

Multiscale Representation for Real-Time Anti-Aliasing Neural Rendering

Dongting Hu
University of Melbourne

Zhenkai Zhang
University of Melbourne

Tingbo Hou
Google

Tongliang Liu
University of Sydney

Huan Fu*
Alibaba Group

Mingming Gong*
University of Melbourne

Abstract

The rendering scheme in neural radiance field (NeRF) is effective in rendering a pixel by casting a ray into the scene. However, NeRF yields blurred rendering results when the training images are captured at non-uniform scales, and produces aliasing artifacts if the test images are taken in distant views. To address this issue, Mip-NeRF proposes a multiscale representation as a conical frustum to encode scale information. Nevertheless, this approach is only suitable for offline rendering since it relies on integrated positional encoding (IPE) to query a multilayer perceptron (MLP). To overcome this limitation, we propose mip voxel grids (Mip-VoG), an explicit multiscale representation with a deferred architecture for real-time anti-aliasing rendering. Our approach includes a density Mip-VoG for scene geometry and a feature Mip-VoG with a small MLP for view-dependent color. Mip-VoG represents scene scale using the level of detail (LOD) derived from ray differentials and uses quadrilinear interpolation to map a queried 3D location to its features and density from two neighboring down-sampled voxel grids. To our knowledge, our approach is the first to offer multiscale training and real-time anti-aliasing rendering simultaneously. We conducted experiments on multiscale dataset, results show that our approach outperforms state-of-the-art real-time rendering baselines.

1. Introduction

The realm of computer vision and graphics is marked by the captivating yet formidable challenge of novel view synthesis. In recent times, neural volumetric representations, most notably the neural radiance field (NeRF) [35], have emerged as a promising breakthrough in reconstructing intricate 3D scenes from multi-view image collections.

*Equal contribution

Method	Multiscale Training	Real-time Rendering	Anti-aliasing Rendering
Mip-NeRF [3]	✓		✓
SNeRG [21]		✓	
MobileNeRF [10]		✓	✓
Ours	✓	✓	✓

Table 1: Our method is the first one concurrently addresses multiscale training, real-time and anti-aliasing rendering.

NeRF employs a coordinate-based multilayer perceptron (MLP) architecture to map a 5D input coordinate (including 3D spatial position and 2D viewing direction) to intrinsic scene attributes (namely, volume density and view-dependent emitted radiance) at that precise location. The pixel rendering process in NeRF involves casting a ray through the pixel into the scene, extracting the scene representation for points sampled along the ray, and ultimately fusing these components to produce the final color output.

While this rendering methodology excels when the training and testing images share a uniform resolution, challenges arise when the training images encompass varying resolutions. This discrepancy in resolutions leads NeRF to produce blurred rendering outputs due to the altered pixel footprints originating from diverse scales. Besides, in case that test viewpoints significantly deviate from the spatial distance of the training views, the sample rate for per pixel would be inadequacy, thereby results in aliasing artifacts.

To surmount this challenge, Mip-NeRF [3] emerges as a noteworthy solution, presenting a continuously-valued scale representation for coordinate-based models. Mip-NeRF introduces a pioneering technique, known as integrated positional encoding (IPE), which facilitate the scene representation with the knowledge of scale. Departing from the conventional ray-based approach, Mip-NeRF adopts a novel rendering strategy involving conical frustums. This representation not only enables effective multiscale training but also tangibly mitigates the persisting issue of aliasing ar-

tifacts. However, it’s important to note that this approach relies on querying the network with the scale-variant IPE. As a consequence, the integration of pre-cache techniques that circumvent positional encoding [21], remains elusive within the current Mip-NeRF’s formulation.

This paper introduces a novel multiscale representation, termed “Mip-VoG” (Mip Voxel Grids, drawing inspiration from “mipmap”), which addresses the challenge of training on images with varying scales and enables real-time anti-aliasing rendering during the inference stage (Tab. 1). Our approach commences by unveiling a “deferred” NeRF variant, wherein both scene geometry and color attributes are explicitly stored within the Mip-VoG framework. The intricate task of decoding view-dependent effects is executed using a compact multilayer perceptron (MLP) that efficiently processes each pixel only once. Instead of pre-training and baking a continuous NeRF into grids [21], we treat voxel values as parameters and directly optimize them as seen in the work by [16, 57]. Despite the “mip” nomenclature, Mip-VoG only maintains one density voxel grid and one feature voxel grid that represent the high-frequency spatial attributes. This structure permits the inference of lower frequency representations through a progressive down-sampling process, achieved via low-pass filters and interpolation algorithms [14]. Given a single 3D point sampled from a camera ray, Mip-VoG intelligently determines the level of detail (LOD) via ray differentials [22]. The LOD calculation is pivotal, as it establishes a pixel-to-voxel ratio that represents the point’s footprint on the full-resolution voxel grids. As a consequence, scene properties corresponding to this point are sampled by interpolating between two adjacent down-sampled level grids. Notably, camera rays cast from a low-resolution frame are spatially represented over a wider area, yielding a higher sample rate and capturing more lower-frequency information. During inference phrase, we pre-compute the voxel grids at each integer level and subsequently convert them to the sparse voxel grid data structure used by SNeRG [21] to increase the rendering speed.

We conducted a comprehensive evaluation of our approach using well-established NeRF datasets, including Synthetic-NeRF [35] and Multiscale-NeRF [3], in line with the multi-scale framework outlined in Mip-NeRF [3]. Our findings underscore the validity of our multiscale representation in effectively addressing complex multiscale training scenarios, while successfully preserving both low and high-frequency components inherent in the multiscale dataset. Comparative analysis against state-of-the-art real-time techniques reveals that our proposed approach excels in mitigating multiscale challenges, yielding impressive results. Furthermore, our method proves instrumental in achieving remarkable accuracy in anti-aliasing rendering, bolstering its applicability and potential impact.

2. Related Works

Scene Representation for View Synthesis Numerous scene representations have been proposed to tackle the intricate task of view synthesis. Approaches such as Light Field Representation [12, 26, 27, 51] and Lumigraph [18, 7] directly interpolate input images, albeit necessitating dense input data for novel view synthesis. In an effort to reduce the demand for exhaustive capture, subsequent studies represent light fields as neural networks [53, 2]. Layered Depth Images [13, 50, 52, 60] alleviate the requirement of input denseness, but their effectiveness hinges on the accuracy of depth maps for rendering photo-realistic images. Recent advancements have introduced methods to estimate Multiplane Images (MPIs)[15, 29, 34, 56, 65, 73] for scenes with forward-facing viewpoints, and voxel grids[54, 31] for inward-facing scenes. Mesh-based representations [13, 50, 52, 60, 37, 38, 19] constitute another notable category within the view synthesis realm, offering real-time rendering potential through optimized rasterization pipelines. However, these methods necessitate template meshes as priors to overcome gradient-based optimization challenges.

Recently, NeRF [35] emerges as a popular method for novel view synthesis. By using a MLP as an implicit and continuous volumetric representation, NeRF maps from a 3D coordinate to the volume density and view-dependent emission at that position. The success of NeRF brings numbers of attention into neural volumetric rendering for view synthesis. Many follow-on works have extended NeRF to generative models [8, 59, 49, 39], generalization [61, 69] dynamic scenes [41, 28, 44, 32, 48], relighting[5, 55, 72], and editing [42, 43, 58, 70, 30, 67], etc. Rendering an image via NeRF necessitates querying an extensive neural network at multiple 3D locations per pixel, resulting in approximately a minute per frame rendering time. Recent advancements seek to enhance NeRF’s rendering efficiency through explicit representation leveraging [9, 57, 16, 65], or by segmenting the scene into sub-regions with smaller neural networks [46, 45].

Real-time Neural Rendering A series of researchers has emerged to address the imperative demand for real-time rendering capabilities. PlenOctrees [68] introduces an innovative spherical harmonic representation of radiance, seamlessly transitioning it into an octree data structure. Fast-NeRF [17] strategically restructures NeRF through refactoring, incorporating a dense voxel grid to efficiently cache the scene of interest for accelerated rendering. iNGP [36] uses a hash-table to store the feature vectors and combining with fully-fused CUDA kernels to accelerate rendering processes. SNeRG [21] proposes a deferred architecture and extracts the scene properties from a pre-trained model into

a sparse grid data structure. On a divergent trajectory, MobileNeRF [10] adopts a unique approach, representing the scene using textured polygons and harnessing polygon rasterization to generate pixel-level features. These features, in turn, are decoded via a compact view-dependent MLP. Nevertheless, it’s important to note that the explicit representations harnessed by these approaches lack scale-agnostic adaptability, thus the efficacy in learning from training images with multiple resolutions is limited. Our primary focus in experimental comparisons resides with SNeRG and MobileNeRF, given their established performance on resource-constrained devices without CUDA access.

Reducing Aliasing in Rendering One straightforward solution to mitigate aliasing for coordinate-based neural representations is supersampling [63], which requires casting multiple rays through pixel during rendering to get the final result. While powerful in its anti-aliasing effects, supersampling exacerbates the already time-intensive rendering procedure of NeRF, thereby confining its utility primarily to offline rendering scenarios. To improve the efficiency, Mip-NeRF [3] proposes to cast a conical frustum into the scene space and render the 3D region instead of a single point. This approach avoids heavy computation burden by approximating the 3D region rendering using gaussian, their algorithm queries the network by IPE of the 3D input region to output the final density and radiance. Due to the heavy reliance on the network to decode the scale information, Mip-NeRF cannot leverage pre-cache techniques to enable real-time capability. Another common technique for reducing aliasing is pre-filtering [40, 23, 66, 6], which pre-filters the maps on a coarse mesh *e.g.* color maps, normal maps, linearly and separately. This strategy involves the pre-filtration of various maps on a coarse mesh, such as color maps and normal maps, independently and linearly. Notably, pre-filtering transfers the computational load to a pre-rendering stage, rendering it well-suited for real-time rendering scenarios. A widely adopted method in 3D rendering applications is mip mapping [14, 64, 11]. A serialization of images or textures, each of which is a progressively lower resolution representation of the previous one, are pre-computed ahead of time to increase rendering speed and reduce aliasing artifacts for real-time inference. Traditionally, mip mapping is integral to the texture mapping process for 3D meshes. Expanding upon this concept, we extend mip mapping to the realm of 3D neural volumetric rendering. Our approach involves applying mip mapping to the voxel grids data structure, rather than a mere 2D map. This extension capitalizes on the advantages of mip mapping to boost rendering efficiency and reduce aliasing artifacts within the domain of neural volumetric rendering.

Relation to DVGO, iNGP and ZipNeRF There are some remarkable concurrent works study the voxel representation for efficient rendering. Still, there are some difference between Mip-VoG and these works. DVGO [57] progressively optimize a higher resolution voxel grid in the training for finer details, but it does not considering a multi-scale representation. Besides, DVGO stores the implicit feature for radiance emission and predict the final pixel color by a shallow MLP. By contrast, our Mip-VoG stores the explicit value for diffuse color and implicit feature for view-dependent specular radiance. iNGP [36] introduces a hash encoding approach based on a multi-resolution structure for speedy high-quality image synthesis. Their approach simultaneously trains several dense grid with different scales and concatenates the feature from each for further prediction. This representation only works for single-scale images since its scale-invariant. In contrast, our method only retains a single voxel grid during training and can progressively sample from Mip-VoG with different LOD. A concurrent work, ZipNeRF [4] operates within a similar setting, where they integrate iNGP’s grid pyramid using multi-sampling within the Mip-NeRF framework, whereas our work takes a distinct approach by incorporating the fundamental concept of “mip” directly into voxel grids.

3. Method

3.1. Review of NeRF

NeRF [35] uses a MLP parameterized by θ as a continuous volumetric function to represent a scene. The network takes input as the view direction d and 3D coordinate $r(t)$ sampled from a camera ray $r(t) = o + td$, and predict the volume density σ at that 3D position together with the view-dependent radiance c from that view direction:

$$\sigma(t), c(t) = \text{MLP}_{\theta}(r(t), d). \quad (1)$$

A vital assumption made by NeRF is to model the density σ only depend on location, while emitted color is conditional on 3D coordinate $r(t)$ and view direction d . In the rendering procedure, NeRF takes the predicted densities and emissions $\{\sigma(t_i), c(t_i)\}_{i=1}^N$ along the ray casted from a pixel, and approximate a volume rendering integral [33] to derive the final color of that pixel:

$$\hat{C}(r) = \sum_{i=1}^N T_i \left(1 - \exp(-\sigma(t_i)\delta(t_i)) \right) c(t_i), \quad (2)$$

with $T_i = \exp\left(-\sum_{j=1}^{i-1} \sigma(t_j)\delta_j\right)$,

where $\delta(t_i) = t_{i+1} - t_i$ is the distance between adjacent samples. One can find that rendering a single ray for each pixel requires evaluating the MLP hundreds of times, resulting in significantly slow rendering speed.

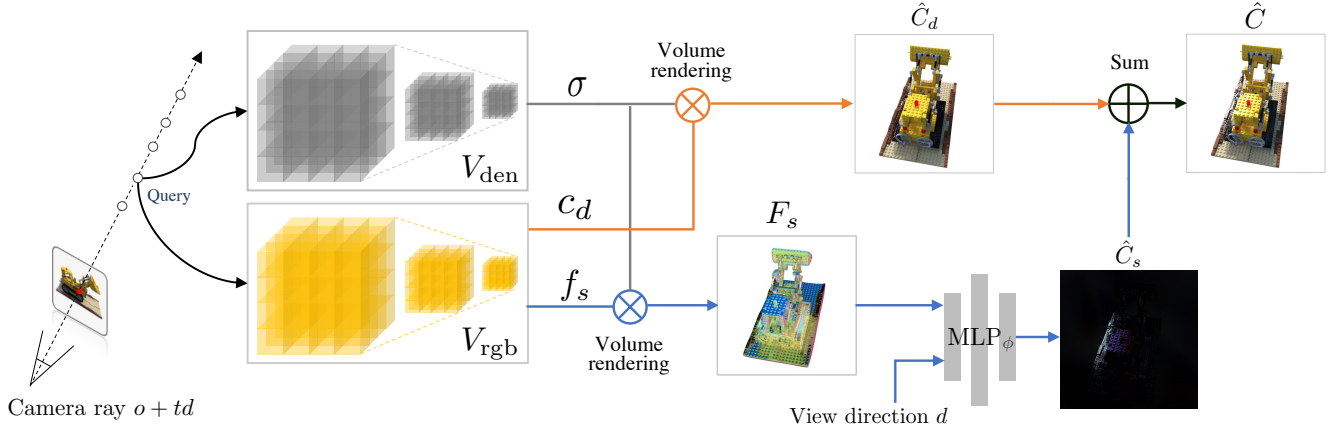


Figure 1: **Rendering framework overview.** We based our pipeline design on deferred NeRF [21] with explicit training of Mip-VoG. Given a point sampled from a camera ray, we query the density Mip-VoG (V_{den}) and color Mip-VoG (V_{rgb}) to predict the 1D volume density σ , 3D diffuse color c_d and 4D feature vector f_s . We then aggregate the diffuse colors and feature vectors along the ray through volume rendering integral (Eq. 3 and 4), resulting in \hat{C}_d and F_s . After that, a tiny MLP is used to predict a pixel-wise view-dependent specular color \hat{C}_s by using the accumulated feature-vector F_s together with the view direction d (Eq. 5). The final color prediction \hat{C} is the summation of diffuse color \hat{C}_d and specular color \hat{C}_s (Eq. 6).

3.2. Review of Deferred NeRF

As discussed previously, real-time rendering can be achieved by pre-computing as many as scene properties. While the scene geometry (volume density) can be directly stored, NeRF relies on a continuous function to represent view-dependent effects. To address this problem, SNeRG [21] introduces a residual architecture that first caches the point-wise pre-trained volume density $\sigma(t)$, diffuse color $c_d(t)$ and 4-dimension feature vector $f_s(t)$. The pixel-wise diffuse color and feature vector are obtained through volume rendering (same as NeRF):

$$\hat{C}_d(r) = \sum_{i=1}^N T_i \left(1 - \exp(-\sigma(t_i)\delta(t_i)) \right) c_d(t_i), \quad (3)$$

$$F_s(r) = \sum_{i=1}^N T_i \left(1 - \exp(-\sigma(t_i)\delta(t_i)) \right) f_s(t_i). \quad (4)$$

Then, a tiny MLP parameterized by ϕ , which forwards once for each pixel based on the feature $F_s(r)$ and view direction d , is applied to predict the pixel-wise specular color as a view-dependent residual:

$$\hat{C}_s(r) = \text{MLP}_\phi(F_s(r), d). \quad (5)$$

The final color of the pixel is obtained by the summation of diffuse color and specular color:

$$\hat{C}(r) = \hat{C}_d(r) + \hat{C}_s(r). \quad (6)$$

Vanilla SNeRG involves pre-training a continuous representation first and caching voxel-wise σ , c_d and f_s into a

sparse voxel grid. In contrast, our method directly learns an explicit representation from scratch, which can be directly used for efficient multiscale representation. We optimize $\sigma \in \mathbb{R}$ in one density Mip-VoG V_{den} and $c_d \in \mathbb{R}^3$ together with $f_s \in \mathbb{R}^4$ in one color Mip-VoG V_{rgb} . The overview of our framework is shown in Fig. 1. In the following section, we present the query algorithms of Mip-VoG.

3.3. Mip Voxel Grids

Mip-VoG harnesses the potential of a sequence of progressively down-sampled "much in little" voxel grids to facilitate scene queries at specific 3D coordinates. When sampling from Mip-VoG for a singular point within the spatial domain, a pivotal initial step involves the calculation of Level of Detail (LOD), which serves as a representative "correct" scale with regard to the complete-resolution voxel grids (section 3.3.1). Subsequent procedures encompass filtering and down-sampling operations applied to the comprehensive voxel grids, resulting in the generation of voxel grids at progressively lower scales. The value sampled from the Mip-VoG, corresponding to the specific LOD, is then obtained through interpolation, which bridges the information across two distinct scales of voxel grids at neighboring levels (section 3.3.2).

To streamline notation, we simplify the representation by omitting subscripts, condensing V_{den} and V_{rgb} into a singular notation: $V^{(0)}$ signifies the original voxel grids at level 0. By analogy, the terms $V^{(1)}$, $V^{(2)}$, $V^{(k)}$ denote the voxel grids after down-sampling, signifying levels 1, 2, and k respectively.

3.3.1 Level of Detail

Given a single ray cast through a pixel, ray differentials represent a pair of differentially offset rays slightly above or to the right of the original ray [22, 1]. We extend this idea to the volumetric rendering with voxel grids. Denote u, v, w as the unit coordinates as regard to the voxel space $V^{(0)}$, for a pixel of the frame with image space coordinates x and y , the ray differentials are defined as derivatives of the ray’s footprint on voxel space with respect to image space coordinates (x and y):

$$\frac{\partial V^{(0)}}{\partial x} = \left\langle \frac{\partial u}{\partial x} \frac{\partial v}{\partial x} \frac{\partial w}{\partial x} \right\rangle, \quad \frac{\partial V^{(0)}}{\partial y} = \left\langle \frac{\partial u}{\partial y} \frac{\partial v}{\partial y} \frac{\partial w}{\partial y} \right\rangle. \quad (7)$$

By applying a first-order Taylor approximation, we can get an expression for the extent of a pixel’s footprint in voxel space based on the voxel-to-pixel spacing:

$$\frac{\partial V^{(0)}}{\partial x} \approx \left\langle \frac{\Delta u}{\Delta x} \frac{\Delta v}{\Delta x} \frac{\Delta w}{\Delta x} \right\rangle, \quad \frac{\partial V^{(0)}}{\partial y} \approx \left\langle \frac{\Delta u}{\Delta y} \frac{\Delta v}{\Delta y} \frac{\Delta w}{\Delta y} \right\rangle. \quad (8)$$

Intuitively, this can be seen as the per-axis offset on the voxel grids made by the deviation of the ray in the image plane. As illustrated in Fig. 2, we adopt $\Delta x, \Delta y$ as the half pixel size along the x and y axis of the image plane, since it can approximate the footprint of a pixel. Given the position of a point $r(t)$ and its neighbors $r_{\Delta x}(t), r_{\Delta y}(t)$ in the world space, the offset in voxel unit coordinate can be derived by the ratio between the distance per axis and the voxel size:

$$\Delta u = \frac{T_u}{V_u}, \quad \Delta v = \frac{T_v}{V_v}, \quad \Delta w = \frac{T_w}{V_w}, \quad (9)$$

where T_u, T_v, T_w are per axis distance to the neighbors in the world space and V_u, V_v, V_w are voxel size on u, v, w along three axes. Then following the convention of mipmapping [20, 64], the LOD λ can be calculated as ¹:

$$\lambda = 1/3 * \log_2(\rho)$$

$$\text{with } \rho = \max \left(\sqrt{\left(\frac{\Delta u}{\Delta x}\right)^2 + \left(\frac{\Delta v}{\Delta x}\right)^2 + \left(\frac{\Delta w}{\Delta x}\right)^2}, \sqrt{\left(\frac{\Delta u}{\Delta y}\right)^2 + \left(\frac{\Delta v}{\Delta y}\right)^2 + \left(\frac{\Delta w}{\Delta y}\right)^2} \right). \quad (10)$$

3.3.2 Filtering and Sampling

Once LOD has been determined, the subsequent step entails sampling the relevant information from the voxel grid at the appropriate scale, as mandated by the Mip-VoG framework. To achieve ”much in little” voxel grids, we progressively down-sample the original voxel grid $V^{(0)}$ into

¹Since the number of voxels drops by 8x each level, we have $\log_8(\rho) = \log_2(\rho) / \log_2(8) = 1/3 * \log_2(\rho)$.

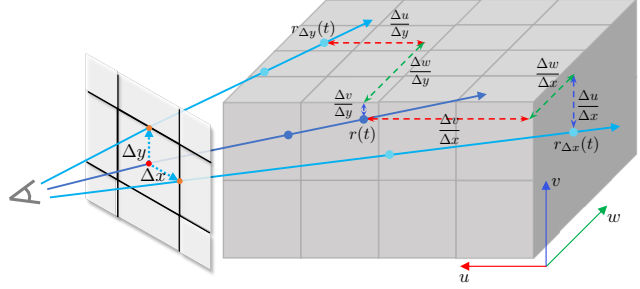


Figure 2: **Demonstration of ray differentials.** To determine the LOD of a point on the camera ray, we first cast two distinct rays, each generated by introducing an offset equivalent to half the pixel size along the x and y axes within the screen space. Subsequently, for every point situated along the original ray trajectory $r(t)$, we compute unit distance between this point its neighbours $r_{\Delta x}(t), r_{\Delta y}(t)$ along three axis on the full-resolution voxel grids $V^{(0)}$, as $\Delta u, \Delta v, \Delta w$. The LOD is finally calculated based on the largest length of the voxel-to-pixel ratio (Eq. 10).

a series of successively lower resolutions. Prior to down-sampling, a pivotal preparatory step involves the application of a low-pass filter denoted as γ . This filter serves to mitigate high-frequency information, ensuring that the down-sampled representations are suitably refined and devoid of artifacts. Following most common mipmap techniques [25], the down-sampling process is conducted in a hierarchical fashion. With each increment in LOD by one level (integer), the default operation involves down-sampling the voxel grid using a scale of $1/2$:

$$V^{(k+1)} = \Downarrow_{1/2} (\gamma(V^{(k)})), \quad (11)$$

where $\Downarrow_{1/2}(\cdot)$ represents the down-sampling with $1/2$ resolution. This process iterates, progressively generating new voxel grids at lower resolutions, thereby accommodating the spatial requirements of the specific LOD and maintaining the coherence of the Mip-VoG representation. In our approach, we employ linear interpolation as the method for down-sample filtering. To illustrate, consider the original voxel grid $V^{(0)}$ with dimensions $D \times N_x \times N_y \times N_z$. When generating $V^{(1)}$, the dimensions are adjusted through scaling to $D \times N_x/2 \times N_y/2 \times N_z/2$. The parameter D signifies the dimension of the modality being considered. This process is extrapolated to achieve a feature voxel grid with higher LOD and correspondingly lower resolution. The resulting structure preserves the hierarchical nature of Mip-VoG and contributes to a coherent representation of the scene, as shown in Fig. 3.

To ensure the smooth continuity across non-integer LOD λ , we adopt quadrilinear interpolation to aggregate samples from two neighboring voxel grids at different levels (upper and lower). Let $f(V, p)$ denote the sampling function of the

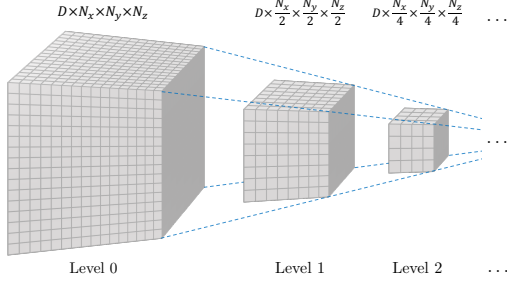


Figure 3: **Structure of Mip-VoG.** At each incremental increase in LOD, we first perform filtering on the previous voxel grids using low-pass filter, to remove high-frequency details. Subsequently, the filtered voxel grids are down-sampled to half of the previous resolution using linear interpolation. When querying the value for a 3D point within the space, Mip-VoG entails the interpolation of results derived from two adjacent voxel grids, selected in correspondence with the LOD of the point.

voxel grids, and $\mu \in \{\sigma, c_d, f_s\}$ represent the stores value, the interpolation process can be mathematically expressed as follows:

$$\mu = (\lceil \lambda \rceil - \lambda) f(V^{(\lfloor \lambda \rfloor)}, r(t)) + (\lambda - \lfloor \lambda \rfloor) f(V^{(\lceil \lambda \rceil)}, r(t)), \quad (12)$$

where $\lfloor \cdot \rfloor$ and $\lceil \cdot \rceil$ are the floor and ceiling function. This quadrilinear interpolation mechanism ensures that the sampled values are seamlessly blended between the two neighboring voxel grids, preserving the continuity of the representation across different LOD values.

3.3.3 Optimization

Similar to the prior work [57], our optimization is mainly divided into two stages: coarse and fine. In the coarse stage, we normally train $V_{\text{den}(c)} \in \mathbb{R}^{1 \times N_x^{(c)} \times N_y^{(c)} \times N_z^{(c)}}$ with $V_{\text{rgb}(c)} \in \mathbb{R}^{3 \times N_x^{(c)} \times N_y^{(c)} \times N_z^{(c)}}$ without using Mip-VoG, as to only obtain a rough 3D geometry $V_{\text{den}(c)}$ for reducing the number of sampling points in the fine stage. In the fine stage, we train a density Mip-VoG $V_{\text{den}(f)} \in \mathbb{R}^{1 \times N_x^{(f)} \times N_y^{(f)} \times N_z^{(f)}}$ and a color Mip-VoG $V_{\text{rgb}(f)} \in \mathbb{R}^{7 \times N_x^{(f)} \times N_y^{(f)} \times N_z^{(f)}}$ with a tiny MLP $_{\phi}$ as introduced before. We use gradient-descent to directly optimize value in voxel grids. As the gradient of the linear interpolation used in Mip-VoG downsampling $\Downarrow_{1/2}$ is tractable, the gradient from voxel grids with different resolution can be naturally aggregated and propagated. The loss function is the square error between the predicted pixel color and the ground truth:

$$\mathcal{L}_r = \sum_i \|C(\mathbf{r}_i) - \hat{C}(\mathbf{r}_i)\|_2^2. \quad (13)$$

4. Experiments

In light of our previous discussion, we primarily compare the results under the real-time rendering setting. We mainly evaluate our method on a simple multiscale synthetic dataset from Mip-NeRF [3] designed to better validate the accuracy on multi-resolution frames. We also conduct the experiment on its single-resolution version blender dataset introduced in the original NeRF paper [35], in order to probe our aliasing performance of the model training on a single-scale dataset. We report the three commonly studied error metrics: PSNR, SSIM [62], and LPIPS [71], and showcase some qualitative results.

4.1. Datasets

1. Synthetic-NeRF [35] presented in the original NeRF paper contains eight scenes. In this single-scale dataset, each scene consists 100 training images and 200 test images with uniform $800 * 800$ resolution. The model trained on this dataset can learn all the high-frequency details from the full resolution images, without being harmed by training images at multiple scales.
2. Multiscale-NeRF [3] is a straightforward conversion to Synthetic-NeRF for analyzing multiscale training and aliasing. It was generated by taking each image in Synthetic-NeRF and box down-sampling it by a factor of 2, 4, and 8 (and modifying the camera intrinsics accordingly). The three down-scaled images along with the original images are then combined into one single dataset. Hence this dataset contains image with four different scales for both training and test set, and the size has been quadrupled. The average evaluation metric is reported as the arithmetic mean of each error metric across all four scales. As suggested by Mip-NeRF [3], we adopt the Area Loss for all the methods which scale the pixel’s loss by the footprint size in the full resolution images, to balance the influence between high and low resolution pixels.

4.2. Implementation Details

In our experiments, we set the same hyperparameters for single-scale and multiscale datasets. In the coarse stage, the resolution of the voxel grid for both density and color is $(128 \times 128 \times 128)$, while in the fine stage, it raises to $(512 \times 512 \times 512)$. The low pass filter is adopted as the Mean Filter with kernel size 5. We use “shifted softplus” mentioned in Mip-NeRF [3] as the density activation. The initial values of alpha is 10^{-6} in the coarse training stage, and 10^{-2} in the fine training stage. Our tiny MLP follows the architecture used in SNeRG [21]. We use the Adam optimizer [24] to train both voxels and the deferred MLP, the learning rate are set to $4 * 10^{-3}$ for the deferred MLP and $1 * 10^{-1}$ for the voxel grids. In addition, we train 10k

Method	PSNR \uparrow				SSIM \uparrow				LPIPS \downarrow			
	Full Res	1/2 Res	1/4 Res	1/8 Res	Full Res	1/2 Res	1/4 Res	1/8 Res	Full Res	1/2 Res	1/4 Res	1/8 Res
Mip-NeRF [3]	32.629	34.336	35.471	35.602	0.958	0.970	0.979	0.983	0.047	0.026	0.017	0.012
SNeRG [21]	27.043	28.405	30.044	28.544	0.912	0.932	0.952	0.950	0.100	0.067	0.047	0.049
MobileNeRF [10]	24.115	25.127	26.633	27.930	0.868	0.885	0.913	0.938	0.141	0.112	0.078	0.050
MobileNeRF [10] w/o SS	23.730	24.425	25.308	25.364	0.861	0.875	0.898	0.910	0.149	0.128	0.104	0.091
Ours	30.333	31.290	31.055	29.014	0.946	0.956	0.960	0.955	0.069	0.049	0.045	0.048

Table 2: **Quantitative results on Multiscale-NeRF.** For comparison of models trained and evaluated on multiscale dataset. All the metrics of the scale are averaged across eight scenes. “w/o SS” removes supersampling from MobileNeRF.

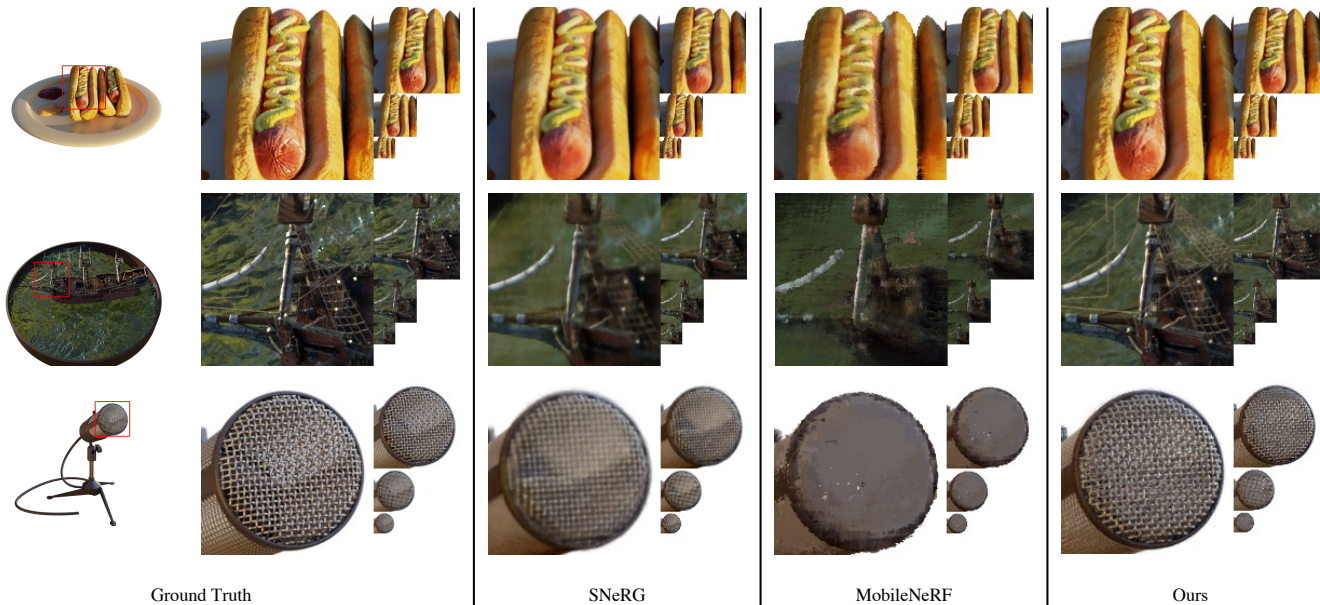


Figure 4: **Qualitative results on Multiscale-NeRF.** We demonstrate Mip-VoG rendering results compared to other baselines on the test set from three scenes, trained and evaluated on multiscale dataset. We visualize a crop region (shown in red box) on a same image at 4 different scales as an image pyramid. MobileNeRF yields over smooth results on all scales, while SNeRG lost high frequencies in high-resolution images and product aliasing in low-resolution frames. Our method surpass the baselines by a large margin as the rendering quality is significantly better.

and 20k iterations for the coarse phase and fine phase with the batch size of 8192, respectively. For real-time web renderer, we convert Mip-VoG to the sparse voxel grid data structure [21] and implement our query procedure in WebGL using the THREE.js library to increase the rendering speed. In terms of a fair comparison, all the methods are trained on a 80GB A100 GPU and tested on laptop GPU. Mip-VoG enables rendering 800×800 images in at 52 FPS on Lenovo Legion 7 (laptop) w/ NVIDIA RTX 2070 SUPER and 71 FPS on Alienware M15R6 (laptop) w/ NVIDIA RTX 3080, with a 104 MB storage footprint.

4.3. Results

Multiscale-NeRF The performance of Mip-VoG for this dataset can be seen in Tab. 2. As shown in the table, our method outperforms baselines on all metrics across all scales. Note that the result is a consequence of both multi-

scale training and anti-aliasing rendering, since this dataset contains multi-resolution images for both training and test set. Hence we defer the ablation into the next section. Since IPE is incompatible with current grid-based approaches that don’t use PE, we only include the result of Mip-NeRF [3] for reference purpose. Our approach, as other real-time methods, sacrifices rendering quality due to network size, resulting in lower performance compared to Mip-NeRF. We visualize some qualitative results in Fig. 4. One can see that other real-time rendering approaches produce blurry results on high resolution images, due to the issue of multiscale training that leads the models fit on low-resolution images. In contrast, our result learns high-frequency information in the full resolution images and output low-frequencies in low resolution frames. Additionally, we visualize the computed LOD in Fig. 5, the pixel-wise results is accumulated through the ray through volume rendering integral.

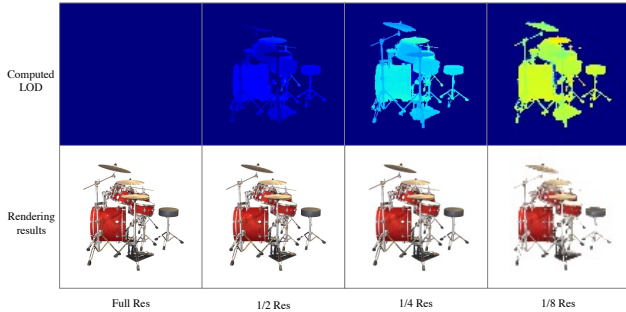


Figure 5: **Visualization of LOD.** We visualize the per-pixel LOD at four different scales. The value is computed using volume rendering integral of the points’ LOD along the ray. We resized the low-resolution rendered images to match the dimensions of the full resolution images, enhancing visibility and facilitating direct comparison. Brighter color indicates higher values.

Synthetic-NeRF Since the baselines models are not compatible with multiscale training, we eliminate the factor of non-uniform training images but to examine the effectiveness on anti-aliasing rendering. For this dataset the model is trained on single scale images and evaluated on the multiscale version, since the testset Multiscale-NeRF contains all the test images in Synthetic-NeRF. This inference scenario can be seen as rendering the single scale dataset but with the distance to the viewpoint has increased by scale factors of 2, 4, and 8 (also known as minification). In Tab. 3, excluding the effect of multiscale training set, our method is still outperforming the baselines on all metrics on high-resolution images. MobileNeRF [10], benefiting from its super-sampling technique, performs better on lower-resolution frames. A potential reason for the large enhancement of low-resolution renderings through super-sampling (MobileNeRF w/o SS vs. MobileNeRF) is that the final view is down-sampled from a higher-resolution rendered image, which improves accuracy and approximates the generation of ground truth low-resolution images. While removing super-sampling from it in the inference phase, our method outperforms the baseline models across all the scales. We showcase some qualitative results in Fig. 6. One can find our method produce more low-frequency details and mitigate aliasing artifact in the low-resolution images if zooming in, while in high-resolution images our method preserve sharp high-frequency details. This results verify that our method effectively learns a multiscale representation since it improves the both multi-resolution training and anti-aliasing.

4.4. Ablations

In this section we analysis the effectiveness of the contribution of Mip-VoG to the model, and give some insights

into the filtering algorithm. We perform all the experiments on Multiscale-NeRF [3], and report the average PSNR over the eight scenes.

Mipmapping To better examine the validity of Mip-VoG, we ablate this technique from training and testing respectively. While produce the rendering result without Mip-VoG reflects the ability of training with the images at multiple resolution, removing it from training effectively shows the improvement on anti-aliasing. As the results shown in Tab. 4, training without Mip-VoG shows lower accuracy in high-resolution test images and higher quality in low-resolution frames. This result is consistent with the studies in Mip-NeRF [3], as the area loss would force model to “overfit” on low-resolution training samples. While training with mip-VoG can help preserve high-frequencies rendering without Mip-VoG would produce aliasing in low-resolution frames, as the metrics of “Ours w/o te-mip” in lower scale is worse than the baseline. Hence, using Mip-VoG in both training and testing contributes the multiscale training and anti-aliasing rendering.

Low-pass Filter One design in the sampling phase is that the voxel grids are pre-filtered by a low-pass filter, which help preserve the high frequency information in the training and abandon them in rendering. We test our model based on three type of options: no filter, gaussian filter and mean filter. We also experiment the filter with different kernel size. The results are shown in Tab. 5. Firstly, using filter yields a better performance on the rendering quality across all the scales. Secondly, Mip-VoG is robust to different filter while the superior performance is achieved when the mean filter with kernel size 5 is chosen. Finally, using too small or too large kernel size tends to slightly harm the filtering outcome, as size 5 surpass the other two for both mean filter and gaussian filter.

5. Conclusion

In the paper, we have presented a multiscale representation for real-time anti-aliasing rendering method. We base our work on the voxel grids representation with a deferred architecture of NeRF. We proposed to use mip voxel grids, which yields the point-wise sampling from the voxel grids of different scales according to the level of detail computed by ray differentials. To generate multiple levels of Mip-VoG, we leverage the low-pass filter and interpolation filtering to downsample the original full resolution voxel grid progressively. The final scene properties of a 3D is sampled from two neighbor voxel grids using quadrilinear interpolation. Experiments show our method effectively learns a multiscale representation from the training images and provides higher accuracy in real-time anti-aliasing rendering.

Method	PSNR \uparrow				SSIM \uparrow				LPIPS \downarrow			
	Full Res	1/2 Res	1/4 Res	1/8 Res	Full Res	1/2 Res	1/4 Res	1/8 Res	Full Res	1/2 Res	1/4 Res	1/8 Res
SNeRG [21]	29.333	30.065	28.355	25.373	0.940	0.949	0.946	0.924	0.134	0.091	0.097	0.144
MobileNeRF [10]	29.448	30.654	31.144	30.000	0.934	0.947	0.957	0.959	0.077	0.054	0.042	0.037
MobileNeRF [10] w/o SS	28.290	28.447	27.317	25.212	0.926	0.935	0.935	0.917	0.093	0.077	0.079	0.094
Ours	30.355	<u>30.467</u>	<u>28.766</u>	<u>26.566</u>	0.949	0.956	<u>0.951</u>	<u>0.935</u>	0.062	0.050	<u>0.058</u>	<u>0.073</u>

Table 3: **Quantitative results on Synthetic-NeRF.** Performance of models that trained on single scale Synthetic-NeRF but evaluated on Multiscale-NeRF. All the metrics of the scale are averaged across eight scenes.

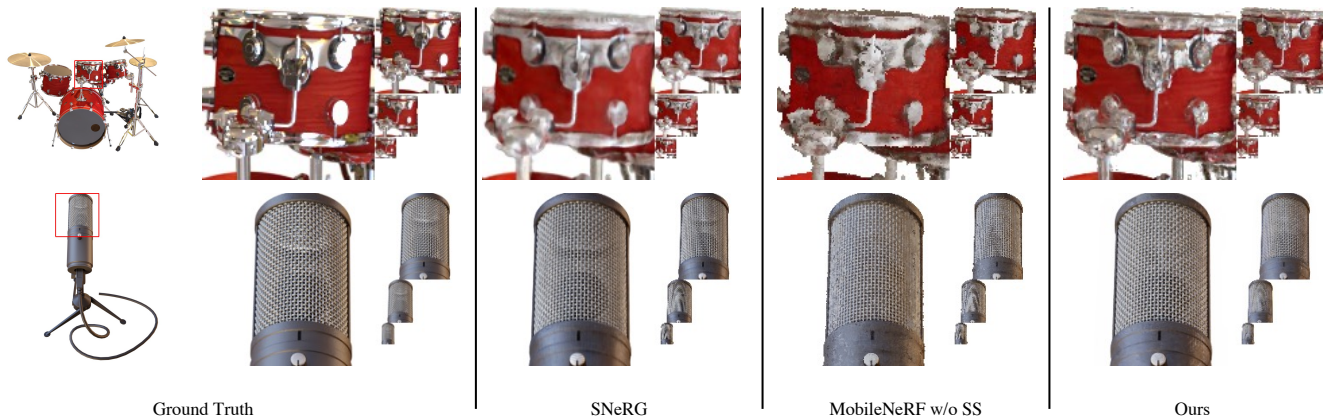


Figure 6: **Qualitative results on Synthetic-NeRF.** We demonstrate Mip-VoG rendering results compared to other baselines on the test set from two scenes, trained on single-scale dataset. One can observe that our method can reduce the aliases on the edge of drum and mic windscreen.

Method	PSNR \uparrow			
	Full Res	1/2 Res	1/4 Res	1/8 Res
Ours w/o tr-mip te-mip	29.690	30.897	30.201	27.371
Ours w/o tr-mip	29.631	30.217	29.461	27.663
Ours w/o te-mip	30.348	31.146	29.581	26.669
Ours	30.333	31.290	31.055	29.014

Table 4: **Result of mipmapping ablation.** We conduct the experiments on the Multiscale-NeRF with respectively dropping the Mip-VoG in training and testing, denoted as “tr-mip” and “te-mip” in the table.

Filter	PSNR \uparrow			
	Full Res	1/2 Res	1/4 Res	1/8 Res
None	29.995	31.043	30.718	28.534
Mean (size 3)	30.211	31.179	30.997	29.005
Mean (size 5)	30.333	31.290	31.055	29.014
Mean (size 7)	30.210	31.177	30.995	29.002
Gaussian (size 3)	30.011	31.068	30.733	28.575
Gaussian (size 5)	30.052	31.078	30.759	28.683
Gaussian (size 7)	30.005	31.059	30.731	28.505

Table 5: **Low-pass filter ablation.** We demonstrate the sensitivity of the low-pass filter with respect to the filter type and the kernel size.

Our main bottleneck of rendering speed is the computation burden of Level of Detail within the web shader. We sacrifice the speed of original SNeRG formulation to provide a multiscale representation. Nonetheless, there is room for improvement through future engineering endeavors, including potential collaboration with state-of-the-art voxel-based real-time rendering engines like [47]. While offering a pioneering multiscale representation for real-time applications, we hope our approach will be valuable to future research on multiscale training and real-time anti-aliasing rendering for neural rendering models.

Acknowledgements This research was mainly undertaken using the LIEF HPC-GPGPU Facility hosted at the University of Melbourne. This Facility was established with the assistance of LIEF Grant LE170100200. This research was also partially supported by the Research Computing Services NCI Access scheme at the University of Melbourne. DH was supported by Melbourne Research Scholarship from the University of Melbourne.

References

- [1] Tomas Akenine-Möller, Jim Nilsson, Magnus Andersson, Colin Barré-Brisebois, Robert Toth, and Tero Karras. Texture level of detail strategies for real-time ray tracing. In *Ray Tracing Gems*, pages 321–345. Springer, 2019. [5](#)
- [2] Benjamin Attal, Jia-Bin Huang, Michael Zollhöfer, Johannes Kopf, and Changil Kim. Learning neural light fields with ray-space embedding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19819–19829, 2022. [2](#)
- [3] 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. [1](#), [2](#), [3](#), [6](#), [7](#), [8](#)
- [4] Jonathan T. Barron, Ben Mildenhall, Dor Verbin, Pratul P. Srinivasan, and Peter Hedman. Zip-nerf: Anti-aliased grid-based neural radiance fields. *ICCV*, 2023. [3](#)
- [5] Sai Bi, Zexiang Xu, Pratul Srinivasan, Ben Mildenhall, Kalyan Sunkavalli, Miloš Hašan, Yannick Hold-Geoffroy, David Kriegman, and Ravi Ramamoorthi. Neural reflectance fields for appearance acquisition. *arXiv preprint arXiv:2008.03824*, 2020. [2](#)
- [6] Eric Bruneton and Fabrice Neyret. A survey of nonlinear pre-filtering methods for efficient and accurate surface shading. *IEEE Transactions on Visualization and Computer Graphics*, 18(2):242–260, 2011. [3](#)
- [7] Chris Buehler, Michael Bosse, Leonard McMillan, Steven Gortler, and Michael Cohen. Unstructured lumigraph rendering. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, pages 425–432, 2001. [2](#)
- [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] Anpei Chen, Zexiang Xu, Andreas Geiger, Jingyi Yu, and Hao Su. Tensorf: Tensorial radiance fields. *arXiv preprint arXiv:2203.09517*, 2022. [2](#)
- [10] Zhiqin Chen, Thomas Funkhouser, Peter Hedman, and Andrea Tagliasacchi. Mobilenerf: Exploiting the polygon rasterization pipeline for efficient neural field rendering on mobile architectures. In *The Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. [1](#), [3](#), [7](#), [8](#), [9](#)
- [11] Cyril Crassin, Fabrice Neyret, Sylvain Lefebvre, Miguel Sainz, and Elmar Eisemann. Beyond triangles: gigavoxels effects in video games. In *SIGGRAPH 2009: Talks*, page 78. ACM, 2009. [3](#)
- [12] Abe Davis, Marc Levoy, and Fredo Durand. Unstructured light fields. In *Computer Graphics Forum*, number 2pt1, pages 305–314. Wiley Online Library, 2012. [2](#)
- [13] Helisa Dhama, Keisuke Tateno, Iro Laina, Nassir Navab, and Federico Tombari. Peeking behind objects: Layered depth prediction from a single image. *Pattern Recognition Letters*, 125:333–340, 2019. [2](#)
- [14] Jon P Ewins, Marcus D Waller, Martin White, and Paul F Lister. Mip-map level selection for texture mapping. *IEEE Transactions on Visualization and Computer Graphics*, 4(4):317–329, 1998. [2](#), [3](#)
- [15] John Flynn, Michael Broxton, Paul Debevec, Matthew DuVall, Graham Fyffe, Ryan Overbeck, Noah Snavely, and Richard Tucker. Deepview: View synthesis with learned gradient descent. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2367–2376, 2019. [2](#)
- [16] 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](#)
- [17] Stephan J Garbin, Marek Kowalski, Matthew Johnson, Jamie Shotton, and Julien Valentin. Fastnerf: High-fidelity neural rendering at 200fps. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14346–14355, 2021. [2](#)
- [18] Steven J Gortler, Radek Grzeszczuk, Richard Szeliski, and Michael F Cohen. The lumigraph. In *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*, pages 43–54, 1996. [2](#)
- [19] Jon Hasselgren, Nikolai Hofmann, and Jacob Munkberg. Shape, light, and material decomposition from images using monte carlo rendering and denoising. *Advances in Neural Information Processing Systems*, 35:22856–22869, 2022. [2](#)
- [20] Paul S Heckbert et al. Texture mapping polygons in perspective. Technical report, Citeseer, 1983. [5](#)
- [21] Peter Hedman, Pratul P Srinivasan, Ben Mildenhall, Jonathan T Barron, and Paul Debevec. Baking neural radiance fields for real-time view synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5875–5884, 2021. [1](#), [2](#), [4](#), [6](#), [7](#), [9](#)
- [22] Homan Igehy. Tracing ray differentials. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*, pages 179–186, 1999. [2](#), [5](#)
- [23] Anton S Kaplanyan, Stephan Hill, Anjul Patney, and Aaron E Lefohn. Filtering distributions of normals for shading anti-aliasing. In *High Performance Graphics*, pages 151–162, 2016. [3](#)
- [24] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. [6](#)
- [25] Sylvain Lefebvre and Hugues Hoppe. Parallel controllable texture synthesis. In *ACM SIGGRAPH 2005 Papers*, pages 777–786. 2005. [5](#)
- [26] Anat Levin and Fredo Durand. Linear view synthesis using a dimensionality gap light field prior. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 1831–1838. IEEE, 2010. [2](#)
- [27] Marc Levoy and Pat Hanrahan. Light field rendering. In *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*, pages 31–42, 1996. [2](#)
- [28] 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. 2
- [29] Zhengqi Li, Wenqi Xian, Abe Davis, and Noah Snavely. Crowdsampling the plenoptic function. In *European Conference on Computer Vision*, pages 178–196. Springer, 2020. 2
- [30] Steven Liu, Xiuming Zhang, Zhoutong Zhang, Richard Zhang, Jun-Yan Zhu, and Bryan Russell. Editing conditional radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5773–5783, 2021. 2
- [31] 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. 2
- [32] 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
- [33] Nelson Max. Optical models for direct volume rendering. *IEEE Transactions on Visualization and Computer Graphics*, 1(2):99–108, 1995. 3
- [34] Ben Mildenhall, Pratul P Srinivasan, Rodrigo Ortiz-Cayon, Nima Khademi Kalantari, Ravi Ramamoorthi, Ren Ng, and Abhishek Kar. Local light field fusion: Practical view synthesis with prescriptive sampling guidelines. *ACM Transactions on Graphics (TOG)*, 38(4):1–14, 2019. 2
- [35] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 1, 2, 3, 6
- [36] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multi-resolution hash encoding. *arXiv preprint arXiv:2201.05989*, 2022. 2, 3
- [37] Jacob Munkberg, Jon Hasselgren, Tianchang Shen, Jun Gao, Wenzheng Chen, Alex Evans, Thomas Müller, and Sanja Fidler. Extracting triangular 3d models, materials, and lighting from images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8280–8290, 2022. 2
- [38] Baptiste Nicolet, Alec Jacobson, and Wenzel Jakob. Large steps in inverse rendering of geometry. *ACM Transactions on Graphics (TOG)*, 40(6):1–13, 2021. 2
- [39] Michael Niemeyer and Andreas Geiger. Giraffe: Representing scenes as compositional generative neural feature fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11453–11464, 2021. 2
- [40] Marc Olano and Dan Baker. Lean mapping. In *Proceedings of the 2010 ACM SIGGRAPH symposium on Interactive 3D Graphics and Games*, pages 181–188, 2010. 3
- [41] Julian Ost, Fahim Mannan, Nils Thuerey, Julian Knodt, and Felix Heide. Neural scene graphs for dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2856–2865, 2021. 2
- [42] 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. 2
- [43] 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. 2
- [44] 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. 2
- [45] Daniel Rebain, Wei Jiang, Soroosh Yazdani, Ke Li, Kwang Moo Yi, and Andrea Tagliasacchi. Derf: Decomposed radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14153–14161, 2021. 2
- [46] 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. 2
- [47] Christian Reiser, Rick Szeliski, Dor Verbin, Pratul Srinivasan, Ben Mildenhall, Andreas Geiger, Jon Barron, and Peter Hedman. Merf: Memory-efficient radiance fields for real-time view synthesis in unbounded scenes. *ACM Transactions on Graphics (TOG)*, 42(4):1–12, 2023. 9
- [48] Viktor Rudnev, Mohamed Elgharib, William Smith, Lingjie Liu, Vladislav Golyanik, and Christian Theobalt. Nerf for outdoor scene relighting. In *Computer Vision—ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XVI*, pages 615–631. Springer, 2022. 2
- [49] Katja Schwarz, Yiyi Liao, Michael Niemeyer, and Andreas Geiger. Graf: Generative radiance fields for 3d-aware image synthesis. *Advances in Neural Information Processing Systems*, 33:20154–20166, 2020. 2
- [50] Jonathan Shade, Steven Gortler, Li-wei He, and Richard Szeliski. Layered depth images. In *Proceedings of the 25th annual conference on Computer graphics and interactive techniques*, pages 231–242, 1998. 2
- [51] Lixin Shi, Haitham Hassanieh, Abe Davis, Dina Katabi, and Fredo Durand. Light field reconstruction using sparsity in the continuous fourier domain. *ACM Transactions on Graphics (TOG)*, 34(1):1–13, 2014. 2
- [52] Meng-Li Shih, Shih-Yang Su, Johannes Kopf, and Jia-Bin Huang. 3d photography using context-aware layered depth inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8028–8038, 2020. 2
- [53] Vincent Sitzmann, Semon Rezkikov, Bill Freeman, Josh Tenenbaum, and Fredo Durand. Light field networks:

- Neural scene representations with single-evaluation rendering. *Advances in Neural Information Processing Systems*, 34:19313–19325, 2021. 2
- [54] Vincent Sitzmann, Justus Thies, Felix Heide, Matthias Nießner, Gordon Wetzstein, and Michael Zollhofer. Deepvoxels: Learning persistent 3d feature embeddings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2437–2446, 2019. 2
- [55] Pratul P Srinivasan, Boyang Deng, Xiuming Zhang, Matthew Tancik, Ben Mildenhall, and Jonathan T Barron. Nerv: Neural reflectance and visibility fields for relighting and view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7495–7504, 2021. 2
- [56] Pratul P Srinivasan, Richard Tucker, Jonathan T Barron, Ravi Ramamoorthi, Ren Ng, and Noah Snavely. Pushing the boundaries of view extrapolation with multiplane images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 175–184, 2019. 2
- [57] Cheng Sun, Min Sun, and Hwann-Tzong Chen. Direct voxel grid optimization: Super-fast convergence for radiance fields reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5459–5469, 2022. 2, 3, 6
- [58] Edgar Tretschk, Ayush Tewari, Vladislav Golyanik, Michael Zollhöfer, Christoph Lassner, and Christian Theobalt. Non-rigid neural radiance fields: Reconstruction and novel view synthesis of a dynamic scene from monocular video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12959–12970, 2021. 2
- [59] Alex Trevithick and Bo Yang. Grf: Learning a general radiance field for 3d representation and rendering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15182–15192, 2021. 2
- [60] Shubham Tulsiani, Richard Tucker, and Noah Snavely. Layer-structured 3d scene inference via view synthesis. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 302–317, 2018. 2
- [61] 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. 2
- [62] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 6
- [63] Turner Whitted. An improved illumination model for shaded display. In *ACM Siggraph 2005 Courses*, pages 4–es. 2005. 3
- [64] Lance Williams. Pyramidal parametrics. In *Proceedings of the 10th annual conference on Computer graphics and interactive techniques*, pages 1–11, 1983. 3, 5
- [65] Suttisak Wizadwongsa, Pakkapon Phongthawee, Jiraphon Yenphraphai, and Supasorn Suwajanakorn. Nex: Real-time view synthesis with neural basis expansion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8534–8543, 2021. 2
- [66] Lifan Wu, Shuang Zhao, Ling-Qi Yan, and Ravi Ramamoorthi. Accurate appearance preserving prefiltering for rendering displacement-mapped surfaces. *ACM Transactions on Graphics (TOG)*, 38(4):1–14, 2019. 3
- [67] Bangbang Yang, Chong Bao, Junyi Zeng, Hujun Bao, Yinda Zhang, Zhaopeng Cui, and Guofeng Zhang. Neumesh: Learning disentangled neural mesh-based implicit field for geometry and texture editing. In *Computer Vision—ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XVI*, pages 597–614. Springer, 2022. 2
- [68] Alex Yu, Ruilong Li, Matthew Tancik, Hao Li, Ren Ng, and Angjoo Kanazawa. Plenotrees for real-time rendering of neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5752–5761, 2021. 2
- [69] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelnerf: Neural radiance fields from one or few images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4578–4587, 2021. 2
- [70] Yu-Jie Yuan, Yang-Tian Sun, Yu-Kun Lai, Yüwen Ma, Rongfei Jia, and Lin Gao. Nerf-editing: geometry editing of neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18353–18364, 2022. 2
- [71] 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
- [72] Xiuming Zhang, Pratul P Srinivasan, Boyang Deng, Paul Debevec, William T Freeman, and Jonathan T Barron. Nerfactor: Neural factorization of shape and reflectance under an unknown illumination. *ACM Transactions on Graphics (TOG)*, 40(6):1–18, 2021. 2
- [73] Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images. *arXiv preprint arXiv:1805.09817*, 2018. 2