MERF: Memory-Efficient Radiance Fields for Real-time View Synthesis in Unbounded Scenes

CHRISTIAN REISER*, University of Tübingen, Tübingen AI Center, Germany and Google Research, United Kingdom
RICHARD SZELISKI, Google Research, United States of America
DOR VERBIN, Google Research, United States of America
PRATUL P. SRINIVASAN, Google Research, United States of America
BEN MILDENHALL, Google Research, United States of America
ANDREAS GEIGER, University of Tübingen, Tübingen AI Center, Germany
JONATHAN T. BARRON, Google Research, United States of America
PETER HEDMAN, Google Research, United Kingdom

Interactive web demo and code at https://creiser.github.io/merf

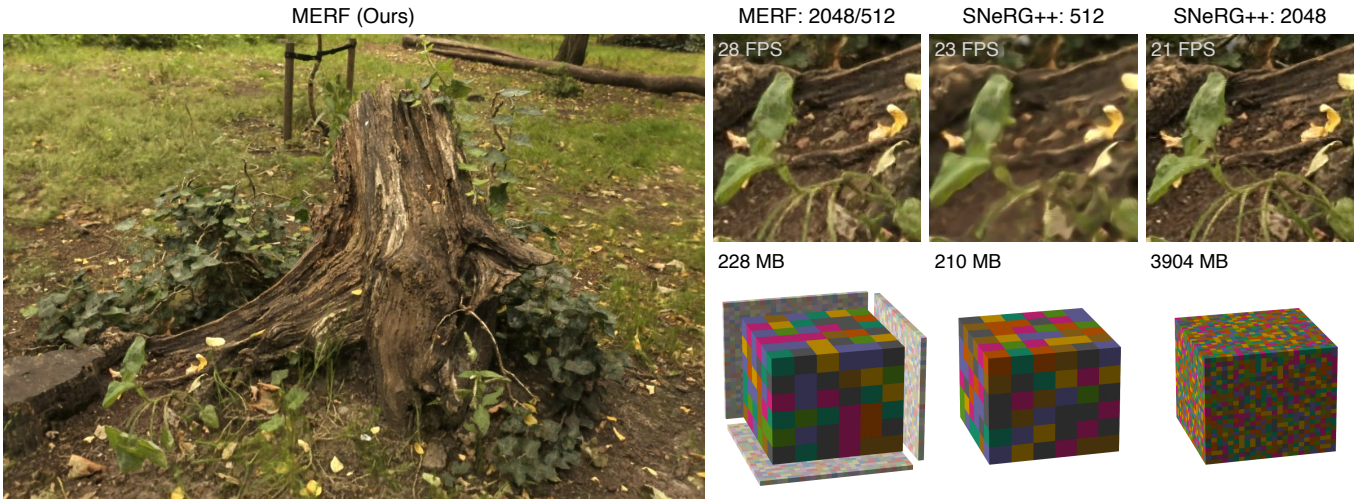


Fig. 1. Our method, *MERF*, allows for real-time view synthesis in the browser (framerates shown here are measured on a Macbook Pro M1). We design a volumetric scene representation that is first optimized to maximize rendering quality and then losslessly baked into a more efficient format for fast rendering. As depicted above, this representation uses a combination of a low-resolution 3D grid and a set of higher-resolution 2D planes (spatial extent of each visualization signifies memory consumption). Compared to prior real-time view synthesis methods such as SNeRG++ (our improved version of Hedman et al. [2021]), MERF achieves better image quality for the same amount of compression, matching the output of a SNeRG++ model that is over 17× larger.

Neural radiance fields enable state-of-the-art photorealistic view synthesis. However, existing radiance field representations are either too compute-intensive for real-time rendering or require too much memory to scale to large scenes. We present a Memory-Efficient Radiance Field (MERF) representation that achieves real-time rendering of large-scale scenes in a browser. MERF reduces the memory consumption of prior sparse volumetric radiance fields using a combination of a sparse feature grid and high-resolution 2D feature planes. To support large-scale unbounded scenes, we introduce a novel contraction function that maps scene coordinates into a bounded volume while still allowing for efficient ray-box intersection. We design a lossless procedure for baking the parameterization used during training into a model that achieves real-time rendering while still preserving the photorealistic view synthesis quality of a volumetric radiance field.

CCS Concepts: • **Computing methodologies** → **Reconstruction**; *Neural networks*; *Volumetric models*.

Additional Key Words and Phrases: Neural Radiance Fields, Volumetric Representation, Image Synthesis, Real-Time Rendering, Deep Learning.

ACM Reference Format:
Christian Reiser, Richard Szeliski, Dor Verbin, Pratul P. Srinivasan, Ben Mildenhall, Andreas Geiger, Jonathan T. Barron, and Peter Hedman. 2023. MERF: Memory-Efficient Radiance Fields for Real-time View Synthesis in

*Work done while interning at Google.

SIGGRAPH 2023, August 2023, Los Angeles, CA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/3592426

Unbounded Scenes. In . ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3592426

1 INTRODUCTION

Neural volumetric scene representations such as Neural Radiance Fields (NeRF) [Mildenhall et al. 2020] enable photorealistic novel view synthesis of scenes with complex geometry and appearance, but the compute required to query such neural representations during volumetric raymarching prohibits real-time rendering. Subsequent works have proposed discretized volumetric representations that can substantially increase rendering performance [Garbin et al. 2021; Hedman et al. 2021; Yu et al. 2021], but these approaches do not yet enable practical real-time rendering for *large-scale* scenes.

These representations struggle to scale to larger scenes primarily due to graphics hardware constraints. Volumetric data is necessarily larger than a 2D surface representation and occupies more space in memory. Similarly, while a camera ray intersects a hard surface at most once, rendering a ray through a volume may require many samples. With state of the art neural or hybrid representations, each of these sample queries is very expensive to evaluate, either in terms of compute or memory bandwidth. As a result, methods that work for scenes with limited extent (single objects in space or forward-facing scenes) typically do not scale up to larger unbounded scenes.

Neural or hybrid volumetric methods must address two fundamental trade-offs that arise from these constraints:

- *Volume vs. surface?* Purely volumetric rendering models are most amenable to gradient-based optimization and produce excellent view synthesis results [Barron et al. 2022]. On the other hand, increasing sparsity and moving closer [Wang et al. 2021; Yariv et al. 2021] or completely [Chen et al. 2023; Munkberg et al. 2022b] to a surface-like representation degrades image quality but results in compact representations that are cheap to render.
- *Memory bound vs. compute bound?* The most compact representations (such as the MLP network in Mildenhall et al. [2020] or the low-rank decomposition in Chen et al. [2022]) require many FLOPS to query, and the fastest representations (such as the sparse 3D data structures used in Yu et al. [2021] and Hedman et al. [2021]) consume large amounts of graphics memory.

One approach to this trade-off is to embrace a slower, more compact volumetric model for optimization and to subsequently "bake" it into a larger but faster representation for rendering. However, baking often affects the representation or rendering model, which can lead to a large drop in image quality. Though this can partially be ameliorated by fine-tuning the baked representation, fine-tuning does not easily scale to larger scenes, as computing gradients for optimization requires significantly more memory than rendering.

The goal of our work is to find a representation that is well suited for both optimization and fast rendering. Our solution is a single unified radiance field representation with two different underlying *parameterizations*. In both stages, our memory-efficient radiance field (MERF) is defined using a combination of a voxel grid [Sun et al. 2022; Yu et al. 2022] and triplane data structure [Chan et al. 2022]. During optimization, we use the NGP hash grid structure [Müller et al. 2022] to compress our parameterization, which allows for differentiable sparsification and provides an inductive bias that aids convergence. After optimization, we query the recovered NGP to explicitly bake out the MERF and create a binary occupancy grid to accelerate rendering. Critically, both the NGP-parameterized and baked MERF *represent the same underlying radiance field function*. This means that the high quality achieved by the optimized MERF carries over to our real-time browser-based rendering engine.

2 RELATED WORK

As our goal is real-time view synthesis in large unbounded scenes, this discussion is focused on approaches that accelerate rendering or reconstruct large spaces. For a comprehensive overview of recent view synthesis approaches, please refer to Tewari et al. [2022].

Early methods for real-time large-scale view synthesis either captured a large number of images and interpolated them with optical flow [Aliaga et al. 2002] or relied heavily on hand-made geometry proxies [Buehler et al. 2001; Debevec et al. 1998]. Later techniques used inaccurate, but automatically reconstructed geometry proxies [Chaurasia et al. 2013], and relied on screen-space neural networks to compensate for this [Hedman et al. 2018; Rückert et al. 2022]. Neural Radiance Fields (NeRF) [Mildenhall et al. 2020] facilitated higher quality reconstructions by representing the full scene volume as a multi-layer perceptron (MLP). This volumetric representation can easily model thin structures and semi-transparent objects and is also well-suited to gradient-based optimization.

NeRF was quickly extended to reconstruct large scenes, using crowdsourced data [Martin-Brualla et al. 2021], tiling the space with NeRF networks [Tancik et al. 2022; Turki et al. 2022], and reconstructing the scene in a warped domain where far-away regions are compressed [Barron et al. 2022; Zhang et al. 2020]. Later, fast radiance field reconstruction was achieved by representing the scene as a grid, stored either densely [Yu et al. 2022], as latent features to be decoded [Karnewar et al. 2022; Liu et al. 2020; Sun et al. 2022], or as latent hash grids [Müller et al. 2022] implemented with specialized CUDA kernels [Li et al. 2022d]. While this dramatically reduces reconstruction time, accurate real-time rendering of large scenes has not yet been demonstrated at high resolutions.

Other methods addressed real-time rendering by precomputing and storing (i.e. *baking*) NeRF's view-dependent colors and opacities in volumetric data structures [Garbin et al. 2021; Hedman et al. 2021; Yu et al. 2021; Zhang et al. 2022] or by splitting the scene into voxels and representing each voxel with a small separate MLP [Reiser et al. 2021]. However, these representations consume a lot of graphics memory and are thus limited to objects, not scenes. Furthermore, these methods incur a quality loss during baking due to the mismatch between the slower rendering procedure used for training and the real-time rendering procedure used for inference.

Alternatively, faster rendering can be achieved by extending the network to work with ray segments rather than points [Lindell et al. 2021; Wu et al. 2022a] or by training a separate sampling network [Barron et al. 2022; Kurz et al. 2022; Neff et al. 2021; Piala and Clark 2021]. However, these approaches have not achieved real-time rates at high resolutions, likely because they require evaluating an MLP for each sample along a ray. Light field coordinates circumvent this problem and require just one MLP evaluation per ray [Attal et al. 2022; Cao et al. 2023; Li et al. 2022c; Sitzmann et al. 2021; Wang et al.

2022b]. However, like traditional light field representations [Gortler et al. 1996; Levoy and Hanrahan 1996], this approach has only been demonstrated to work well within small viewing volumes. Similarly, multi-plane image [Flynn et al. 2019; Mildenhall et al. 2019; Wizadwongsa et al. 2021; Zhou et al. 2018] or multi-sphere image [Attal et al. 2020; Broxton et al. 2020] representations map well to graphics hardware and can be rendered in real-time, but also support only restricted camera motion.

It is also possible to speed up NeRF rendering by post-processing the output image with a convolutional neural network. This makes it possible to perform an expensive volumetric rendering step at a lower resolution and then upsample that result to the final desired resolution [Li et al. 2022a; Wang et al. 2022a]. Wu et al. [2022b] combined this approach with baking and showed high-quality real-time rendering of large scenes. However, to achieve this, they required a 3D scan of the scene as input, and they used a CUDA implementation designed for workstation-class hardware. In contrast, our method only needs posed images as input and runs in a browser on commodity hardware such as laptops.

Mesh-based approaches achieve the fastest rendering. These methods either directly optimize a mesh [Munkberg et al. 2022a], extract a mesh after training [Kellnhofer et al. 2021; Yariv et al. 2023] and/or restrict NeRF evaluation to planes or polygons [Chen et al. 2023; Lin et al. 2022]. While these surface-based techniques are faster than our volume-based one, their view synthesis quality is lower. Fuzzy materials, thin structures like hair or foliage, or underdetermined regions (e.g. background) are often not captured well by surface-based approaches.

The problem of compressing NeRF reconstructions has also been explored in prior work. Several methods achieve this by post-processing an existing reconstruction through incremental pruning [Deng and Tartaglione 2023] with vector quantization [Li et al. 2022b]. Takikawa et al. [2022] directly optimize for a compressed codebook-based representation of the scene. While these methods all report impressive compression ratios, they all rely on evaluating an MLP for each volume sample and are therefore too slow for real-time rendering of large scenes.

Our approach works by projecting 3D samples onto three 2D projections that correspond to the cardinal axes. Similar representations, often referred to as *tri-planes*, have been explored for surface reconstruction from point clouds [Peng et al. 2020] and generative modelling of 3D scenes [DeVries et al. 2021] or faces [Chan et al. 2022]. Recently TensoRF [Chen et al. 2022] use tri-planes for NeRF reconstruction. TensorRF decomposes the 3D scene volume into a sum of vector-matrix outer products, which makes it possible to directly train a compressed and high quality radiance field. However, TensoRF trades off memory footprint for more expensive queries that involve a large matrix multiplication. Our representation significantly speeds up the query time by removing the need for the matrix product while simultaneously halving the memory bandwidth consumption.

3 PRELIMINARIES

We begin with a short review of relevant prior work on radiance fields for unbounded scenes. A radiance field maps every 3D position $\mathbf{x} \in \mathbb{R}^3$ and viewing direction $\mathbf{d} \in \mathbb{S}^2$ to the volumetric density $\tau \in \mathbb{R}_+$ at that location and the RGB color emitted from it along the view direction, $\mathbf{c} \in \mathbb{R}^3$. The color of the ray emitted from point $\mathbf{o}$ in the direction $\mathbf{d}$ can then be computed using the radiance field by sampling points along the ray, $\mathbf{x}_i = \mathbf{o} + t_i\mathbf{d}$, and compositing the corresponding densities $\{\tau_i\}$ and colors $\{\mathbf{c}_i\}$ according to the numerical quadrature approach of Max [1995]:

$$\mathbf{C} = \sum_i w_i \mathbf{c}_i\,, \quad w_i = \alpha_i T_i\,, \quad T_i = \prod_{j=1}^{i-1}(1-\alpha_j)\,, \quad \alpha_i = 1 - e^{-\tau_i \delta_i}\,, \tag{1}$$

where T_i and α_i denote transmittance and alpha values of sample i, and $\delta_i = t_{i+1} - t_i$ is the distance between adjacent samples.

The original NeRF work parameterized a radiance field using a Multilayer Perceptron (MLP), which outputs the volume density and view-dependent color for any continuous 3D location. In order to reduce the number of MLP evaluations to one per ray, SNeRG uses a deferred shading model in which the radiance field is decomposed into a 3D field of densities τ, diffuse RGB colors $\mathbf{c_d}$, and feature vectors $\mathbf{f}$ [Hedman et al. 2021]. SNeRG's deferred rendering model volumetrically accumulates the diffuse colors $\{\mathbf{c}_{d,i}\}$ and features $\{\mathbf{f}_i\}$ along the ray, similar to Equation 1:

$$\mathbf{C}_d = \sum_i w_i \mathbf{c}_{d,i}, \quad \mathbf{F} = \sum_i w_i \mathbf{f}_i, \tag{2}$$

and computes the ray's color as the sum of the accumulated diffuse color $\mathbf{C}_d$ and the view-dependent color computed using a small MLP h that takes as input $\mathbf{C}_d$, $\mathbf{F}$, and the viewing direction $\mathbf{d}$:

$$\mathbf{C} = \mathbf{C}_d + h(\mathbf{C}_d, \mathbf{F}, \mathbf{d}). \tag{3}$$

SNeRG uses a large MLP during training and bakes it after convergence into a block-sparse grid for real-time rendering.

In order for radiance fields to render high quality unbounded scenes containing nearby objects as well as objects far from the camera, mip-NeRF 360 [Barron et al. 2022] uses a contraction function to warp the unbounded scene domain into a finite sphere:

$$\text{contract}(\mathbf{x}) = \begin{cases} \mathbf{x} & \text{if } \|\mathbf{x}\|_2 \le 1 \\ \left(2 - \frac{1}{\|\mathbf{x}\|_2}\right)\frac{\mathbf{x}}{\|\mathbf{x}\|_2} & \text{if } \|\mathbf{x}\|_2 > 1 \end{cases} \tag{4}$$

4 SCENE REPRESENTATION

In this section, we describe the MERF scene representation, which is designed to enable real-time volumetric rendering of unbounded scenes while maintaining a low memory footprint.

4.1 Volume Parameterization

MERF represents a scene using a 3D field of volume densities $\tau \in \mathbb{R}_+$, diffuse RGB colors $\mathbf{c_d} \in \mathbb{R}^3$, and feature vectors $\mathbf{f} \in \mathbb{R}^K$, as shown in Figure 2. These quantities are rendered using the deferred shading model from SNeRG, described in Section 3.

We parameterize this field with a low-resolution 3D $L \times L \times L$ voxel grid V and three high-resolution 2D $R \times R$ grids P_x, P_y, and P_z, one for each of the cardinal yz, xz, and xy planes. Each element of the low-resolution 3D grid and the three high-resolution 2D grids stores a vector with $C = 4 + K$ channels In our experiments, we use $C = 8$ and default to $L = 512$ and $R = 2048$.

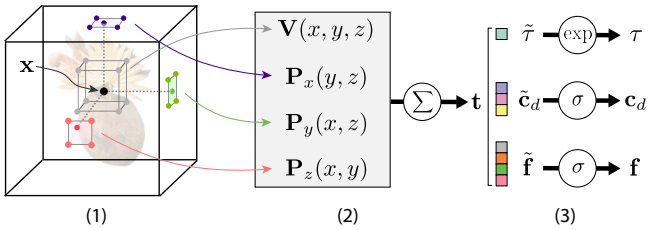


Fig. 2. Our scene representation. For a location $\mathbf{x}$ along a ray: (1) We query its eight neighbors on a low-resolution 3D grid; and we project it onto each of the three axis-aligned planes, and then query each projection's four neighbors on a high-resolution 2D grid. (2) The eight low-resolution 3D neighbors are evaluated and trilinearly interpolated while the three sets of four high-resolution 2D neighbors are evaluated and bilinearly interpolated, and the resulting features are summed into a single feature vector $\mathbf{t}$. (3) The feature vector is split and nonlinearly mapped into three components: density τ, RGB color $\mathbf{c}_d$, and a feature vector $\mathbf{f}$ encoding view dependence effects.

We define the continuous field of C-vectors as the sum of trilinearly interpolated vectors from the 3D grid and bilinearly interpolated vectors from the three 2D grids:

$$\mathbf{t}(x, y, z) = \mathbf{V}(x, y, z) + \mathbf{P}_x(y, z) + \mathbf{P}_y(x, z) + \mathbf{P}_z(x, y), \tag{5}$$

where $\mathbf{V}\colon \mathbb{R}^3 \to \mathbb{R}^C$ is a trilinear interpolation operator using the 3D grid values, and $\mathbf{P}_i\colon \mathbb{R}^2 \to \mathbb{R}^C$ is a bilinear interpolation operator using the grid perpendicular to the ith axis, for $i \in \{x, y, z\}$.

We split the C-vector at any 3D location into three components corresponding to density $\tilde{\tau} \in \mathbb{R}$, diffuse color $\tilde{\mathbf{c}}_d \in \mathbb{R}^3$, and view-dependence feature $\tilde{\mathbf{f}} \in \mathbb{R}^K$, and then apply nonlinear functions to obtain the three values:

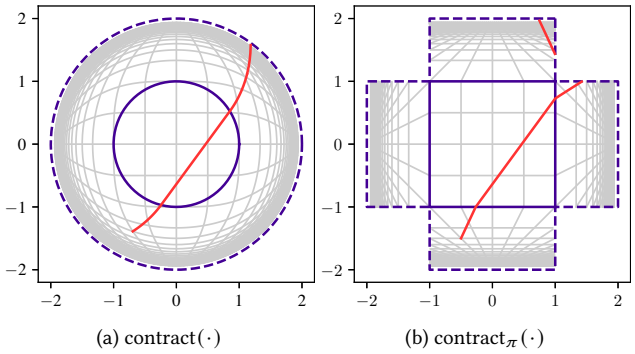
$$\tau = \exp(\tilde{\tau}), \quad \mathbf{c_d} = \sigma(\tilde{\mathbf{c}}_d), \quad \mathbf{f} = \sigma(\tilde{\mathbf{f}}), \tag{6}$$

where σ is the standard logistic sigmoid function, which constrains colors and features to lie within $(0, 1)$. Note that we apply the nonlinearities after interpolation and summation, which has been shown to greatly increase the representational power of grid representations [Karnewar et al. 2022; Sun et al. 2022].

4.2 Piecewise-projective Contraction

Mip-NeRF 360 [Barron et al. 2022] demonstrated the importance of applying a contraction function to input coordinates when representing large-scale scenes with unbounded extent. The contraction maps large far-away regions of space into small regions in contracted space, which has the effect of allocating model capacity towards representing high-resolution content near the input camera locations.

The contraction function used in mip-NeRF 360 (Equation 4) nonlinearly maps any point in space $\mathbf{x} \in \mathbb{R}^3$ into a radius-2 ball and represents the scene within this contracted space. While mip-NeRF 360's contraction function is effective and efficient to evaluate, it cannot be easily used in a real-time rendering pipeline with discretized voxels. This is because empty space skipping is critical for efficient volume rendering, and requires a method for analytically computing the intersection between a ray and the axis-aligned



(a) contract$(\cdot)$ (b) contract$_\pi(\cdot)$

Fig. 3. A 2D visualization of (a) the spherical contraction function used by Barron et al. [2022]; and (b) our piecewise projective contraction, both applied to the same ray (in red). The spherical contraction maps a straight line to a curve, which makes empty space skipping more expensive, while our contraction maps a straight line to a small number of line segments. The gray lines show the result of applying each contraction function to a regular grid.

bounding box (AABB) of any "active" (i.e., occupied) content. As can be seen in Figure 3a, mip-NeRF 360's contraction function does not preserve straight lines and thereby makes ray-AABB intersections challenging to compute.

To address this, we propose a novel contraction function for which ray-AABB intersections can be computed trivially. The jth coordinate of a contracted point is defined as follows:

$$\text{contract}_\pi(\mathbf{x})_j = \begin{cases} x_j & \text{if } \|\mathbf{x}\|_\infty \le 1 \\ \frac{x_j}{\|\mathbf{x}\|_\infty} & \text{if } x_j \ne \|\mathbf{x}\|_\infty > 1 \\ \left(2 - \frac{1}{|x_j|}\right)\frac{x_j}{|x_j|} & \text{if } x_j = \|\mathbf{x}\|_\infty > 1 \end{cases}, \tag{7}$$

where $\|\cdot\|_\infty$ is the L_∞ norm ($\|\mathbf{x}\|_\infty = \max_j |x_j|$). This contraction function is piecewise-projective: it contains seven regions, and within each region, it is a projective transformation. The unit cube $\|\mathbf{x}\|_\infty$ is preserved (i.e., the contraction function is the identity), and the other six regions defined by the coordinate maximizing $|x_j|$ and its sign each get mapped by a different projective transformation. Because projective transformations preserve straight lines, any contracted ray must be piecewise-linear, and the only discontinuities in its direction are on the boundaries between the seven regions (see Figure 3b for a 2D example with 5 regions). The origin and direction of a ray can therefore be computed in contracted space, which allows us to use standard ray-AABB intersection tests. Table 2 (c) shows that our proposed contraction function performs on par with the original spherical contraction.

Our contraction function is inspired by the normalized device coordinate (NDC) mapping [Mildenhall et al. 2020] of a camera, but with the far plane at infinity. This mapping takes a point in an unbounded domain - the camera's frustum - and maps it to the bounded NDC space. To extend this idea to 360° scenes, we mentally instanitate six cameras that face either in the positive or negative direction of the cardinal axes. Each camera has a horizontal and vertical field of view of 90° degree and the camera's near planes are at 1. As a result the unit cube is not penetrated by any camera

Table 1. Quantitative results of our model on all scenes from Mip-NeRF 360 [Barron et al. 2022]. Models that render in real-time are highlighted. Mobile-NeRF did not evaluate on the "indoor" scenes, so those metrics are absent.

	Outdoor			Indoor		
	PSNR ↑	SSIM ↑	LPIPS ↓	PSNR ↑	SSIM ↑	LPIPS ↓
NeRF [Mildenhall et al. 2020]	21.46	0.458	0.515	26.84	0.790	0.370
NeRF++ [Zhang et al. 2020]	22.76	0.548	0.427	28.05	0.836	0.309
SVS [Riegler and Koltun 2021]	23.01	0.662	**0.253**	28.22	0.907	**0.160**
Mip-NeRF 360 [Barron et al. 2022]	**24.47**	**0.691**	0.283	**31.72**	**0.917**	0.180
Instant-NGP [Müller et al. 2022]	22.90	0.566	0.371	29.15	0.880	0.216
Deep Blending [Hedman et al. 2018]	21.54	0.524	0.364	26.40	0.844	0.261
Mobile-NeRF [Chen et al. 2023]	21.95	0.470	0.470	–	–	–
Ours	**23.19**	**0.616**	**0.343**	**27.80**	**0.855**	**0.271**

frustum and the camera's frusta subdivide the remaining space into six non-overlapping regions. Any point outside the unit cube is contracted by applying the NDC mapping of the camera in whose frustum the point lies.

5 TRAINING AND BAKING

In this section, we describe how to efficiently optimize a MERF and bake it into a representation that can be used for high-quality real-time rendering. To achieve our goal of not losing quality during this baking process, we take care to ensure that the baked representation and the scene representation we use during optimization both describe the same radiance field. This requires a training procedure that accounts for any baking discretization or quantization.

5.1 Efficient Training

Modeling high-resolution large-scale scenes requires high-capacity representations that may consume prohibitive amounts of memory during training. Significantly more memory is consumed during training than during rendering because training requires that intermediate activations be stored for the sake of backpropagation and that higher-precision per-parameter statistics be accumulated for the Adam optimizer. In our case, we found that training requires more than twelve times as much video memory as rendering. We thus optimize a compressed representation of MERF's low-resolution voxel grid and high-resolution 2D planes by parameterizing them as MLPs with a multi-resolution hash encoding [Müller et al. 2022].

However, baking the MLPs' outputs onto discrete grids for rendering introduces a mismatch between the representations used for training and for rendering. Prior work accounted for this by fine-tuning the baked representation [Yu et al. 2022], but fine-tuning requires the entire representation to be stored in memory and suffers from the aforementioned scalability issues. Instead, we simulate finite grid resolutions during training by querying the MLPs at virtual grid corners and interpolating the resulting outputs using bilinear interpolation for the high-resolution 2D grids and trilinear interpolation for the low-resolution voxel grid.

In addition to requiring high-capacity representations, high-resolution large-scale scenes require dense samples along rays during volume rendering, which also significantly contributes to the training memory footprint. To address this, mip-NeRF 360 introduced a hierarchical sampling technique that uses "proposal" MLPs to represent coarse versions of scene geometry. A proposal MLP maps 3D positions to density values, which are converted into probability distributions along rays that are supervised to be consistent with the densities output by the NeRF MLP. These proposal distributions are used in an iterative resampling procedure that produces a small number of samples that are concentrated around visible scene content. While this proposal MLP hierarchical sampling strategy is effective for reducing the number of samples along each ray during training, the proposal MLP is too expensive to evaluate for real-time rendering purposes. Instead, we use traditional empty space skipping during rendering, which also concentrates representation queries around surfaces. To avoid introducing a mismatch between training and rendering, we only bake content in regions considered occupied by the proposal MLP during training, as detailed in Section 5.3.

5.2 Quantization

To reduce our system's memory consumption at render time, we wish to quantize each of the C dimensions at every location in the grid to a single byte (see further discussion in Section 6). However, simply quantizing the optimized grid values after training creates mismatches between the optimized model and the one used for rendering, which leads to a drop in rendering quality as shown in Table 2 (b).

Our solution to this is to quantize the C values at every location during optimization. That is, we nonlinearly map them to lie in $[0, 1]$ using a sigmoid σ, then quantize them to a single byte using a quantization function q, and finally affinely map the result to the range $[-m, m]$, as:

$$\tilde{\mathbf{t}}' = 2m \cdot q\big(\sigma(\tilde{\mathbf{t}})\big) - m\,, \tag{8}$$

where we choose $m = 14$ for densities (which are computed using an exponential nonlinearity), and $m = 7$ for diffuse colors and features. Note that this only quantizes the values stored in the grid, and the non-linearities in Equation 6 are subsequently applied after linearly interpolating and summing these values.

We implement the byte quantization function q as:

$$q(x) = x + \not{\nabla}\left(\frac{\lfloor (2^8 - 1)x + 1/2 \rfloor}{2^8 - 1} - x\right), \tag{9}$$

where $\not\nabla(\cdot)$ is a stop-gradient, which prevents gradients from back-propagating to its input. This use of a stop-gradient allows us to obtain gradients for the non-differentiable rounding function by treating q as the identity function during the backward pass, which is referred to as the straight-through estimator, as used in [Bengio et al. 2013; Yin et al. 2019].

5.3 Baking

After training, we evaluate and store the MLP's outputs on discrete grids for real-time rendering. First, we compute a binary 3D grid $\mathbf{A}$ indicating voxels that contributed to any training image (i.e., voxels should *not* be stored if they correspond to occluded content, are not sampled by any training ray, or have low opacity). To populate $\mathbf{A}$, we render all training rays and extract from them a set of weighted points $\{(\mathbf{x}_i, w_i)\}$, where $\mathbf{x}_i$ is the point's position, and w_i is the associated volume rendering weight from Equation 1. Note that these points cluster around surfaces in the scene as they are sampled with a proposal-MLP [Barron et al. 2022].

We mark the eight voxels surrounding a given point $\mathbf{x}_i$ as occupied if both the volume rendering weight w_i and the opacity, α_i, exceed a threshold set to 0.005. To cull as aggressively as possible, we compute α_i based on the distance between samples δ_i used by the real-time renderer — recall that it steps through contracted space with a small uniform step size. As the proposal-MLP often suggests steps larger than δ_i, computing α_i this way leads to better culling. However, we still guarantee voxels which contribute a significant opacity value ($\alpha_i > 0.005$) are not culled in the final sampling scheme. Note that while the opacity α_i only depends on the density at $\mathbf{x}_i$, the weight w_i also depends on densities along the entire ray, making usage of w_i necessary to account for visibility.

We observe that the opacity check based on the real-time renderer's step size significantly decreases the fraction of the volume marked as occupied. Note that this is in addition to the sparsity already achieved by only considering voxels in locations that have been sampled by the proposal-MLP. In contrast, existing baking pipelines often do not consider the proposal-MLP and perform visibility culling with uniformly-spaced sample points. This often results in fog-like artifacts and floating blobs because the underlying 3D field can have arbitrary values in regions not sampled by the proposal-MLP. Table 2 demonstrates that our Proposal-MLP-aware baking pipeline is almost lossless.

After computing the binary grid $\mathbf{A}$, we bake the three high-resolution 2D planes and the low-resolution 3D voxel grid. Following SNeRG, we store this voxel grid in a block-sparse format, where we only store data blocks that contain occupied voxels. For empty space skipping, we create multiple lower resolution versions of the binary occupancy grid $\mathbf{A}$ with max-pooling. To reduce on-disk storage, we encode textures as PNGs.

6 REAL-TIME RENDERING

We implement our real-time viewer as a Javascript 3D (three.js) web application, based on SNeRG's implementation, where rendering is orchestrated by a single GLSL fragment shader.

For efficient ray marching, we employ a multi-resolution hierarchy of occupancy grids. The set of occupancy grids is created by max-pooling the full-resolution binary mask $\mathbf{A}$ with filter sizes 16, 32 and 128. For instance, if the base resolution is 4096, this results in occupancy grids of size 256, 128 and 32, occupying a total of 18 MB of video memory. We leverage this multi-resolution hierarchy of occupancy grids for faster space skipping. Given any sample location, we query the occupancy grids in a coarse-to-fine manner. If any level indicates the voxel as empty, we can skip the corresponding volume until the ray enters a new voxel at that level and compute the new sample location using the efficient ray-AABB intersection discussed in Section 4.2. We only access the MERF scene representation for samples where all occupancy grid levels are marked as occupied. Finally, we terminate ray marching along a ray once the transmittance value T_i (defined in Equation 1) falls below 2×10^{-4}.

To further decrease the number of memory accesses during rendering, we split textures into density and appearance (containing diffuse RGB colors and feature vectors) components. When accessing the MERF representation at any location, we first query the density component and only read the corresponding appearance component if the voxel opacity computed from the returned density is nonzero. Moreover, we obtain an additional 4× speed-up by optimizing the deferred rendering MLP. More specifically, we conduct loop unrolling, modify the memory layout to facilitate linear accesses, and exploit fast *mat4*-multiplication.

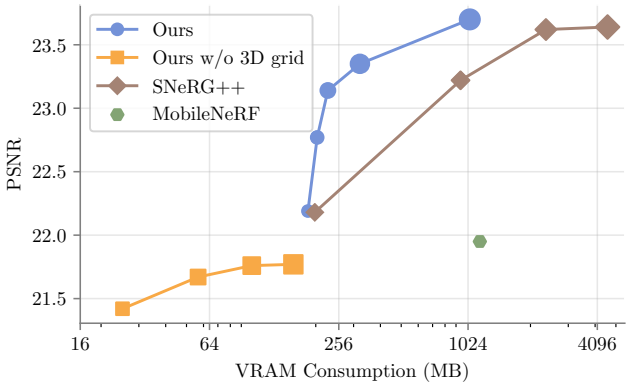


Fig. 4. PSNR (higher is better) vs VRAM consumption (lower is better) for our model, our improved SNeRG baseline, MobileNeRF, and an ablation of our model without 3D grids. Each line besides MobileNeRF represents the same model with a varying resolution of its underlying spatial grid, indicated by marker size.

Table 2. We compare our final model after baking to the model before baking (a) demonstrating that our Proposal-MLP-aware baking pipeline is almost lossless. Omitting quantization-aware training (b) leads to a drop in rendering quality. Our proposed contraction function performs on par with the original spherical contraction function (c), while enabling efficient ray-AABB intersection tests. Results are averaged over all outdoor scenes.

	PSNR ↑	SSIM ↑	LPIPS ↓
(a) Pre-baking	23.20	0.620	0.336
(b) w/o quant.-aware training	22.64	0.603	0.347
(c) Spherical contraction	23.22	0.619	0.341
Ours (Post-baking)	23.19	0.616	0.343

Table 3. Performance comparison on the "outdoor" scenes.

	PSNR ↑	SSIM ↑	LPIPS ↓	VRAM ↓	DISK ↓	FPS ↑
MacBook M1 Pro, 1280×720						
Mobile-NeRF	21.95	0.470	0.470	1162	345	65.7
BakedSDF	22.47	0.585	0.349	**463**	561	**121.5**
SNeRG++	**23.64**	**0.672**	**0.285**	4571	3785	18.7
Ours	23.19	0.616	0.343	524	**188**	28.3
NVIDIA RTX 3090, 1920×1080						
Instant-NGP	22.90	0.566	0.371	–	**107**	4
Ours	**23.19**	**0.616**	**0.343**	524	188	**119**

7 EXPERIMENTS

We experimentally evaluate our model in terms of rendering quality, video memory consumption, and real-time rendering performance. We compare MERF to a variety of offline view synthesis methods (NeRF [Mildenhall et al. 2020], mip-NeRF 360 [Barron et al. 2022], Stable View Synthesis [Riegler and Koltun 2021], and Instant-NGP [Müller et al. 2022]), and real-time ones (Deep Blending [Hedman et al. 2018], Mobile-NeRF [Chen et al. 2023], and SNeRG [Hedman et al. 2021]). To make this evaluation as rigorous as possible we evaluate against an improved version of SNeRG, which we call SNeRG++. SNeRG++ uses many components of our approach: multi-level empty space skipping, an optimized MLP implementation, our improved baking pipeline, and post-activation interpolation (which increases model expressivity by allowing for intra-voxel discontinuities [Karnewar et al. 2022; Sun et al. 2022]). Unless otherwise stated, for MERF we set the triplane resolution R to 2048 and the sparse grid resolution L to 512, and for SNeRG++ we set the grid resolution to 2048.

For evaluation, we use the challenging mip-NeRF 360 dataset [Barron et al. 2022] which contains five outdoor and four indoor scenes. All scenes are unbounded and require a high resolution representation (2048^3) to be faithfully reproduced. If not indicated otherwise, reported metrics are averaged over runs from the five outdoor scenes. We evaluate rendering quality using peak-signal-to-noise-ratio (PSNR), SSIM [Wang et al. 2004], and LPIPS [Zhang et al. 2018].

7.1 Quality Comparison

We first compare our method to a variety of offline and real-time view synthesis methods in terms of rendering quality on both outdoor and indoor scenes. As can be seen in Table 1, MERF achieves better scores than the real-time competitors DeepBlending and Mobile-NeRF on all three metrics. On the outdoor scenes, MERF even achieves higher scores than the offline methods Instant-NGP, NeRF and NeRF++ and performs on par with SVS in terms of PSNR and SSIM, despite offline rendering quality not being the primary focus of this work. In contrast, on indoor scenes MERF achieves slightly lower scores than Instant-NGP, SVS, and NeRF++. This might be a result of the indoor scenes containing more artificial materials with stronger view-dependent effects that our shallow decoder MLP h designed for real-time rendering cannot model well. Qualitatively, Figure 6 shows that our reconstructions of the background are much sharper compared to Mobile-NeRF and Instant-NGP, which we attribute to using a scene contraction function.

7.2 MERF vs. SNeRG++

We begin our analysis with a comparison of MERF with SNeRG++: both use the same training pipeline, view-dependent appearance model, and rendering engine implementation, as this enables controlled experiments. To support our claim that MERF provides a favorable trade-off between memory consumption and rendering quality, we train models with varying resolutions: For MERF, we vary the feature plane resolution R from 512 and 3072 while setting the grid resolution L to either 512 or 1024. Likewise, for SNeRG++, we vary the grid resolution L from 512 to 2048. To demonstrate the benefit of adding a low-resolution sparse 3D grid, we also train our models without the 3D grid. As can be seen in Figure 4, the memory consumption of SNeRG++ quickly rises to multiple gigabytes, whereas our model scales better to higher resolutions. For models without 3D grids we observe that quality saturates well below the other models. In Figure 5 we see that our model ($R = 2048$, $L = 512$) achieves similar quality to SNeRG++, while requiring a fraction of the memory. Additionally, in Figure 7 we see that ablating the 3D grid from MERF leads to a significant loss in quality.

7.3 Real-time Rendering Evaluation

Finally, we evaluate the rendering speed of MERF, MobileNeRF, SNeRG++, Instant-NGP, and our concurrent work BakedSDF in frames per second (FPS). Note that MERF, Mobile-NeRF and SNeRG++ all run in the browser and use the view-dependence model introduced by Hedman et al. [Hedman et al. 2021]. In contrast, Instant-NGP uses a different view-dependence model and is implemented in CUDA and is therefore less portable across devices. For benchmarking the methods that include web viewers (MERF, Mobile-NERF, SNeRG++, BakedSDF) we use an M1 MacBook Pro and set the rendering resolution to 1280×720. When evaluating against Instant-NGP, to make the comparison fair we use an RTX 3090 GPU (which Instant-NGP is optimized for) and increase the rendering resolution to 1920×1080 to demonstrate MERF's scalability on high-powered devices. As can be seen in Table 3, our method runs faster than SNeRG++ while consuming only one fifth of the memory. While MobileNeRF achieves higher frame rates on the Macbook than MERF, it requires twice as much video memory and reduces rendering quality (a 1.24 dB reduction in PSNR). This reduced quality is especially evident in background regions, as shown in Figure 6. From our experiment with the RTX 3090, we see that Instant-NGP does not achieve real-time performance (4 FPS), while MERF renders at frame rates well above 100. While the concurrent work BakedSDF demonstrates that the quality gap between volume-based and mesh-based representations can be further reduced, MERF still outperforms BakedSDF in terms of fidelity.

7.4 Limitations

Since we use the view-dependence model introduced in SNeRG [Hedman et al. 2021], we also inherit its limitations: By evaluating view-dependent color once per ray, we are unable to faithfully model view-dependent appearance for rays that intersect with semi-transparent objects. Furthermore, since the tiny MLP has limited capacity, it may struggle to scale to much larger scenes or objects with complex reflections.

Moreover, our method still performs volume rendering, which limits it to devices equipped with a sufficiently powerful GPU such as laptops, tablets or workstations. Running our model on smaller, thermally limited devices such as mobile phones or headsets will require further reductions in memory and runtime.

8 CONCLUSION

We have presented MERF, a compressed volume representation for radiance fields, which enables real-time rendering of large-scale scenes in a browser. By using novel hybrid volumetric parameterization, a novel contraction function that preserves straight lines, and a baking procedure that ensures that our real-time representation describes the same radiance field as was used during optimization, MERF is able to achieve faster and more accurate real-time rendering of large and complicated real-world scenes than prior real-time NeRF-like models. Out of all real-time methods, ours produces the highest-quality renderings for any given memory budget. Not only does it achieve 31.6% (MSE) higher quality in the outdoor scenes compared to MobileNeRF, the previous state-of-the-art, it also requires less than half of the GPU memory.

ACKNOWLEDGMENTS

We thank Marcos Seefelder, Julien Philip and Simon Rodriguez for their suggestions on shader optimization. This work was supported by the ERC Starting Grant LEGO3D (850533) and the DFG EXC number 2064/1 - project number 390727645.

REFERENCES

D.G. Aliaga, T. Funkhouser, D. Yanovsky, and I. Carlbom. 2002. Sea of images. *IEEE Visualization* (2002).

Benjamin Attal, Jia-Bin Huang, Michael Zollhöfer, Johannes Kopf, and Changil Kim. 2022. Learning Neural Light Fields with Ray-Space Embedding Networks. *CVPR* (2022).

Benjamin Attal, Selena Ling, Aaron Gokaslan, Christian Richardt, and James Tompkin. 2020. MatryODShka: Real-time 6DoF Video View Synthesis using Multi-Sphere Images. *ECCV* (2020).

Jonathan T. Barron, Ben Mildenhall, Dor Verbin, Pratul P. Srinivasan, and Peter Hedman. 2022. Mip-NeRF 360: Unbounded Anti-Aliased Neural Radiance Fields. *CVPR* (2022).

Yoshua Bengio, Nicholas Léonard, and Aaron Courville. 2013. Estimating or propagating gradients through stochastic neurons for conditional computation. *arXiv:1308.3432* (2013).

Michael Broxton, John Flynn, Ryan Overbeck, Daniel Erickson, Peter Hedman, Matthew DuVall, Jason Dourgarian, Jay Busch, Matt Whalen, and Paul Debevec. 2020. Immersive Light Field Video with a Layered Mesh Representation. *ACM Transactions on Graphics* (2020).

Chris Buehler, Michael Bosse, Leonard McMillan, Steven Gortler, and Michael Cohen. 2001. Unstructured Lumigraph Rendering. *SIGGRAPH* (2001).

Junli Cao, Huan Wang, Pavlo Chemerys, Vladislav Shakhrai, Ju Hu, Yun Fu, Denys Makoviichuk, Sergey Tulyakov, and Jian Ren. 2023. Real-Time Neural Light Field on Mobile Devices. *CVPR* (2023).

Eric R. Chan, Connor Z. Lin, Matthew A. Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas Guibas, Jonathan Tremblay, Sameh Khamis, Tero Karras, and Gordon Wetzstein. 2022. Efficient Geometry-aware 3D Generative Adversarial Networks. *CVPR* (2022).

Gaurav Chaurasia, Sylvain Duchene, Olga Sorkine-Hornung, and George Drettakis. 2013. Depth Synthesis and Local Warps for Plausible Image-Based Navigation. *ACM Transactions on Graphics* (2013).

Anpei Chen, Zexiang Xu, Andreas Geiger, Jingyi Yu, and Hao Su. 2022. TensoRF: Tensorial Radiance Fields. *ECCV* (2022).

Zhiqin Chen, Thomas Funkhouser, Peter Hedman, and Andrea Tagliasacchi. 2023. MobileNeRF: Exploiting the polygon rasterization pipeline for efficient neural field rendering on mobile architectures. *CVPR* (2023).

Paul Debevec, Yizhou Yu, and George Borshukov. 1998. Efficient view-dependent image-based rendering with projective texture-mapping. *EGSR* (1998).

Chenxi Lola Deng and Enzo Tartaglione. 2023. Compressing Explicit Voxel Grid Representations: Fast NeRFs Become Also Small. *WACV* (2023).

Terrance DeVries, Miguel Angel Bautista, Nitish Srivastava, Graham W. Taylor, and Joshua M. Susskind. 2021. Unconstrained Scene Generation with Locally Conditioned Radiance Fields. *ICCV* (2021).

John Flynn, Michael Broxton, Paul Debevec, Matthew DuVall, Graham Fyffe, Ryan Overbeck, Noah Snavely, and Richard Tucker. 2019. DeepView: View synthesis with learned gradient descent. *CVPR* (2019).

Stephan J. Garbin, Marek Kowalski, Matthew Johnson, Jamie Shotton, and Julien Valentin. 2021. FastNeRF: High-Fidelity Neural Rendering at 200FPS. *ICCV* (2021).

Steven J. Gortler, Radek Grzeszczuk, Richard Szeliski, and Michael F. Cohen. 1996. The lumigraph. *SIGGRAPH* (1996).

Peter Hedman, Julien Philip, True Price, Jan-Michael Frahm, George Drettakis, and Gabriel Brostow. 2018. Deep blending for free-viewpoint image-based rendering. *SIGGRAPH Asia* (2018).

Peter Hedman, Pratul P. Srinivasan, Ben Mildenhall, Jonathan T. Barron, and Paul Debevec. 2021. Baking Neural Radiance Fields for Real-Time View Synthesis. *ICCV* (2021).

Animesh Karnewar, Tobias Ritschel, Oliver Wang, and Niloy Mitra. 2022. ReLU fields: The little non-linearity that could. *SIGGRAPH* (2022).

Petr Kellnhofer, Lars Jebe, Andrew Jones, Ryan Spicer, Kari Pulli, and Gordon Wetzstein. 2021. Neural Lumigraph Rendering. *CVPR* (2021).

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *ICLR*.

Andreas Kurz, Thomas Neff, Zhaoyang Lv, Michael Zollhöfer, and Markus Steinberger. 2022. AdaNeRF: Adaptive Sampling for Real-time Rendering of Neural Radiance Fields. *ECCV* (2022).

Marc Levoy and Pat Hanrahan. 1996. Light field rendering. *SIGGRAPH* (1996).

Lingzhi Li, Zhen Shen, Zhongshu Wang, Li Shen, and Liefeng Bo. 2022b. Compressing Volumetric Radiance Fields to 1 MB. *arXiv:2211.16386* (2022).

Ruilong Li, Matthew Tancik, and Angjoo Kanazawa. 2022d. NerfAcc: A General NeRF Accleration Toolbox. *arXiv:2210.04847* (2022).

Sicheng Li, Hao Li, Yue Wang, Yiyi Liao, and Lu Yu. 2022a. SteerNeRF: Accelerating NeRF Rendering via Smooth Viewpoint Trajectory. *arXiv:2212.08476* (2022).

Zhong Li, Liangchen Song, Celong Liu, Junsong Yuan, and Yi Xu. 2022c. NeuLF: Efficient Novel View Synthesis with Neural 4D Light Field. *EGSR* (2022).

Zhi-Hao Lin, Wei-Chiu Ma, Hao-Yu Hsu, Yu-Chiang Frank Wang, and Shenlong Wang. 2022. NeurMiPs: Neural Mixture of Planar Experts for View Synthesis. *CVPR* (2022).

David B. Lindell, Julien N.P. Martel, and Gordon Wetzstein. 2021. AutoInt: Automatic Integration for Fast Neural Rendering. *CVPR* (2021).

Lingjie Liu, Jiatao Gu, Kyaw Zaw Lin, Tat-Seng Chua, and Christian Theobalt. 2020. Neural Sparse Voxel Fields. *NeurIPS* (2020).

Ricardo Martin-Brualla, Noha Radwan, Mehdi S. M. Sajjadi, Jonathan T. Barron, Alexey Dosovitskiy, and Daniel Duckworth. 2021. NeRF in the Wild: Neural Radiance Fields for Unconstrained Photo Collections. *CVPR* (2021).

Nelson Max. 1995. Optical models for direct volume rendering. *IEEE Transactions on Visualization and Computer Graphics* (1995).

Ben Mildenhall, Pratul P. Srinivasan, Rodrigo Ortiz-Cayon, Nima Khademi Kalantari, Ravi Ramamoorthi, Ren Ng, and Abhishek Kar. 2019. Local Light Field Fusion: Practical View Synthesis with Prescriptive Sampling Guidelines. *ACM Transactions on Graphics* (2019).

Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. 2020. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. *ECCV* (2020).

Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. 2022. Instant neural graphics primitives with a multiresolution hash encoding. *SIGGRAPH* (2022).

Jacob Munkberg, Wenzheng Chen, Jon Hasselgren, Alex Evans, Tianchang Shen, Thomas Müller, Jun Gao, and Sanja Fidler. 2022a. Extracting Triangular 3D Models, Materials, and Lighting From Images. *CVPR* (2022).

Jacob Munkberg, Jon Hasselgren, Tianchang Shen, Jun Gao, Wenzheng Chen, Alex Evans, Thomas Müller, and Sanja Fidler. 2022b. Extracting Triangular 3D Models, Materials, and Lighting From Images. *CVPR* (2022).

Thomas Neff, Pascal Stadlbauer, Mathias Parger, Andreas Kurz, Joerg H. Mueller, Chakravarty R. Alla Chaitanya, Anton Kaplanyan, and Markus Steinberger. 2021. DONeRF: Towards Real-Time Rendering of Compact Neural Radiance Fields using Depth Oracle Networks. *Computer Graphics Forum* (2021).

Songyou Peng, Michael Niemeyer, Lars Mescheder, Marc Pollefeys, and Andreas Geiger. 2020. Convolutional Occupancy Networks. *ECCV* (2020).

Martin Piala and Ronald Clark. 2021. TermiNeRF: Ray Termination Prediction for Efficient Neural Rendering. *3DV* (2021).

Christian Reiser, Songyou Peng, Yiyi Liao, and Andreas Geiger. 2021. KiloNeRF: Speeding up neural radiance fields with thousands of tiny MLPs. *ICCV* (2021).

Gernot Riegler and Vladlen Koltun. 2021. Stable view synthesis. *CVPR* (2021).

Darius Rückert, Linus Franke, and Marc Stamminger. 2022. ADOP: Approximate differentiable one-pixel point rendering. *SIGGRAPH* (2022).

Vincent Sitzmann, Semon Rezchikov, William T. Freeman, Joshua B. Tenenbaum, and Fredo Durand. 2021. Light Field Networks: Neural Scene Representations with Single-Evaluation Rendering. *NeurIPS* (2021).

Table 4. Varying the resolution of the 3D grid in our representation.

Resolution L	PSNR ↑	SSIM ↑	LPIPS ↓	VRAM ↓
–	20.59	0.509	0.430	101
64	21.81	0.562	0.388	102
128	22.48	0.585	0.368	107
256	22.81	0.600	0.355	134
512	23.19	0.616	0.343	524

Cheng Sun, Min Sun, and Hwann-Tzong Chen. 2022. Direct voxel grid optimization: Super-fast convergence for radiance fields reconstruction. *CVPR* (2022).

Towaki Takikawa, Alex Evans, Jonathan Tremblay, Thomas Müller, Morgan McGuire, Alec Jacobson, and Sanja Fidler. 2022. Variable Bitrate Neural Fields. *ACM Transactions on Graphics* (2022).

Matthew Tancik, Vincent Casser, Xinchen Yan, Sabeek Pradhan, Ben Mildenhall, Pratul Srinivasan, Jonathan T. Barron, and Henrik Kretzschmar. 2022. Block-NeRF: Scalable Large Scene Neural View Synthesis. *CVPR* (2022).

Ayush Tewari, Justus Thies, Ben Mildenhall, Pratul Srinivasan, Edgar Tretschk, W Yifan, Christoph Lassner, Vincent Sitzmann, Ricardo Martin-Brualla, Stephen Lombardi, et al. 2022. Advances in neural rendering. *Computer Graphics Forum* (2022).

Haithem Turki, Deva Ramanan, and Mahadev Satyanarayanan. 2022. Mega-NERF: Scalable Construction of Large-Scale NeRFs for Virtual Fly-Throughs. *CVPR* (2022).

Huan Wang, Jian Ren, Zeng Huang, Kyle Olszewski, Menglei Chai, Yun Fu, and Sergey Tulyakov. 2022b. R2L: Distilling Neural Radiance Field to Neural Light Field for Efficient Novel View Synthesis. *ECCV* (2022).

Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. 2021. NeuS: Learning Neural Implicit Surfaces by Volume Rendering for Multi-view Reconstruction. *NeurIPS* (2021).

Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE TIP* (2004).

Zhongshu Wang, Lingzhi Li, Zhen Shen, Li Shen, and Liefeng Bo. 2022a. 4K-NeRF: High Fidelity Neural Radiance Fields at Ultra High Resolutions. *arXiv:2212.04701* (2022).

Suttisak Wizadwongsa, Pakkapon Phongthawee, Jiraphon Yenphraphai, and Supasorn Suwajanakorn. 2021. NeX: Real-time View Synthesis with Neural Basis Expansion. *CVPR* (2021).

Liwen Wu, Jae Yong Lee, Anand Bhattad, Yuxiong Wang, and David Forsyth. 2022a. DIVeR: Real-time and Accurate Neural Radiance Fields with Deterministic Integration for Volume Rendering. *CVPR* (2022).

Xiuchao Wu, Jiamin Xu, Zihan Zhu, Hujun Bao, Qixing Huang, James Tompkin, and Weiwei Xu. 2022b. Scalable Neural Indoor Scene Rendering. *ACM TOG* (2022).

Lior Yariv, Jiatao Gu, Yoni Kasten, and Yaron Lipman. 2021. Volume rendering of neural implicit surfaces. *NeurIPS* (2021).

Lior Yariv, Peter Hedman, Christian Reiser, Dor Verbin, Pratul P. Srinivasan, Richard Szeliski, Jonathan T. Barron, and Ben Mildenhall. 2023. BakedSDF: Meshing Neural SDFs for Real-Time View Synthesis. *SIGGRAPH* (2023).

Penghang Yin, Jiancheng Lyu, Shuai Zhang, Stanley Osher, Yingyong Qi, and Jack Xin. 2019. Understanding straight-through estimator in training activation quantized neural nets. *ICLR* (2019).

Alex Yu, Sara Fridovich-Keil, Matthew Tancik, Qinhong Chen, Benjamin Recht, and Angjoo Kanazawa. 2022. Plenoxels: Radiance fields without neural networks. *CVPR* (2022).

Alex Yu, Ruilong Li, Matthew Tancik, Hao Li, Ren Ng, and Angjoo Kanazawa. 2021. PlenOctrees for real-time rendering of neural radiance fields. *ICCV* (2021).

Jian Zhang, Jinchi Huang, Bowen Cai, Huan Fu, Mingming Gong, Chaohui Wang, Jiaming Wang, Hongchen Luo, Rongfei Jia, Binqiang Zhao, and Xing Tang. 2022. Digging into Radiance Grid for Real-Time View Synthesis with Detail Preservation. *ECCV* (2022).

Kai Zhang, Gernot Riegler, Noah Snavely, and Vladlen Koltun. 2020. NeRF++: Analyzing and Improving Neural Radiance Fields. *arXiv:2010.07492* (2020).

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. 2018. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. *CVPR* (2018).

Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. 2018. Stereo Magnification: Learning View Synthesis using Multiplane Images. *SIGGRAPH* (2018).

A TRAINING DETAILS

All SNeRG++ and MERF models use the same training hyperparameters and architectures. We train for 25000 iterations with a batch size of 2^{16} pixels, which takes approximately 3.5 hours. Visibility calculation takes approximately an hour but could be reduced to a few minutes with a better implementation. The final baking step requires less than two minutes. A training batch is created by sampling pixels from all training images. We use the Adam [Kingma and Ba 2015] optimizer with an exponentially decaying learning rate. The learning rate is warmed up during the first 100 iterations where it is increased from 1e−4 to 1e−2. Then the learning rate is decayed to 1e−3. Adam's hyperparameters β_1, β_2 and ϵ are set to 0.9, 0.99 and 1e−15, respectively. To regularize the hash grids we use a weight decay of 0.03 on its grid values.

Table 5. VRAM consumption in MB of individual components of SNeRG++ and the proposed representation. The multi-resolution occupancy grid used for empty space skipping is referred to as "occupancy". Results are averaged over all scenes from Mip-NeRF 360 [Barron et al. 2022].

	3D grid	2D planes	occupancy	total
SNeRG++	4751	–	2	4753
Ours	299	101	2	402

B ARCHITECTURE

For parameterizing all grids (i.e. 3D voxel grids and 2D planes) we use an MLP with a multi-resolution hash encoding [Müller et al. 2022]. The hash encoding uses 20 levels and the individual hash tables have 2^{21} entries. Following Müller et al. [2022], hash collisions are resolved with a 2-layer MLP with 64 hidden units. This MLP outputs an 8-dimensional vector representing density, diffuse RGB and the 4-dimensional view-dependency feature vector. For our deferred view-dependency model we closely follow SNeRG [Hedman et al. 2021] and use a 3-layer MLP with 16 hidden units. As in SNeRG viewing directions are encoded with 4 frequencies. Following MipNeRF360 [Barron et al. 2022] we use hierarchical sampling with three levels and therefore require two Proposal-MLPs. The Proposal-MLPs consist of 2 layers with 64 hidden units and use a hash encoding. Since a Proposoal-MLP merely needs to model coarse geometry, for the Proposal-MLPs' hash encodings we use only 10 levels, a maximum grid resolution of 512 and a hash table size of 2^{16}.

C BENCHMARKING

For each test scene we define a camera pose that is used for benchmarking. Camera poses are set programmatically in each viewer. For fair comparison, we ensure that resolutions and camera intrinsics (i.e. field of view) are identical across viewers. We compute the average frame rate across 150 frames.

By default, Instant-NGP makes use of progressive upsampling, where the image is first rendered at a lower resolution. When the camera rests, additional low resolutions images are rendered that are dynamically combined into the final high resolution image. We also implemented progressive rendering as part of our webviewer. Independent of the method, progressive upsampling speeds up rendering in proportion to the ratio of target resolution and initial render resolution. For simplicity, we disable progressive rendering for benchmarking.

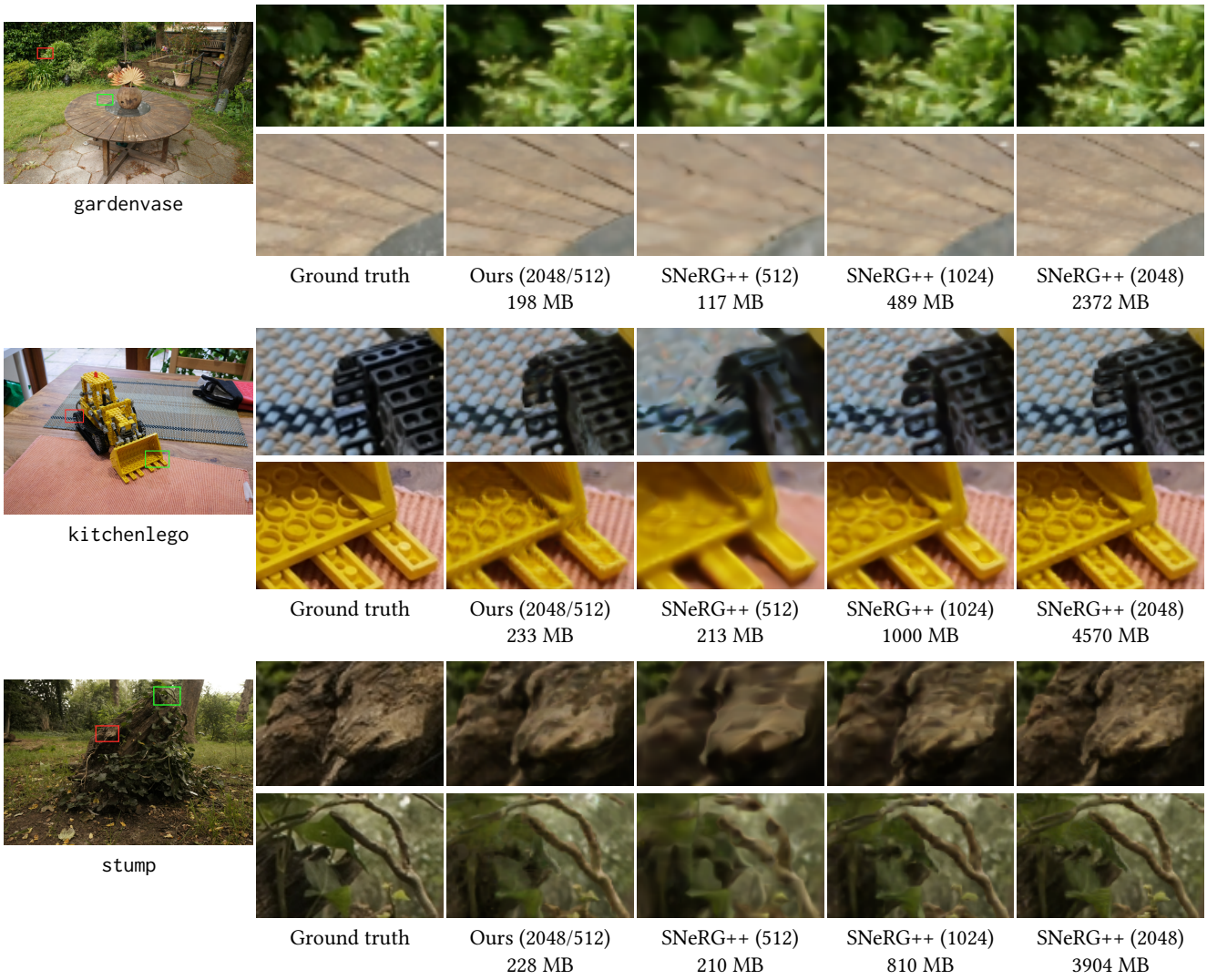


Fig. 5. Visual comparison between MERF and various resolution SNeRG++ [Hedman et al. 2021] models. Total VRAM (GPU memory) usage during rendering is listed beneath each method name. Only SNeRG++ (512) has comparable size to our model, whereas SNeRG++ (1024) and SNeRG++ (2048) are significantly larger.

		Outdoor						Indoor				
		bicycle	flower	garden	stump	tree	mean	living	kitchen	lego	bonsai	mean
PSNR ↑	SNeRG++	**23.80**	**21.12**	**26.31**	24.54	**22.44**	**23.64**	**29.79**	**26.35**	**28.72**	**30.04**	**28.72**
	Ours	22.62	20.33	25.58	**25.04**	22.39	23.19	29.28	25.82	27.42	28.68	27.80
SSIM ↑	SNeRG++	**0.684**	**0.564**	**0.814**	**0.715**	**0.582**	**0.672**	**0.892**	**0.849**	**0.872**	**0.914**	**0.882**
	Ours	0.595	0.492	0.763	0.677	0.554	0.616	0.874	0.819	0.842	0.884	0.855
LPIPS ↓	SNeRG++	**0.292**	**0.349**	**0.165**	**0.267**	**0.353**	**0.285**	**0.254**	**0.255**	**0.184**	**0.219**	**0.228**
	Ours	0.371	0.406	0.215	0.309	0.414	0.343	0.292	0.307	0.224	0.262	0.271
VRAM ↓	SNeRG++	5041	4727	2372	3904	6810	4571	5536	4091	4570	5707	4976
	Ours	**291**	**271**	**198**	**228**	**1630**	**524**	**197**	**254**	**233**	**295**	**245**
DISK ↓	SNeRG++	4173	3876	1888	3226	5766	3785	3781	2834	3175	3209	3250
	Ours	**207**	**189**	**114**	**156**	**273**	**188**	**92**	**136**	**118**	**167**	**129**

Table 6. Per-scene comparison between SNeRG++ and our method on the dataset from Mip-NeRF 360 [Barron et al. 2022].

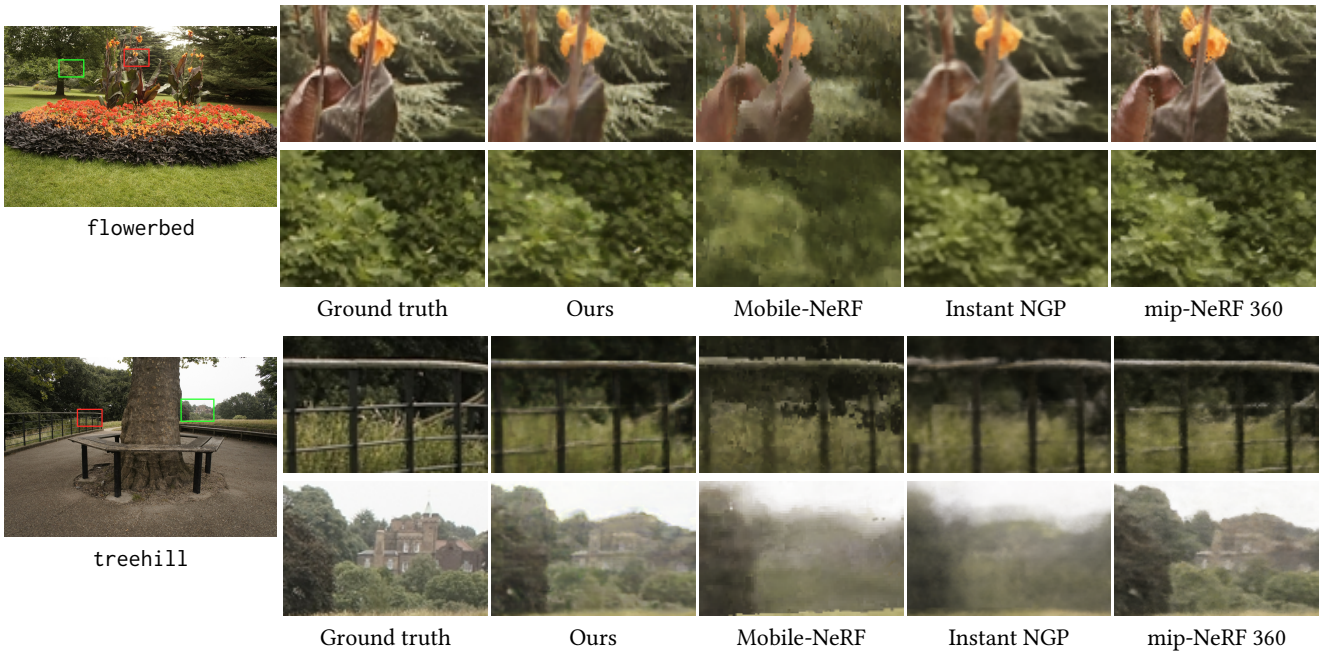


Fig. 6. Visual comparison between MERF and other view synthesis methods. Mobile-NeRF [Chen et al. 2023] is the only other real-time method (30fps or better). Instant NGP [Müller et al. 2022] runs at interactive rates (around 5fps) and mip-NeRF 360 [Barron et al. 2022] is an extremely heavyweight offline method (around 30 seconds to render a single frame), representing the current state-of-the-art view synthesis quality.

Fig. 7. Visual comparison between MERF with and without a low-resolution 3D voxel grid (both models use three high-resolution 2D grids). Omitting the 3D grid often results in parts of the scene being poorly reconstructed; note the missing background in this example.

(a) Render only 3D grid (b) Render only 2D planes (c) Normal rendering

Fig. 8. To visualize what information is stored in the respective part of our representations we render only the 3D grid (a) respectively only the 2D planes (b).