for real-world scenes in total with separated optimizers for static components and deformation network. More details can be referred to Sec. A in the supplementary material.

4.2 Evaluation on Synthetic Datasets

D-NeRF Dataset Firstly introduced in [45] by synthesizing through Blender [12] in a monocular setting, this dataset contains eight scenes in total by recording various objects in complex motion. The number of training images ranges from 50 to 200 while the test set of each scene has 20 images. For fair comparison, we train baselines from scratch using the released codes and their default configurations. The quantitative results are listed in Tab. 1 and canonical geometry visualization comparison with D-NeRF [45] is shown in Fig. 4b. Following previous works, three metrics are selected for quantitative evaluation, including peak signal-to-noise ratio (PSNR), structural similarity (SSIM) [55], and learned perceptual image patch similarity (LPIPS) [66]. Our method achieves high-fidelity rendering results, outperforming across all metrics..

Multi-scale D-NeRF Dataset Following [6, 22], we downsample images from the original D-NeRF dataset with a factor of 2, 4, and 8 respectively to provide observations across different scales. In the meantime, the focal lengths of downsampled images are downscaled according to the perspective camera projection. In this setting, metrics at different scales are calculated for better anti-aliasing evaluation, as listed in Tab. 3, where K-Planes-H and K-Planes-E denote hybrid

Table 3: Quantitative comparison of our method against several SOTA methods on the multi-scale D-NeRF dataset.

Methods	Train. ↓	PSNR ↑					SSIM ↑					LPIPS$_v$ ↓				
		Full Res.	1/2 Res.	1/4 Res.	1/8 Res.	Avg.	Full Res.	1/2 Res.	1/4 Res.	1/8 Res.	Avg.	Full Res.	1/2 Res.	1/4 Res.	1/8 Res	Avg.
D-NeRF [45]	1 d	28.01	28.37	29.49	28.94	28.70	0.935	0.947	0.951	0.947	0.944	0.065	0.052	0.063	0.064	0.062
TiNeuVox [15]	**28 m**	31.24	32.12	32.55	30.50	31.60	0.962	0.969	0.975	0.966	0.968	0.059	0.045	0.035	0.046	0.046
K-Planes-H [16]	1 h	27.26	27.67	27.97	27.92	27.61	0.955	0.954	0.953	0.949	0.952	0.062	0.056	0.054	0.063	0.059
K-Planes-E [16]	1 h	26.80	27.18	27.58	27.34	27.23	0.951	0.949	0.948	0.941	0.947	0.069	0.065	0.064	0.068	0.067
Tensor4D [48]	10 h	25.25	25.67	26.04	25.15	25.49	0.932	0.934	0.938	0.921	0.931	0.101	0.093	0.071	0.050	0.079
4DGS [58]	40 m	30.13	30.36	30.84	30.63	30.49	0.963	0.966	0.968	0.967	0.966	0.048	0.042	0.042	0.038	0.042
D3DGS [64]	35 m	32.40	32.60	32.96	32.76	32.68	0.976	0.978	0.979	0.979	0.978	0.032	0.027	0.027	0.025	0.028
Ours	30 m	**34.15**	**35.04**	**35.81**	**36.09**	**35.27**	**0.980**	**0.984**	**0.987**	**0.988**	**0.985**	**0.029**	**0.020**	**0.014**	**0.010**	**0.019**

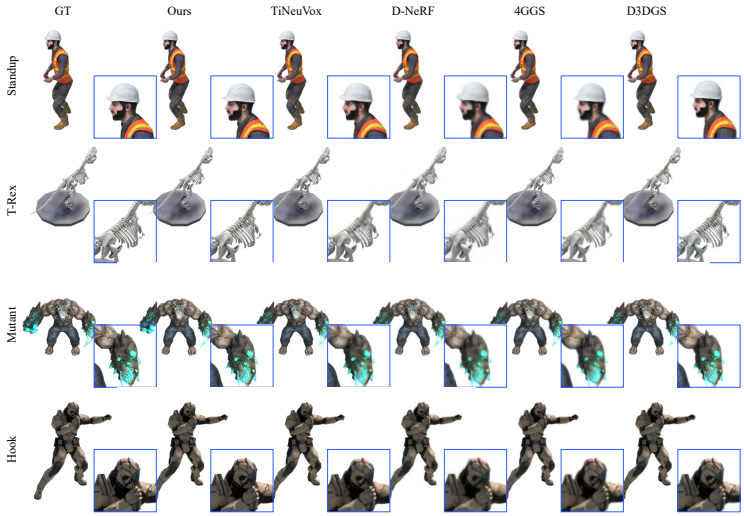


Fig. 5: Qualitative comparison of reconstruction results between our method and baseline methods on the multi-scale D-NeRF dataset.

and explicit architecture of [16] respectively. Our method outperforms others in all three metrics across various resolutions while achieving efficient training in 30 minutes. Besides, we show examples of the rendering results of full-resolution with close-up views in Fig. 5 for qualitative comparison. Our results are the closest to the ground truth with details restored such as the thin skeleton of *T-Rex* and facial features of *Standup*, while other SOTA methods only achieve blurry results. See supplementary for more anti-aliasing results.

4.3 Evaluation on Real-world Datasets

HyperNeRF Dataset To validate the applicability of our method in practice, we adopt the dataset from [42]. It contains dynamic scenes captured by one or two cameras with varying topology. Following [42], every 4th frame is selected for training, while the midlle frame is used as a test frame for the interpolation task. For the novel view synthesis task, frame IDs for validation have been pro-

Table 4: Quantitative comparison between our method and SOTA methods on the HyperNeRF dataset and NeRF-DS dataset.

Methods	HyperNeRF dataset			NeRF-DS dataset		
	PSNR ↑	MS-SSIM ↑	LPIPS$_a$ ↓	PSNR ↑	SSIM ↑	LPIPS$_a$ ↓
HyperNeRF [42]	27.35	**0.927**	**0.144**	23.45	0.849	0.181
TiNeuVox [15]	26.78	0.882	0.374	21.80	0.825	0.213
NeRF-DS [62]	-	-	-	23.61	0.853	0.128
4DGS [58]	27.05	0.900	0.266	22.10	0.789	0.207
D3DGS [64]	25.57	0.859	0.224	23.74	0.844	0.127
Ours	**27.80**	0.926	0.153	**24.48**	**0.866**	**0.115**

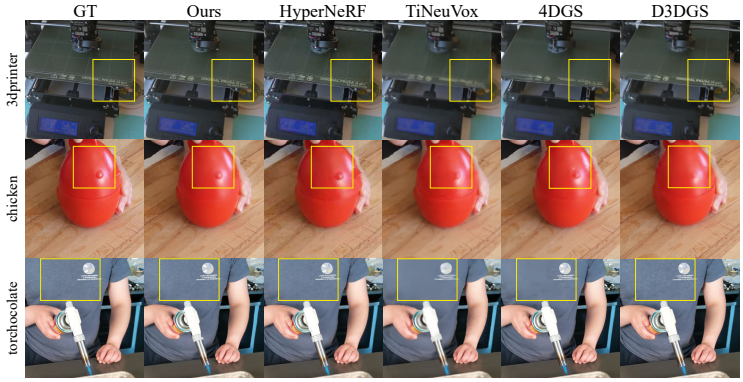


Fig. 6: Qualitative comparison of reconstruction results between our method and baseline methods on the HyperNeRF dataset. For a clearer view of the high-frequency and glossy details, refer to the highlighted area.

vided beforehand. We conduct the same experiments as [42]. For interpolation, we test on three scenes, *i.e.*, *torchocolate, hand* and *cut-lemon*. For novel view synthesis, *3d printer* and *chicken* are tested. For the evaluation benchmark, we use multi-scale structural similarity (MS-SSIM) [56] instead of SSIM following [42]. We use reported metrics of [42] while other results are based on their official implementation. Tab. 4 and Fig. 6 show quantitative and qualitative comparisons with SOTA methods, respectively. Though our method does not always outperform HyperNeRF (*e.g.*, MS-SSIM and LPIPS), intricate details such as tiny text patterns and glossy appearance can be restored while others generate blurry results. Meanwhile, [58] cannot represent complex clothing textures and [64] fails to recover fine-frained details as opposed to the proposed method.

NeRF-DS Dataset NeRF-DS [62] proposed a dataset consisting of 8 forward-facing scenes with deforming and moving objects in the real world, registered with more accurate camera poses compared to HyperNeRF dataset [42]. Similar to vrig datasets from [42], images for training are captured from one camera to prevent teleporting issues, while images for validating the novel view synthesis task are from another camera. Detailed quantitative results and qualitative comparisons can be found in Tab. 4 and Fig. 7. All baseline results are obtained through our experiments using their official implementation except for [42]. [62]

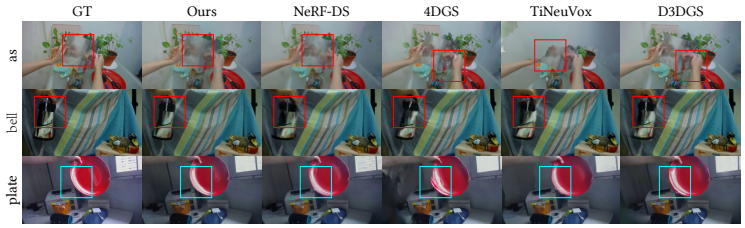


Fig. 7: Qualitative comparison of reconstruction results between our method and baseline methods on the NeRF-DS dataset.

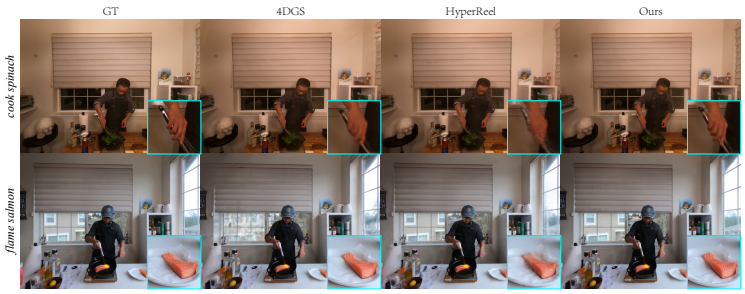


Fig. 8: Qualitative comparison of reconstruction results between our method and baseline methods on the PlenopticVideo dataset.

fails to model rapid translation of the moving object in *bell* scene. GS-based methods [58, 64] suffer from under-reconstruction artifacts in heavily obscured areas, while [15] produces foggy geometry and cannot model motion correctly as in [58]. Our method can better reconstruct complex specular and reflective appearances, as well as achieve more stable deformation, compared to other baselines.

PlenopticVideo Dataset This is a multi-view real-world scene dataset proposed by [30]. It includes 6 complex forward-facing scenes captured with 21 cameras. The central view is reserved for testing and other 18 synchronized views are chosen for training. Each view contains a 10-second video (300 frames) in resolution 2028×2704 (2.7K). Please refer to Tab. 2 for quantitative results and Fig. 8 for qualitative comparison. We directly use reported metrics from baselines that adopted PSNR, MS-SSIM, LPIPS (AlexNet) as benchmarks, except [49] that only reported SSIM metrics (marked as * in Tab. 2) and [4, 58]. We train [4, 58] from scratch using the official implementation. Besides, [30, 37] only reported metrics of *flame salmon* scene. We conduct all experiments on half resolution (1014×1352) of all scenes except an unsynchronized scene *coffee-martini*. Our method has achieved competitive results compared to SOTA methods across all three metrics, and it excels in restoring fine-grained details, as illustrated in the zoom-in view in Fig. 8.

Table 5: Ablation results on the multi-scale D-NeRF dataset.

Methods	PSNR ↑	SSIM ↑	LPIPS_v ↓	LPIPS_a ↓
Ours w/o JR	32.22	<u>0.977</u>	<u>0.030</u>	<u>0.018</u>
Ours w/o $\mathcal{M}$	<u>32.54</u>	0.973	0.038	0.027
Ours	**35.27**	**0.985**	**0.019**	**0.011**

4.4 Ablation Study

To verify the effectiveness of our Tri-Mip encoding and joint geometry and deformation refinement scheme, we conducted two ablation experiments on the multi-scale D-NeRF dataset. For the former, we design a variant of our method without mipmapping, namely "Ours w/o $\mathcal{M}$", following Tri-MipRF [22]. We replace the three pre-defined mipmap features with three plane features of the same size, thus the total trained parameters remain the same except that the mipmap and pre-filtering techniques are removed. For the latter, we derive another variant "Ours w/o JR", in which the resolution of the tri-miplane features is set to match the final resolution after the refinement stage. The quantitative results are listed in Tab. 5. More ablation results can be referred to the supplementary.

5 Conclusion & Future Work

In this work, we propose a novel NeRF-based framework called **DMiT**, which is the first to introduce anti-aliasing into dynamic scene representation. Our framework incorporates a deformation network for the mapping between the observation space and the hierarchical mipmapped tri-planes of the canonical geometry, as well as a joint refinement procedure that enables the decoupling of motion and geometry. Owing to the compactness and efficiency of tri-planes along with the well-designed mipmapping module, our method has achieved impressive training speed and high-fidelity rendering quality at various distances, while the blurry and aliasing artifacts in previous work are largely reduced. The qualitative and quantitative results on synthetic and real-world datasets have demonstrated the effectiveness and applicability of **DMiT**.

Since NeRF-based methods heavily rely on the accuracy of camera registration, imprecise camera parameters would inevitably affect our results. Also, because the tri-plane representation is based on orthogonal projection factorization, our methods may struggle to represent large-scale or unbounded scenes where spatial information cannot be fully encoded in a resolution-limited architecture. As our proposed method is a general dynamic framework, predicting accurate appearance without decomposition in more complex specular, reflective, or refractive scenes remains challenging due to multi-view inconsistencies across captured images, especially when using only monocular RGB images as input. Furthermore, our proposed method may restore false colors when using unconstrained input. Please see the supplementary materials for further discussion. We leave the limitations listed above for future work.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (No. 62322210 and No. 62302484), the Beijing Municipal Science and Technology Commission (No. Z231100005923031), the China Postdoctoral Science Foundation (No. 2023M743568), the Royal Society International Exchanges 2023 Cost Share (NSFC) (IEC\NSFC\233698), and Tencent AI Lab Rhino-Bird Focused Research Program (No. RBFR2023001).

References

1. Akeley, K.: Reality engine graphics. In: Proceedings of the 20th Annual Conference on Computer Graphics and Interactive Techniques. p. 109–116. SIGGRAPH '93, Association for Computing Machinery, New York, NY, USA (1993)
2. Akenine-Moller, T., Haines, E., Hoffman, N.: Real-time rendering. AK Peters/crc Press (2019)
3. Alexiadis, D.S., Zarpalas, D., Daras, P.: Real-time, full 3-d reconstruction of moving foreground objects from multiple consumer depth cameras. IEEE Trans. Multim. **15**(2), 339–358 (2013)
4. Attal, B., Huang, J.B., Richardt, C., Zollhoefer, M., Kopf, J., O'Toole, M., Kim, C.: Hyperreel: High-fidelity 6-dof video with ray-conditioned sampling. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16610–16620 (2023)
5. Attal, B., Laidlaw, E., Gokaslan, A., Kim, C., Richardt, C., Tompkin, J., O'Toole, M.: Törf: Time-of-flight radiance fields for dynamic scene view synthesis. Advances in Neural Information Processing Systems **34** (2021)
6. Barron, J.T., Mildenhall, B., Tancik, M., Hedman, P., Martin-Brualla, R., Srinivasan, P.P.: Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In: IEEE/CVF International Conference on Computer Vision. pp. 5835–5844 (2021)
7. Barron, J.T., Mildenhall, B., Verbin, D., Srinivasan, P.P., Hedman, P.: Zip-nerf: Anti-aliased grid-based neural radiance fields. IEEE/CVF International Conference on Computer Vision (2023)
8. Cao, A., Johnson, J.: Hexplane: A fast representation for dynamic scenes. IEEE/CVF Conference on Computer Vision and Pattern Recognition (2023)
9. Chan, E.R., Lin, C.Z., Chan, M.A., Nagano, K., Pan, B., Mello, S.D., Gallo, O., Guibas, L.J., Tremblay, J., Khamis, S., Karras, T., Wetzstein, G.: Efficient geometry-aware 3d generative adversarial networks. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition,. pp. 16102–16112. IEEE (2022)
10. Chen, A., Xu, Z., Geiger, A., Yu, J., Su, H.: Tensorf: Tensorial radiance fields. In: European Conference on Computer Vision (2022)
11. Chen, L., Peng, S., Zhou, X.: Towards efficient and photorealistic 3d human reconstruction: a brief survey. Visual Informatics **5**(4), 11–19 (2021)
12. Community, B.O.: Blender - a 3D modelling and rendering package. Stichting Blender Foundation, Amsterdam (2023)
13. Deering, M., Winner, S., Schediwy, B., Duffy, C., Hunt, N.: The triangle processor and normal vector shader: a VLSI system for high performance graphics. In: 15th Annual Conference on Computer Graphics and Interactive Techniques. pp. 21–30 (1988)

14. Du, Y., Zhang, Y., Yu, H.X., Tenenbaum, J.B., Wu, J.: Neural radiance flow for 4d view synthesis and video processing. In: IEEE/CVF International Conference on Computer Vision. pp. 14304–14314 (2021)
15. Fang, J., Yi, T., Wang, X., Xie, L., Zhang, X., Liu, W., Nießner, M., Tian, Q.: Fast dynamic radiance fields with time-aware neural voxels. In: SIGGRAPH Asia 2022 Conference Papers (2022)
16. Fridovich-Keil, S., Meanti, G., Warburg, F.R., Recht, B., Kanazawa, A.: K-planes: Explicit radiance fields in space, time, and appearance. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12479–12488 (2023)
17. Fridovich-Keil, S., Yu, A., Tancik, M., Chen, Q., Recht, B., Kanazawa, A.: Plenoxels: Radiance fields without neural networks. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5501–5510 (2022)
18. Fuchs, H., Goldfeather, J., Hultquist, J.P., Spach, S., Austin, J.D., Brooks Jr, F.P., Eyles, J.G., Poulton, J.: Fast spheres, shadows, textures, transparencies, and imgage enhancements in pixel-planes. In: 12th Annual Conference on Computer Graphics and Interactive Techniques. pp. 111–120. ACM New York, NY, USA (1985)
19. Gan, W., Xu, H., Huang, Y., Chen, S., Yokoya, N.: V4d: Voxel for 4d novel view synthesis. IEEE Transactions on Visualization and Computer Graphics (2023)
20. Haeberli, P., Akeley, K.: The accumulation buffer: hardware support for high-quality rendering. In: 17th Annual Conference on Computer Graphics and Interactive Techniques. pp. 309–318 (1990)
21. Hasenfratz, J., Lapierre, M., Sillion, F.X.: A real-time system for full body interaction with virtual worlds. In: EGVE. pp. 147–156. Eurographics Association (2004)
22. Hu, W., Wang, Y., Ma, L., Yang, B., Gao, L., Liu, X., Ma, Y.: Tri-miprf: Tri-mip representation for efficient anti-aliasing neural radiance fields. In: ICCV (2023)
23. Huang, Y.H., Sun, Y.T., Yang, Z., Lyu, X., Cao, Y.P., Qi, X.: Sc-gs: Sparse-controlled gaussian splatting for editable dynamic scenes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4220–4230 (2024)
24. Kaplanyan, A.S., Hill, S., Patney, A., Lefohn, A.E.: Filtering distributions of normals for shading antialiasing. High Performance Graphics **151**, 162 (2016)
25. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G.: 3d gaussian splatting for real-time radiance field rendering. ACM Transactions on Graphics **42**(4) (July 2023)
26. Klenk, S., Koestler, L., Scaramuzza, D., Cremers, D.: E-nerf: Neural radiance fields from a moving event camera. IEEE Robotics and Automation Letters (2023)
27. Kuznetsov, A.: Neumip: Multi-resolution neural materials. ACM Transactions on Graphics **40**(4) (2021)
28. Laine, S., Hellsten, J., Karras, T., Seol, Y., Lehtinen, J., Aila, T.: Modular primitives for high-performance differentiable rendering. ACM Transactions on Graphics **39**(6) (2020)
29. Leler, W.: Human vision, anti-aliasing, and the cheap 4000 line display. In: 7th Annual Conference on Computer Graphics and Interactive Techniques. pp. 308–313 (1980)
30. Li, T., Slavcheva, M., Zollhoefer, M., Green, S., Lassner, C., Kim, C., Schmidt, T., Lovegrove, S., Goesele, M., Newcombe, R., et al.: Neural 3d video synthesis from multi-view video. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5521–5531 (2022)

31. Li, Z., Niklaus, S., Snavely, N., Wang, O.: Neural scene flow fields for space-time view synthesis of dynamic scenes. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6498–6508 (2021)
32. Liu, J.W., Cao, Y.P., Mao, W., Zhang, W., Zhang, D.J., Keppo, J., Shan, Y., Qie, X., Shou, M.Z.: Devrf: Fast deformable voxel radiance fields for dynamic scenes. Advances in Neural Information Processing Systems **35**, 36762–36775 (2022)
33. Luiten, J., Kopanas, G., Leibe, B., Ramanan, D.: Dynamic 3d gaussians: Tracking by persistent dynamic view synthesis. In: 2024 International Conference on 3D Vision (3DV). pp. 800–809. IEEE (2024)
34. Luo, F., Zhu, Y., Fu, Y., Zhou, H., Chen, Z., Xiao, C.: Sparse rgb-d images create a real thing: A flexible voxel based 3d reconstruction pipeline for single object. Visual Informatics **7**(1), 66–76 (2023)
35. Mammen, A.: Transparency and antialiasing algorithms implemented with the virtual pixel maps technique. IEEE Computer Graphics and Applications **9**(4), 43–55 (1989)
36. Max, N.: Optical models for direct volume rendering. IEEE Transactions on Visualization and Computer Graphics **1**(2), 99–108 (1995)
37. Mildenhall, B., Srinivasan, P.P., Ortiz-Cayon, R., Kalantari, N.K., Ramamoorthi, R., Ng, R., Kar, A.: Local light field fusion: Practical view synthesis with prescriptive sampling guidelines. ACM Transactions on Graphics (TOG) **38**(4), 1–14 (2019)
38. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. In: European Conference on Computer Vision. vol. 12346, pp. 405–421 (2020)
39. Müller, T.: tiny-cuda-nn (4 2021), https://github.com/NVlabs/tiny-cuda-nn
40. Olano, M., Baker, D.: Lean mapping. In: 2010 ACM SIGGRAPH symposium on Interactive 3D Graphics and Games. pp. 181–188 (2010)
41. Park, K., Sinha, U., Barron, J.T., Bouaziz, S., Goldman, D.B., Seitz, S.M., Martin-Brualla, R.: Nerfies: Deformable neural radiance fields. In: IEEE/CVF International Conference on Computer Vision. pp. 5865–5874 (2021)
42. Park, K., Sinha, U., Hedman, P., Barron, J.T., Bouaziz, S., Goldman, D.B., Martin-Brualla, R., Seitz, S.M.: HyperNeRF: a higher-dimensional representation for topologically varying neural radiance fields. ACM Transactions on Graphics **40**(6), 1–12 (2021)
43. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems **32** (2019)
44. Peng, S., Yan, Y., Shuai, Q., Bao, H., Zhou, X.: Representing volumetric videos as dynamic mlp maps. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4252–4262 (2023)
45. Pumarola, A., Corona, E., Pons-Moll, G., Moreno-Noguer, F.: D-NeRF: Neural radiance fields for dynamic scenes. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10318–10327 (2021)
46. Samavati, T., Soryani, M.: Deep learning-based 3d reconstruction: a survey. Artifical Intelligence Review **56**(9), 9175–9219 (2023)
47. Schönberger, J.L., Frahm, J.M.: Structure-from-motion revisited. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4104–4113 (2016)
48. Shao, R., Zheng, Z., Tu, H., Liu, B., Zhang, H., Liu, Y.: 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. pp. 16632–16642 (2023)
49. Song, L., Chen, A., Li, Z., Chen, Z., Chen, L., Yuan, J., Xu, Y., Geiger, A.: Nerfplayer: A streamable dynamic scene representation with decomposed neural radiance fields. IEEE Transactions on Visualization and Computer Graphics **29**(5), 2732–2742 (2023)
50. Sun, J.M., Wu, T., Gao, L.: Recent advances in implicit representation-based 3d shape generation. Visual Intelligence **2**(1), 9 (2024)
51. Tretschk, E., Tewari, A., Golyanik, V., Zollhofer, M., Lassner, C., Theobalt, C.: Non-rigid neural radiance fields: Reconstruction and novel view synthesis of a dynamic scene from monocular video. In: IEEE/CVF International Conference on Computer Vision. pp. 12959–12970 (2021)
52. Verbin, D., Hedman, P., Mildenhall, B., Zickler, T., Barron, J.T., Srinivasan, P.P.: Ref-NeRF: Structured view-dependent appearance for neural radiance fields. IEEE/CVF Conference on Computer Vision and Pattern Recognition pp. 5481–5490 (2022)
53. Wang, F., Tan, S., Li, X., Tian, Z., Song, Y., Liu, H.: Mixed neural voxels for fast multi-view video synthesis. in 2023 ieee. In: CVF International Conference on Computer Vision (ICCV). pp. 19649–19659 (2023)
54. Wang, P., Liu, L., Liu, Y., Theobalt, C., Komura, T., Wang, W.: Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. In: Advances in Neural Information Processing Systems 34. pp. 27171–27183 (2021)
55. Wang, Z., Bovik, A., Sheikh, H., Simoncelli, E.: Image quality assessment: from error visibility to structural similarity. IEEE Transactions on Image Processing **13**(4), 600–612 (2004)
56. Wang, Z., Simoncelli, E.P., Bovik, A.C.: Multiscale structural similarity for image quality assessment. In: The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003. vol. 2, pp. 1398–1402. Ieee (2003)
57. Whitted, T.: An improved illumination model for shaded display. In: 6th Annual Conference on Computer Graphics and Interactive Techniques. p. 14. ACM (1979)
58. Wu, G., Yi, T., Fang, J., Xie, L., Zhang, X., Wei, W., Liu, W., Tian, Q., Wang, X.: 4d gaussian splatting for real-time dynamic scene rendering. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20310–20320 (2024)
59. Wu, L., Zhao, S., Yan, L.Q., Ramamoorthi, R.: Accurate appearance preserving prefiltering for rendering displacement-mapped surfaces. ACM Transactions on Graphics **38**(4), 1–14 (2019)
60. Wu, T., Yuan, Y.J., Zhang, L.X., Yang, J., Cao, Y.P., Yan, L.Q., Gao, L.: Recent advances in 3d gaussian splatting. Computational Visual Media pp. 1–30 (2024)
61. Xian, W., Huang, J.B., Kopf, J., Kim, C.: Space-time neural irradiance fields for free-viewpoint video. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 9421–9431 (2021)
62. Yan, Z., Li, C., Lee, G.H.: Nerf-ds: Neural radiance fields for dynamic specular objects. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8285–8295 (2023)
63. Yang, Z., Yang, H., Pan, Z., Zhang, L.: Real-time photorealistic dynamic scene representation and rendering with 4d gaussian splatting. In: International Conference on Learning Representations (ICLR) (2024)
64. Yang, Z., Gao, X., Zhou, W., Jiao, S., Zhang, Y., Jin, X.: Deformable 3d gaussians for high-fidelity monocular dynamic scene reconstruction. In: Proceedings of the

IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 20331–20341 (2024)
65. Zhang, K., Riegler, G., Snavely, N., Koltun, V.: Nerf++: Analyzing and improving neural radiance fields. arXiv preprint arXiv:2010.07492 (2020)
66. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 586–595 (2018)
67. Zhu, H., Nie, Y., Yue, T., Cao, X.: The role of prior in image based 3d modeling: a survey. Frontiers of Computer Science **11**, 175–191 (2017)
68. Zhuang, Y., Zhang, Q., Feng, Y., Zhu, H., Yao, Y., Li, X., Cao, Y.P., Shan, Y., Cao, X.: Anti-aliased neural implicit surfaces with encoding level of detail. In: SIGGRAPH Asia 2023 Conference Papers. pp. 1–10 (2023)