
CeRF: Convolutional Neural Radiance Fields for New View Synthesis with Derivatives of Ray Modeling

Xiaoyan Yang* Dingbo Lu* Yang Li† Chenhui Li Changbo Wang†

School of Computer Science and Technology, East China Normal University
 {51215901035, 51215901103}@stu.ecnu.edu.cn
 {yli, chli, cbwang}@cs.ecnu.edu.cn

Abstract

In recent years, novel view synthesis has gained popularity in generating high-fidelity images. While demonstrating superior performance in the task of synthesizing novel views, the majority of these methods are still based on the conventional multi-layer perceptron for scene embedding. Furthermore, light field models suffer from geometric blurring during pixel rendering, while radiance field-based volume rendering methods have multiple solutions for a certain target of density distribution integration. To address these issues, we introduce the Convolutional Neural Radiance Fields to model the derivatives of radiance along rays. Based on 1D convolutional operations, our proposed method effectively extracts potential ray representations through a structured neural network architecture. Besides, with the proposed ray modeling, a proposed recurrent module is employed to solve geometric ambiguity in the fully neural rendering process. Extensive experiments demonstrate the promising results of our proposed model compared with existing state-of-the-art methods.

1 Introduction

Recently, novel view synthesis with neural implicit representations has rapidly advanced due to its surprisingly high-quality generated image with different camera poses. Based on the original neural radiance field (NeRF) method [19], various research directions have been explored to further improve rendering quality, rendering speed, and other aspects [4, 6, 36, 45]. In addition, as NeRF bridges the radiance and geometrics information, many other applications exploit NeRF as their building block to achieve other graphic goals, such as extracting geometrical, semantic, and material information from the scene[22, 40, 48], and extending static setting into dynamic scenes [1].

As most of these methods are based on the NeRF framework, the volume rendering strategy remains in the processing pipeline yet has multi-solution ambiguity. As shown in the left of Figure 1, the same integral color is obtained under completely different density distributions. Similarly, for light field-based methods [24], the geometric configuration is ambiguous since both the point on the surface \mathbf{P}_T and the point in space \mathbf{P}_F have the same colors viewing from two pose, as shown in the right of Figure 1. Specifically, the different points \mathbf{P}_T and \mathbf{P}_F receive same radiance $L(\mathbf{P}_T, \mathbf{o}_i) = L(\mathbf{P}_F, \mathbf{o}_i)$ for $i = \{1, 2\}$ from two camera observation \mathbf{o}_1 and \mathbf{o}_2 . This indicates that there are potentially multiple position solutions to satisfy the same view-dependent radiance, which is hard to optimize for neural networks. NeuS [35] introduces neural implicit surfaces with surface constraints. However, surface-based approaches have complicated conversion from SDF to volume density.

*Equal contribution.

†Yang Li and Changbo Wang are the corresponding authors.

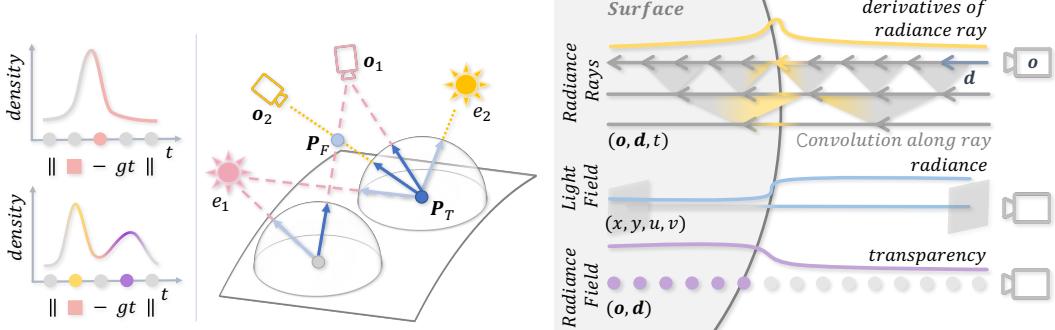


Figure 1: Ambiguity in light field and radiance field

Figure 2: Visualization for different fields

Although SDF-based methods have complicated data structures, the network structure is very simple as they mainly employ Multi-Layer Perceptron (MLP) for implicit representation. In the field of deep learning, MLP is initially the simplest design, but many architectures have been developed since then, including Convolutional Neural networks (CNN) [10] and Transformers [30]. Furthermore, CNNs have been shown to be successful in image classification tasks, such as ResNet [11]. Additionally, recurrent neural networks (RNNs) have proved to be efficient in handling sequential data.

In this paper, we model the novel view synthesis task as an implicit neural representation of the derivatives of radiance in the scene. As depicted in Figure 2, the radiance remains constant until the rays hit the surface, and the radiance remains zero internally for homogeneous materials. Compared to modeling in NeRF and light field methods, our proposed method only requires the network to store information on the surface, rather than the entirety of the empty space from origin to intersection. Therefore, modeling the radiance rays as derivatives yields sparse solutions along individual rays and the optimization problem changes from a regression task to a classification-like task.

To further leverage the advantages of the modeling, we propose the Convolutional Neural Radiance Fields (CeRF) to extract local features along the ray with a fully neural rendering scheme. Our key ideas are to incorporate a sophisticated CNN structure to achieve a more streamlined and uninterrupted ray representation and formulate a gated recurrent unit (GRU) in the rendering network. By approximating the render equation with a fully neural network process, the theoretical upper limit of our proposed approach exceeds the expression capability of the volume-based rendering scheme in conventional NeRF. Extensive experiments demonstrate that, by modeling radiance derivatives, our proposed CeRF exhibits a more elegant and effective encoding of scenes, and achieves comparable results with state-of-the-art methods. Our proposed structure can be easily incorporated into all NeRF-based methods designed for specific scenes, leading to improved rendering results. Our contributions can be summarized as follows:

- (1) A theoretical analysis of the rendering equation with novel radiance model. To the best of our knowledge, we are the first to model the derivatives of radiance along a ray in novel view synthesis.
- (2) A novel CeRF framework to encode ray features and employs a neural ray rendering based on our proposed rendering scheme.
- (3) Experiments demonstrate that CeRF attains comparable results to the state-of-the-art models and accurately fit complex geometric details. We will release our source code for further research.

2 Related work

2.1 Neural light fields

Neural light fields have been well-received in synthesizing new views due to their good results on scenes with complex optical phenomena and fast rendering speed [25]. NeX [39] utilizes a linear combination of multiple neural network-based basis functions to obtain pixel color values. RSE [2] maps the input rays to a high-dimensional space, and the rendering of each pixel requires only one visit to the network, resulting in faster view synthesis. For dense input images, light field-based rendering methods [9, 15] can be used to synthesize new views. However, for sparse input images,

light field-based interpolation alone fails to generate images with scene coherence due to the large spacing between viewpoints. Wang *et al.* [37] propose a progressively connected neural network for light field view synthesis, improving both efficiency and quality. LFN [24] proposes a method for learning a neural representation of the scene that can be rendered by a single forward pass to a new view. LFNR [25] further improves upon this by introducing constraints on the geometry of multiple viewpoints in the light field to learn the representation of a scene in a sparse set of inputs. R2L [34] distills neural radiance fields into neural light fields for efficient view synthesis, converting the former into the latter for more efficient view synthesis. Compared with light field-based approaches, our proposed CeRF is capable of dealing with geometric information and has a very different light ray modeling scheme.

2.2 Neural radiance fields

Mildenhall *et al.* [19] proposed NeRF, which represents the scene as a continuous function of 3D spatial coordinates, allowing for high-quality rendering of new views. Subsequently, many improvements to NeRF have been proposed. Some methods aim to improve rendering quality [4, 31, 12, 41, 21, 29, 43]. Others focus on optimizing inference speed or model size [20, 16, 5, 6, 28, 17]. Among these, Mip-NeRF[4] extends the rays-based representation to frustums, significantly improving the rendering effectiveness of NeRF. Plenoxels [44] uses a sparse voxel grid to explicitly represent a scene, speeding up convergence and rendering. DVGO [26] proposes the post-activation interpolation on voxel density and incorporates a priori to solve suboptimal geometry solutions, achieving fast convergence of the network. In addition, some methods not only generate rendered images from implicit representations but also model more about geometry [42], semantic [33], or material information [31, 48, 3]. NeuRay [18] predicts the visibility of 3D points in input views, allowing the network to focus on visible image features when constructing the radiance field. Methods that simultaneously estimate SDF can alleviate the problem of multiple solutions, like PhySG [46], an SDF-based hybrid representation of specular BRDFs and environmental illumination, to enable physics-based editing of material and lighting radiance. Ref-NeRF[32] parameterizes the scene as a view-dependent outgoing radiance, improving the realism of specular reflections. However, all these methods employ MLP as the main neural network. In contrast, our proposed CeRF focuses on ray derivatives modeling and its corresponding structure of neural networks. Mukund *et al.* [27] proposed to use a transformer instead of a volume rendering function, which is closest to our method. However, our architecture utilizes a simpler convolutional network, while achieving better performance.

3 Approach

Given a set of images with known camera poses and intrinsic parameters, our target is rendering new views for unseen poses. To this end, we introduce CeRF, as shown in Figure 3. Specifically, for each pixel in an image, CeRF emits a ray \mathbf{r} from a viewing origin \mathbf{o} with a normalized view direction \mathbf{d} and a set of distances $t_i, i \in [1, D]$ along rays, and predicts the pixel color $C(\mathbf{r})$.

Viewing points along a ray as a whole, the position embedding and direction embedding are inputted into Convolutional Ray Feature Extractor to extract ray feature encoding $\mathbf{V}_\mathbf{r}$. A GRU-based Geometry Attribute Network with Unique Surface Constraint α to estimate geometry attribute coefficients $\mathbf{w}_\mathbf{r} = [w_1, \dots, w_D]$. Meanwhile, Radiance Attribute Network predicts radiance color $\mathbf{C}_\mathbf{r} = [c_1, \dots, c_D]$. The final pixel color $C(r)$ is obtained by calculating the expectation and introducing epipolar points. The coarse-to-fine structure is used as the same as other NeRF-based methods [4].

In the following sections, we first derivate the radiance ray from the render equation, followed by an overview of the Convolutional Ray Feature Extractor. Then, a detailed description of our proposed neural radiance ray rendering. Finally, the loss function for parameters optimization is introduced.

3.1 Derivatives of radiance ray

We begin with the render equations in computer graphics to build our model. To render a synthesized image, we need to calculate the radiance $L(\mathbf{p}, \omega_o)$ leaving point \mathbf{p} in direction ω_o along the view ray. For new view synthesis settings, the pose, and intensity of the light sources in the scene do not change with time or wavelength. Therefore, the formula of the render equation at one surface point \mathbf{p}

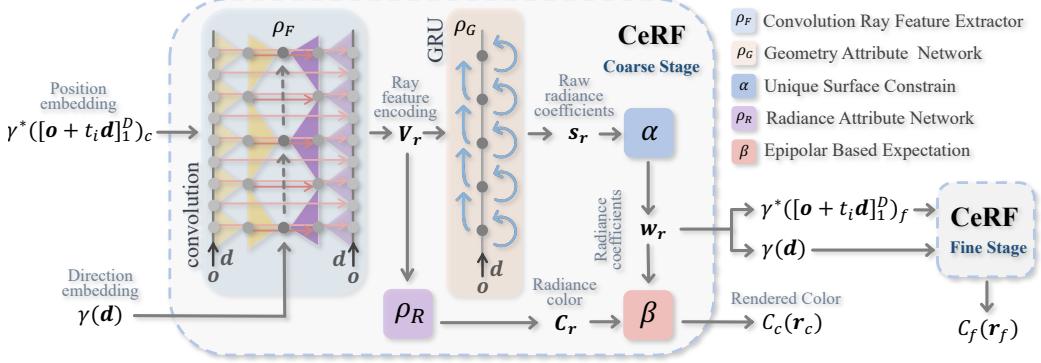


Figure 3: **Overview.** Based on the modeling of the derivatives of radiance ray, we proposed a fully neural rendering framework. Two key ideas guide our design: 1) encoding the light using the convolutional operation on sampled points along a ray, and 2) using a neural network to fit the rendering process, which can render new views.

can be simplified to scattering equation [8] as follows:

$$L(\mathbf{p}, \omega_o) = L^{\text{out}}(\mathbf{p}, \omega_o) + \int_{\omega_i \in \mathbf{S}^2(\mathbf{p})} L^{\text{in}}(\mathbf{p}, -\omega_i) f_s(\mathbf{p}, \omega_i, \omega_o) |\omega_i \cdot \mathbf{n}_p| d\omega_i, \quad (1)$$

where $L^{\text{out}}(\mathbf{p}, \omega_o)$ symbolized the light emitted by point \mathbf{p} in direction ω_o , and $L^{\text{in}}(\mathbf{p}, -\omega_i)$ represents the light coming to the surface. $f_s(\mathbf{p}, \omega_i, \omega_o)$ is the bidirectional scattering distribution functions (BSDFs) that characterizes the scattering behavior of light, for given incoming and outgoing directions ω_i and ω_o , respectively. The integrable region \mathbf{S}^2 is the direction of all incident light coming to a point \mathbf{p} .

Given a ray emitted from the camera origin \mathbf{o} in the direction \mathbf{d} , its intersection with a surface can be represented as $\mathbf{p} = \mathbf{o} + t\mathbf{d}$, where t is the distance from the intersection to the origin of the view ray. The direction of the outgoing radiance ω_o is opposite to the view direction \mathbf{d} , i.e., $\mathbf{d} = -\omega_o$. As we only consider the radiance flux along a single ray, the radiance can be rewritten as $L(\mathbf{o}, \mathbf{d}, t)$, which only depends on three variables $\mathbf{o}, \mathbf{d}, t$. Based on the universal function fitting capability of neural networks, we believe that a well-designed network \mathbf{F} can approximate the render function for different materials. Hence, we propose an approximation factorization of the scattering equation as

$$L(\mathbf{o}, \mathbf{d}, t) = \mathbf{F}(\mathbf{o}, \mathbf{d}, t; \Theta), \quad (2)$$

where Θ is network parameters encoding the constant properties for diverse scenes. In general, we can consider that the radiance of the entire static scene, including both global and local illumination, has been precomputed and baked into a texture map on the object surface [23]. Consequently, we can use a powerful function to fit the rendering effects.

Light field and radiance field can model the rendering equation, but both of them are multiple possibilities from different perspectives. In addition, the neural network needs to learn the surface radiance even at empty space, leading to increased complication in network fitting. In order to obtain a single mapping space, we model the radiance ray as the derivatives of radiance $\partial L(\mathbf{o}, \mathbf{d}, t)/\partial t$ instead of the radiance $L(\mathbf{o}, \mathbf{d}, t)$ itself. Thus, the network directly predicts the position and color of the surface intersection points. The radiance along a ray r from origin \mathbf{o} in direction \mathbf{d} can be expressed as the integral of the derivatives of the radiance ray as

$$L(\mathbf{o}, \mathbf{d}, t) = \int_t^{+\infty} \frac{\partial L(\mathbf{o}, \mathbf{d}, x)}{\partial x} dx. \quad (3)$$

As the radiance only changes significantly near the surface and remains zeros in empty space or inside objects, we introduce Dirac's delta function $\delta(\mathbf{o}, \mathbf{d}, t)$ to approximate the derivatives of radiance $\partial L(\mathbf{o}, \mathbf{d}, t)/\partial t$ as

$$\frac{\partial L(\mathbf{o}, \mathbf{d}, t)}{\partial t} \propto \delta(\mathbf{o}, \mathbf{d}, t) L(\mathbf{o}, \mathbf{d}, t), \quad (4)$$

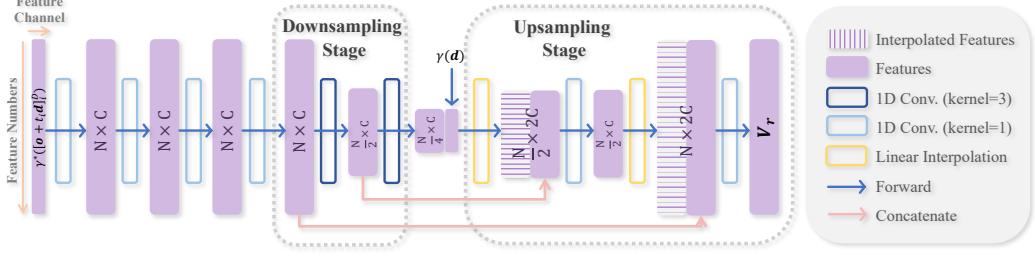


Figure 4: Architecture of Convolutional Ray Feature Extractor

where Dirac’s delta function only has value on the peak located at the intersection of the light and the surface. To implement this radiance ray in practice, we discretize Eq. 3 and Eq. 4 as

$$L(\mathbf{o}, \mathbf{d}, t) \approx \sum_{i=t}^D \frac{\Delta L(\mathbf{o}, \mathbf{d}, i)}{\Delta i} = \sum_{i=1}^D \mathbb{I}[\mathbf{o}, \mathbf{d}, i] L(\mathbf{o}, \mathbf{d}, i), \quad (5)$$

where the indicator function $\mathbb{I}[\mathbf{o}, \mathbf{d}, i]$ is equivalent to the delta function $\delta(\mathbf{o}, \mathbf{d}, t)$ in the non-continuous situation. As the indicator function $\mathbb{I}[\mathbf{o}, \mathbf{d}, t]$ is only non-zero at the sampling points near the surface, the range of summation can be extended from $[t, D]$ to $[1, D]$, hence the radiance is not relative with distance t . Then, the rendering equation can be split into two parts: $\mathbb{I}[\mathbf{o}, \mathbf{d}, i]$ that expresses the intersection between light ray from the indicator function and the scene and $L(\mathbf{o}, \mathbf{d}, i)$ records the radiance at different locations in the scene.

To approximate Eq.5 with a neural function, we first propose a convolutional Ray Feature Extractor $\mathbf{V}_r = \rho_F(\mathbf{o}, \mathbf{d}, \mathbf{t}; \theta_F)$, where $\mathbf{t} = [t_1, \dots, t_D]$ for ray information embedding. Then we denote the intersection position by the Geometry Attribute Network $\rho_G(\mathbf{V}_r; \theta_G)$ with the Unique Surface Constrain α for approximating single-peak indicator function \mathbb{I} . Meanwhile, the proposed Radiance Attribute Network $\rho_R(\mathbf{V}_r; \theta_R)$ conveys the intersection radiance. Finally, by substituting Eq. 5, we obtain an approximation expression of the scattering equation in radiance ray as

$$L(\mathbf{o}, \mathbf{d}) \propto \sum_{i=1}^D \alpha [\rho_G(\mathbf{V}_r; \theta_G)] \rho_R(\mathbf{V}_r; \theta_R). \quad (6)$$

Please note that compared to the traditional NeRF, the Radiance Attribute Network ρ_R of CeRF only needs to have the correct values on the surface, whereas the former needs to accurately predict the density values of all points in the entire space. Due to the introduction of the indicator function $\mathbb{I}[\mathbf{o}, \mathbf{d}, i]$, CeRF only selects the strongest signal points. Even if the network predicts noise values at other locations, they will be suppressed by the Unique Surface Constrain. Therefore, the network is more likely to learn corresponding relationships of unimodal distribution.

3.2 Convolutional Ray Feature Extractor

After taking the derivatives, only one peak exists in the target distribution of each ray. In order to design a network that parameterizes this model more effectively, we first consider treating the space of each ray as a complete scene representation and utilize 1D convolution to extract features for the one-hot spot of the intersection surface. These sparse interactions enable the network to have a more continuous input representation by parameter sharing and have been proven effective in many previous tasks [13]. Specifically, inspired by ray parameterize in radiance field [14], we define the sequential sampling points on one ray as a whole and model it as our scene representation $\mathbf{r} = [\mathbf{o} + t_i \mathbf{d}]_1^D$, to achieve more continuous and compact representation \mathbf{V}_r in Eq. 6. MLPs only map individual points into high-dimensional feature representations, which has limitations in modeling interdependency between adjacent points. CNN encodes the characteristics of neighbor points on the ray through the kernel convolution, allowing adjacent sampled points to be interdependent as shown in Figure 2. We propose the Convolutional Ray Feature Extractor $\rho_F(\mathbf{r}; \theta_F)$ with a U-shaped CNN architecture.

As shown in Figure 4, when a ray \mathbf{r} is input into the Convolutional Ray Feature Extractor, the spatial position encoding is low-dimension at first. Therefore, the first four layers of convolution

with a kernel of 1, as same as MLP, are applied to extract the high-dimensional deep features of individual points. The subsequent convolutional layers learn the local and global information of the light ray in the latent representation. The next two convolutional downsampling layers, which have a kernel size of 3 and stride of 2, consider the positions and features of neighboring points when producing feature values. This feature embedding architecture implicitly encodes the observation view direction from sampling positions along a ray. The convolutional structure compresses the number of sampling points along the light direction, producing bottleneck features. Through this layer-by-layer convolution, the receptive field gradually increases, and the bottleneck feature embeds the entire light ray, as shown in Figure 2. Finally, the last two upsampling layers utilize linear interpolation for upsampling and concatenate the interpolated feature vectors with relative vectors from the downsampling stage. The output ray feature encoding \mathbf{V}_r contains both position-dependent local features and light-dependent global features.

3.3 Neural radiance ray rendering

For the two parts in Eq.6, we use the Geometry Attribute Network ρ_G and Radiance Attribute Network ρ_R to predict the most possible position of the surface intersection and the radiance of the surface, respectively. The Geometry Attribute Network outputs the raw geometry attribute coefficient $\mathbf{s}_r = [s_1, s_2, \dots, s_D] = \rho_G(\mathbf{V}_r; \theta_G)$, which ranges from zero to one. Meanwhile, given the ray feature encoding \mathbf{V}_r as input, the Radiance Attribute Network outputs the radiance color $\mathbf{C}_r = [c_1, c_2, \dots, c_D] = \rho_R(\mathbf{V}_r; \theta_R)$.

Given a ray \mathbf{r} , the Geometry Attribute Network primarily learns the intersection position and local incoming radiance. To handle occlusion relationships among points on a light ray, we utilize the GRU [7], which has demonstrated strong performance in sequence learning. During each recursive step, all points sampled before the current one are used as priors. The GRU-based module outputs raw geometry attribute coefficient s_r through a three-layer fully connected network with *sigmoid* as the activation function of the last layer.

In order to solve the problem of ambiguity in the fitting problem, we introduce the Unique Surface Constrain α to approximate the indicator function in Eq. 5. Because the emissivity is at its maximum only at the first intersection point and is negligible for the rest of the ray, we propose to use *softmax* which is a differentiable and moderately variable manner. In order to handle the situation of background in the dataset, we propose an epipolar-based expectation mechanism β . We hypothesize an epipolar point \mathbf{p}_e at infinity with a raw geometry attribute coefficient s_e and the same color c_e as the background. When the inputs are all zeros, the outputs after *softmax* calculation are $1/N$. Therefore, we set $s_e = 1/N$ to handle the case where the space traversed by the ray is completely empty. The formula for geometry attribute coefficient of sampled points $\mathbf{w}_r = [w_1, w_2, \dots, w_D]$ and geometry attribute coefficient of the epipolar point w_e are shown in the following equation:

$$\mathbf{w}_{r,e} = \alpha(\mathbf{s}_{r,e}; \theta_\alpha) = softmax(\theta_\alpha * \mathbf{s}_{r,e}), \quad (7)$$

where $\mathbf{w}_{r,e} = concat(\mathbf{w}_r, w_e)$, $\mathbf{s}_{r,e} = concat(\mathbf{s}_r, s_e)$. To avoid numerical instability when all inputs are small, we multiply the raw geometry attribute coefficient $\mathbf{s}_{r,e}$ by a rescale parameter θ_α . With *sigmoid* as an activation function, Radiance Attribute Network ρ_R is a two-layer MLP outputting radiance color \mathbf{C}_r from ray feature embedding \mathbf{V}_r . Thus, the ray radiance is equivalent to the sum of the radiance variation of all sampling points and epipolar point

$$C(\mathbf{r}) = \beta(\mathbf{w}_{r,e}, \mathbf{c}_{r,e}) = \sum_{i=1}^N (w_i * c_i) + w_e * c_e. \quad (8)$$

Compared to traditional NeRF methods, our approach is a pure neural rendering method that omits the need to compute intermediate opacity and accumulated opacity values, resulting in easy implementation and straightforward usage.

3.4 Loss with Empty Space Regularization

For both the coarse and refinement stages, we utilize the square of \mathcal{L}_2 loss function as the costs. To handle the situation that a ray does not intersect with the surface of an object, we employ \mathcal{L}_1 distance on the geometry attribute coefficient of the sampling points on the ray as an Empty Space

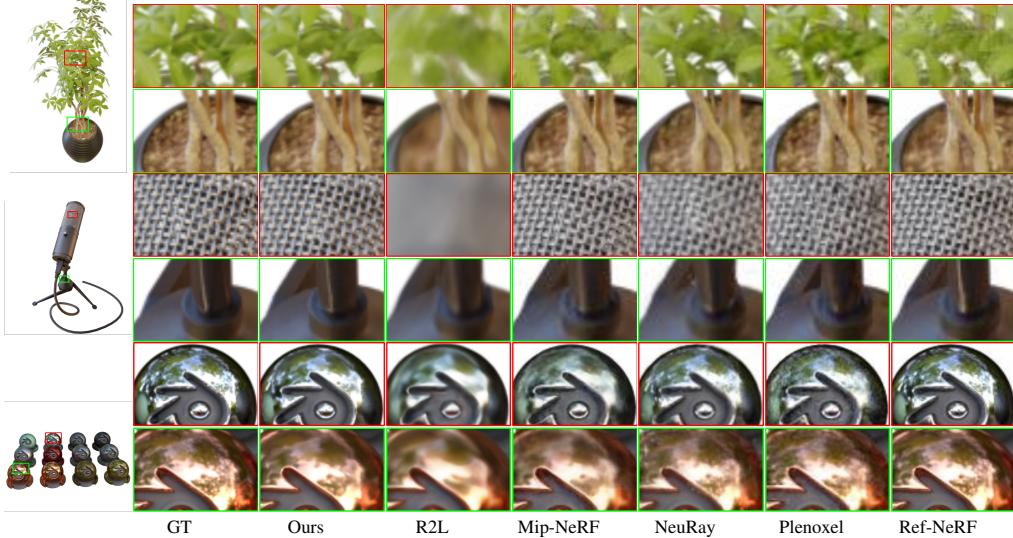


Figure 5: Qualitative results on the Blender dataset. CeRF excels at capturing complex geometric structures and fine details. In the ficus scene, our CeRF has the ability to accurately render the subsurface scattering of leaves and displays clear separation of layered and occluded branches. Despite Ref-NeRF performing well in scenes with strong specular reflections, it falls in accurately modeling light refraction and subsurface scattering. In the mic scene, CeRF demonstrates an aptitude for small mesh surfaces and accurately reconstructs subtle blue reflections on the base and small white highlights on the rods. In the metal ball scene, our specular reflections showcase rich and intricate details comparable to the results of Ref-NeRF without decoupled materials.

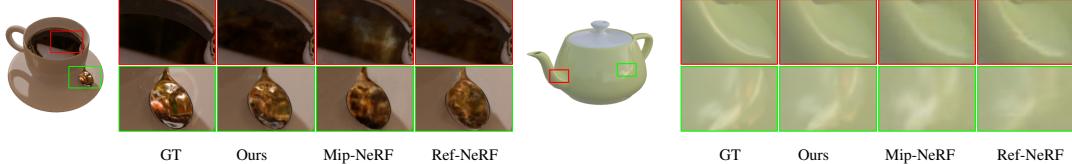


Figure 6: Qualitative results on the Shiny Blender dataset. In the lack of explicitly modeling mirror reflections, CeRF produces better results for the reflections in the coffee and spoon. For the ceramic teapot, our method is capable of rendering high-gloss patterns with more precise edge contours.

Regularization \mathcal{L}_e , to make the raw geometry attribute coefficients tend to lie on the epipolar line. Our final loss function for the model can be expressed as

$$\mathcal{L} = \sum_{\mathbf{r} \in \mathbf{r}_c, \mathbf{r}_f} \lambda \left\| C(\mathbf{r}) - \hat{C}(\mathbf{r}) \right\|_2^2 + \lambda_w \cdot \|\mathbf{w}_r\|_1, \quad (9)$$

where $\mathbf{w}_r = \alpha[\rho_G(\rho_F(\mathbf{r}))]$. $C(\mathbf{r})$ is the prediction and $\hat{C}(\mathbf{r})$ is the ground-truth. \mathbf{r}_c corresponds to the sampled rays in the coarse stage, while \mathbf{r}_f represents to the sampled rays in the fine stage.

4 Experiments

4.1 Experiments setting

Datasets and metrics Our proposed method was evaluated using both the Blender dataset [19] and the Shiny Blender dataset [31]. Both datasets provided 100 training images, 200 testing images, and 100 validation images with a resolution of 800×800 for each scene. To assess the performance of our method, we employed three commonly used metrics in the field of novel view synthesis: Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index (SSIM) [38], and Learned Perceptual Image Patch Similarity (LPIPS) [47].

Table 1: Comparison with SOTA on the Blender dataset

	PSNR↑	SSIM↑	LPIPS↓
Plenoxel [44]	31.71	0.958	0.049
R2L [34]	31.87	0.995	0.034
NeuRay [18]	32.35	0.960	0.048
Mip-NeRF [4]	33.11	0.962	0.042
Ref-NeRF [32]	33.99	0.966	0.038
CeRF	34.57	0.969	0.034

Table 2: Comparison with SOTA on the Shiny Blender dataset

	PSNR↑	SSIM↑	LPIPS↓
Plenoxel [44]	28.81	0.8967	0.1557
DVGO [26]	29.87	0.9258	0.1444
PhySG [46]	26.2133	0.9212	0.2077
Mip-NeRF [4]	29.7600	0.9417	0.0920
Ref-NeRF [32]	35.9617	0.9670	0.0587
CeRF	33.6015	0.9644	0.0597

Table 3: Ablation study on the drum scene

	PSNR↑	SSIM↑	LPIPS↓
baseline	26.4104	0.9420	0.0500
w/o ρ_F	26.7711	0.9453	0.0467
w/o ρ_G	27.1833	0.9497	0.0427
w/o α	23.6040	0.8996	0.1546
w/o β	27.1475	0.9509	0.0411
w/o \mathcal{L}_e	27.3703	0.9533	0.0391
CeRF	27.3900	0.9526	0.0391

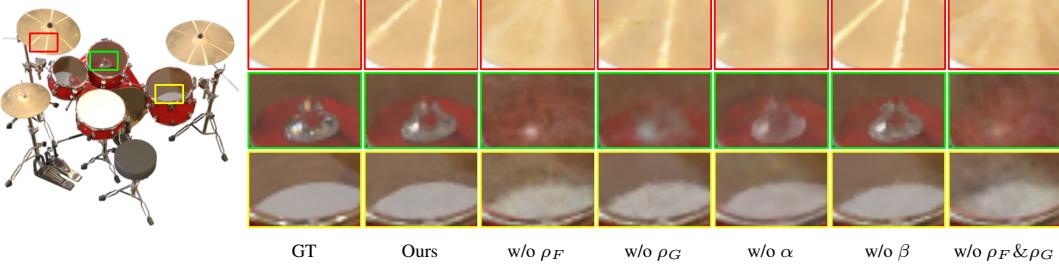


Figure 7: Ablation studies on drums scene

Hyperparameters setting The implementation of our CeRF is based on the Mip-NeRF [4] in PyTorch. To construct the networks, we use convolutional layers with a kernel size of 1 for all linear layers. In Convolutional Ray Feature Extractor, a convolutional layer with a kernel size of 3, a stride of 2 and padding of 1 is employed to perform downsampling in the ray direction. Linear interpolation is used in the upsampling layer and is followed by a layer of MLP to recover the channel size of the feature. The hidden state dimension of the GRU in ρ_G is set to 64 with one recurrent layer. For α , the rescale parameter θ_α is set to 10. Moreover, the value of the raw geometry attribute coefficient s_e at the epipolar is set to $1/N$, and the color c_e value at the epipolar is set to the same white color as the background. These parameters are selected based on our experimentation and analysis of the results.

Training details We utilize the Adam optimizer with a logarithmically annealed learning rate that ranged from 2×10^{-3} to 2×10^{-5} . The coefficients used for computing the running averages of the gradient and its square are set to 0.8 and 0.888, respectively. In our experiments, unless otherwise specified, we train the network using a batch size of 16384 for 500,000 iterations on 8 NIVIDIA RTX3080Ti GPUs. The weight $\lambda = 0.1$ is assigned to the coarse stage, and it is assigned to 1 for the fine stage. The suggested weight for the Empty Space Regularization λ_w is set to 0.01.

4.2 Comparison with State-of-the-Art Methods

In this section, we compare our CeRF with other neural field representations ³, including NeuRay [18], Mip-NeRF [4], Plenoxel [44], Ref-NeRF [32], as well as R2L [34]. The qualitative and quantitative results on the Blender dataset are presented in Figure 5 and Table 1, respectively. Compared to the baseline, our method outperforms in all metrics on the Blender dataset. Without requiring complex distillation training, CeRF achieves an 8.5% increase in PSNR compared to R2L. Furthermore, CeRF also outperforms the SOTA, Ref-NeRF, in all three metrics.

Figure 6 and Table 2 demonstrate the qualitative and quantitative results of the Shiny Blender dataset. CeRF achieves the second rank after Ref-NeRF and outperforms all the other methods. However, our method still outperforms the current state-of-the-art method, Ref-NeRF, in scenes with subsurface reflections, such as coffee and teapot. Furthermore, we choose not to model reflections because our goal was to create a model that is as simple and general as possible, and we opt for a pure implicit representation. Moreover, we believe that designing structures with separated materials, such as Ref-NeRF [32] and I2-SDF [48], can further enhance network performance.

³Results from https://huggingface.co/nrtf/nerf_factory.

Table 4: Different ρ_F setting on the drum scene

	PSNR↑	SSIM↑	LPIPS↓	GPU Mem (GB)↓	Training Speed (s/epoch)↓
W128U4K3D8	26.2041	0.9387	0.0580	8.649	190
W256U4K5D8	25.4161	0.9289	0.0739	11.109	300
W256U2K3D8	27.1174	0.9497	0.0424	11.867	350
W256U6K3D8	27.3028	0.9521	0.0403	10.577	330
W256U4K3D6	27.1732	0.9507	0.0417	9.351	310
W256U4K3D10	27.4342	0.9538	0.0379	11.849	380
W256U4K3D8(ours)	27.3900	0.9526	0.0391	11.039	340

Table 5: ρ_G variance on the lego scene

	PSNR↑	SSIM↑	LPIPS↓
$\theta_\alpha = 1$	23.0863	0.8916	0.1452
$\theta_\alpha = 20$	27.2061	0.9526	0.0394
learnable θ_α	27.3559	0.9526	0.0394
$s_e = 2/N$	27.2547	0.9518	0.0402
learnable s_e	14.9650	0.8358	0.1665
$\theta_\alpha = 10, s_e = 1/N$	27.3900	0.9526	0.0391

4.3 Ablation studies

In order to evaluate the effectiveness of the individual components in our proposed CeRF model, we conducted ablation experiments on the drum scene using half resolution, a batch size of 14,744, and an iteration step size of 250,000. Figure 7 and Table 3 present the qualitative and quantitative results of our ablation experiments, which demonstrate the effectiveness of our proposed modules from both a metric and visual perspective.

The absence of the **Convolutional Ray Feature Extractor** ρ_F results in the loss of directionally-correlated visual representations, making it difficult to extract the specular properties of the scene. Removing the **Radiance Attribute Network** ρ_G causes obstacles in capturing information about the sampled light point sequences and results in noise on the rendered object surface. We found that removing the **Unique Surface Constraint** α gives the worst results, as the radiance of the highlight cannot be correctly reconstructed. Ablation experiments based on the **Epipolar Expectation** show that if β is removed, the model will struggle to accurately compute the radiance of the empty space that rays pass through in the background of the dataset. This would have a negative impact on the accuracy and quality of the rendered output. In the absence of both the ρ_F and ρ_G modules, the mirror reflections in the drum scene become unacceptably blurred, as shown in Figure 7. The **Empty Space Regularization** L_e ensures the correct matching of empty rays on epipolar lines. Although the metric may slightly decrease, we demonstrate through depth visualization that regularization is important in fitting geometry at infinity, which is described in the supplementary materials.

4.4 Parameters experiments

In this subsection, we conduct a series of experiments to explore the hyperparameters configurations for ρ_F , ρ_G , α , and β . The testing was performed on a half-resolution lego scene with an iteration step of 250,000. The notations in Table 4, such as W128U4K3D8, signify a setting with a width of 128, 4 layers of up-sampling and down-sampling, a kernel size of 3, and a total module depth of 8.

Our experimental results, as presented in Table 4, demonstrate that our parameter settings achieve a balance between resource consumption and effectiveness. Although the setting of a total module depth of 8 achieves better results, it requires significantly more computational resources.

Table 5 displays the results of scaling θ_α and s_e in CeRF. We proposed a new mechanism to achieve learnable s_e . However, experiments showed that this method was not effective, and thus fixed s_e was used instead. Further details can be found in the supplementary materials.

5 Conclusion and discussion

In conclusion, we proposed CeRF as a novel approach for new view synthesis based on the derivatives of the radiance model. Our proposed model, with a CNN-based feature extractor and GRU-based neural rendering, has reduced the learning difficulties for implicit representation and recovered high-quality rendering results even under complex geometry. Moreover, CeRF can be easily integrated with other deep learning frameworks, making it a promising alternative to NeRF for a variety of high-level tasks. Extensive experiments showed that our proposed CeRF yields promising results with a simple network structure.

Discussion. While CeRF extends linear layers to convolution layers, CNNs are still a simple and preliminary type of network. There are still many ways to optimize convolutional operations, as well as utilizing superior structure like Transformers to improve current models.

Limitations. The convolution operation increases the computational complexity, and the RNN introduces more parameters. Therefore, the calculation speed and cost are higher than NeRF. There is

no modeling of physically-based materials, nor is there any supervision or constraints placed on the geometry, such as depths and normals, or extended to semantic information. Further future work is needed to demonstrate the generality of this framework.

Negative Social Impact. The training of the program requires a long time to run on a GPU, and the power consumption is not environmentally friendly.

References

- [1] Benjamin Attal, Jia-Bin Huang, Christian Richardt, Michael Zollhoefer, Johannes Kopf, Matthew O’Toole, and Changil Kim. Hyperreal: High-fidelity 6-dof video with ray-conditioned sampling. *arXiv preprint arXiv:2301.02238*, 2023.
- [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.
- [3] Chong Bao, Yinda Zhang, Bangbang Yang, Tianxing Fan, Zesong Yang, Hujun Bao, Guofeng Zhang, and Zhaopeng Cui. Sine: Semantic-driven image-based nerf editing with prior-guided editing field. *arXiv preprint arXiv:2303.13277*, 2023.
- [4] 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, 2021.
- [5] Jonathan T Barron, Ben Mildenhall, Dor Verbin, Pratul P Srinivasan, and Peter Hedman. Zip-nerf: Anti-aliased grid-based neural radiance fields. *arXiv preprint arXiv:2304.06706*, 2023.
- [6] Anpei Chen, Zexiang Xu, Andreas Geiger, Jingyi Yu, and Hao Su. Tensorf: Tensorial radiance fields. In *European Conference on Computer Vision (ECCV)*, 2022.
- [7] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*, 2014.
- [8] James D Foley, Foley Dan Van, Andries Van Dam, Steven K Feiner, and John F Hughes. Computer graphics: principles and practice. volume 12110, chapter 29, page 790. Addison-Wesley Professional, 1996.
- [9] Steven J. Gortler, Radek Grzeszczuk, Richard Szeliski, and Michael F. Cohen. The lumigraph. *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*, 1996.
- [10] Juxiang Gu, Zhenhua Wang, Jason Kuen, Liyang Ma, Amir Shahroudy, Bing Shuai, Ting Liu, Xingxing Wang, Gang Wang, Jianfei Cai, et al. Recent advances in convolutional neural networks. *Pattern recognition*, 77:354–377, 2018.
- [11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
- [12] Xin Huang, Qi Zhang, Ying Feng, Xiaoyu Li, Xuan Wang, and Qing Wang. Local implicit ray function for generalizable radiance field representation. *arXiv preprint arXiv:2304.12746*, 2023.
- [13] Serkan Kiranyaz, Onur Avci, Osama Abdeljaber, Turker Ince, Moncef Gabbouj, and Daniel J Inman. 1d convolutional neural networks and applications: A survey. *Mechanical systems and signal processing*, 151:107398, 2021.
- [14] Abiramy Kuganesan, Shih-yang Su, James J Little, and Helge Rhodin. Unerf: Time and memory conscious u-shaped network for training neural radiance fields. *arXiv preprint arXiv:2206.11952*, 2022.
- [15] Marc Levoy and Pat Hanrahan. Light field rendering. *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*, 1996.
- [16] David B Lindell, Dave Van Veen, Jeong Joon Park, and Gordon Wetzstein. Bacon: Band-limited coordinate networks for multiscale scene representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16252–16262, 2022.
- [17] Jia-Wei Liu, Yan-Pei Cao, Weijia Mao, Wenqiao Zhang, David Junhao Zhang, Jussi Keppo, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. Devrf: Fast deformable voxel radiance fields for dynamic scenes. *ArXiv*, abs/2205.15723, 2022.
- [18] Yuan Liu, Sida Peng, Lingjie Liu, Qianqian Wang, Peng Wang, Christian Theobalt, Xiaowei Zhou, and Wenping Wang. Neural rays for occlusion-aware image-based rendering. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7814–7823, 2021.
- [19] 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.
- [20] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Transactions on Graphics (ToG)*, 41(4):1–15, 2022.
- [21] Yicong Peng, Yichao Yan, Shengqi Liu, Y. Cheng, Shanyan Guan, Bowen Pan, Guangtao Zhai, and Xiaokang Yang. Cagenerf: Cage-based neural radiance field for generalized 3d deformation and animation. In *Neural Information Processing Systems*, 2022.
- [22] Marie-Julie Rakotosaona, Fabian Manhardt, Diego Martin Arroyo, Michael Niemeyer, Abhijit Kundu, and Federico Tombari. Nerfmeshing: Distilling neural radiance fields into geometrically-accurate 3d meshes, 2023.
- [23] Ari Silvennoinen and Peter-Pike Sloan. Ray guiding for production lightmap baking. In *SIGGRAPH Asia 2019 Technical Briefs*, pages 91–94. 2019.

- [24] Vincent Sitzmann, Semon Rezchikov, 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.
- [25] Mohammed Suhail, Carlos Esteves, Leonid Sigal, and Ameesh Makadia. Light field neural rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8269–8279, 2022.
- [26] Cheng Sun, Min Sun, and Hwann-Tzong Chen. Improved direct voxel grid optimization for radiance fields reconstruction. *ArXiv*, abs/2206.05085, 2022.
- [27] Mukund Varma T, Peihao Wang, Xuxi Chen, Tianlong Chen, Subhashini Venugopalan, and Zhangyang Wang. Is attention all nerf needs? *arXiv preprint arXiv:2207.13298*, 2022.
- [28] Jiaxiang Tang, Xiaokang Chen, Jingbo Wang, and Gang Zeng. Compressible-composable nerf via rank-residual decomposition. *ArXiv*, abs/2205.14870, 2022.
- [29] Jiapeng Tang, Lev Markhasin, Bi Wang, Justus Thies, and Matthias Nießner. Neural shape deformation priors. *ArXiv*, abs/2210.05616, 2022.
- [30] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.
- [31] Dor Verbin, Peter Hedman, Ben Mildenhall, Todd Zickler, Jonathan T Barron, and Pratul P Srinivasan. Ref-nerf: Structured view-dependent appearance for neural radiance fields. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5481–5490. IEEE, 2022.
- [32] Dor Verbin, Peter Hedman, Ben Mildenhall, Todd E. Zickler, Jonathan T. Barron, and Pratul P. Srinivasan. Ref-nerf: Structured view-dependent appearance for neural radiance fields. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5481–5490, 2021.
- [33] Bing Wang, Lu Chen, and Bo Yang. Dm-nerf: 3d scene geometry decomposition and manipulation from 2d images. *arXiv preprint arXiv:2208.07227*, 2022.
- [34] Huan Wang, Jian Ren, Zeng Huang, Kyle Olszewski, Menglei Chai, Yun Fu, and Sergey Tulyakov. R2l: Distilling neural radiance field to neural light field for efficient novel view synthesis. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXI*, pages 612–629. Springer, 2022.
- [35] Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. *NeurIPS*, 2021.
- [36] Peng Wang, Yuan Liu, Zhaoxi Chen, Lingjie Liu, Ziwei Liu, Taku Komura, Christian Theobalt, and Wenping Wang. F2-nerf: Fast neural radiance field training with free camera trajectories. *CVPR*, 2023.
- [37] Peng Wang, Yuan Liu, Guying Lin, Jiatao Gu, Lingjie Liu, Taku Komura, and Wenping Wang. Progressively-connected light field network for efficient view synthesis. *arXiv preprint arXiv:2207.04465*, 2022.
- [38] Zhou Wang, Alan Conrad Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13:600–612, 2004.
- [39] Suttisak Wizadwongsu, 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.
- [40] Yu-Shiang Wong and Niloy J Mitra. Factored neural representation for scene understanding. *arXiv preprint arXiv:2304.10950*, 2023.
- [41] Tianhao Wu, Hanxue Liang, Fangcheng Zhong, Gernot Riegler, Shimon Vainer, and Cengiz Oztireli. Implicit surface reconstruction for semi-transparent and thin objects with decoupled geometry and opacity. *arXiv preprint arXiv:2303.10083*, 2023.
- [42] 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.
- [43] Fukun Yin, Wen Jing Liu, Zilong Huang, Pei Cheng, Tao Chen, and Gang Yu. Coordinates are not lonely - codebook prior helps implicit neural 3d representations. *ArXiv*, abs/2210.11170, 2022.
- [44] Alex Yu, Sara Fridovich-Keil, Matthew Tancik, Qinhong Chen, Benjamin Recht, and Angjoo Kanazawa. Plenoxels: Radiance fields without neural networks. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5491–5500, 2021.
- [45] Zehao Yu, Songyou Peng, Michael Niemeyer, Torsten Sattler, and Andreas Geiger. Monosdf: Exploring monocular geometric cues for neural implicit surface reconstruction. *arXiv preprint arXiv:2206.00665*, 2022.
- [46] Kai Zhang, Fujun Luan, Qianqian Wang, Kavita Bala, and Noah Snavely. Physg: Inverse rendering with spherical gaussians for physics-based material editing and relighting. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5449–5458, 2021.
- [47] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 586–595, 2018.
- [48] Jingsen Zhu, Yuchi Huo, Qi Ye, Fujun Luan, Jifan Li, Dianbing Xi, Lisha Wang, Rui Tang, Wei Hua, Hujun Bao, et al. I2-sdf: Intrinsic indoor scene reconstruction and editing via raytracing in neural sdf. *arXiv preprint arXiv:2303.07634*, 2023.

Supplementary Material for CeRF: Convolutional Neural Radiance Fields for New View Synthesis with Derivatives of Ray Modeling

1 A Empty Space Regularization

2 NeRF utilizes classical volume rendering to render the 3D radiance field into a 2D image, as

$$C(\mathbf{r}) = \sum_{i=1}^N T_i (1 - \exp(-\sigma_i \delta_i)) \mathbf{c}_i, \text{ where } T_i = \exp \left(- \sum_{j=1}^{i-1} \sigma_j \delta_j \right), \quad (1)$$

3 where σ_i is the volume density of the i th point along the ray, δ_i is the distance between adjacent
4 samples. The accumulated transmittance T_i introduced in the formula constrains the sum of the volume
5 density over the ray to be no greater than 1 and allows NeRF to handle the occlusion relationship
6 between objects to some extent. In cases where rays pass through empty space, NeRF can learn that
7 the volume density at all sampled points is 0, resulting in a total sum of 0 for the entire ray.

8 CeRF proposes a Neural Radiance Ray Rendering technique that can effectively handle complex
9 layers between objects. Unlike the rendering method used in NeRF, the neural renderer does not
10 explicitly model the rendering function, allowing it to fit more complex optical phenomena such as
11 subsurface scattering. When dealing with rays in empty space, the neural renderer does not impose
12 any constraints on the location of selected points along these rays. However, by implementing Unique
13 Surface Constrain, which is implemented with *softmax*, we force the sum of radiance coefficients
14 for all sampled points on the rays in empty space to be 1 instead of learning an incorrect result where
15 this sum is 0. In other words, to satisfy this constraint imposed by softmax, random points are chosen
16 along these rays by the Neural Radiance Ray Renderer.

17 As depicted in Figure 1, the bluer area indicates a larger depth value and vice versa. We set the
18 epipolar point at a depth of 120, where the maximum depth of objects in the scene is 6. It is evident
19 that CeRF’s results are displayed as white in empty spaces whereas missing Loss with Empty Space
20 Regularization results appear in varying shades of gray at their respective positions. This demonstrates
21 that rays directed towards empty spaces select points located farthest away - i.e., the epipolar point.
22 The epipolar mechanism and Loss with Empty Space Regularization functioned as expected.

23 B CeRF with learnable s_e

24 For learnable s_e , as shown in the Figure 2, we try a new mechanism to compute the corresponding s_e
25 value for each ray. Specifically, we obtain the bottleneck layer features obtained in Convolutional
26 Ray Feature Extractor for each ray and input that feature into a Learnable Epipolar Decoder φ .
27 The structure of Learnable Epipolar Decoder consists of two MLP layers with feature channels of
28 256 and 32 respectively, along with an additional sigmoid.

29 Our main text explains that our U-shaped CNN architecture treats each ray’s space as a complete
30 scene representation for feature extraction. Therefore, we believe that these compact bottleneck layer
31 features contain simple information such as whether the ray passes through an object. However,
32 through the experiment, we discover that it is challenging to extract information about a single
33 epipolar point from the bottleneck layer features. This is because these features encode complex
34 spatial information for an entire ray. Despite compressing feature channels to 1/4 of the original
35 ray features (which has N 256-dimensional features), the bottleneck layer features still have high
36 dimensionality and are difficult to reduce using a small MLP. As such, we opt for a fixed-value
37 s_e instead of the learnable s_e mechanism. Our experiment shows that this approach resulted in
38 significant performance improvements with minimal additional computational overhead.

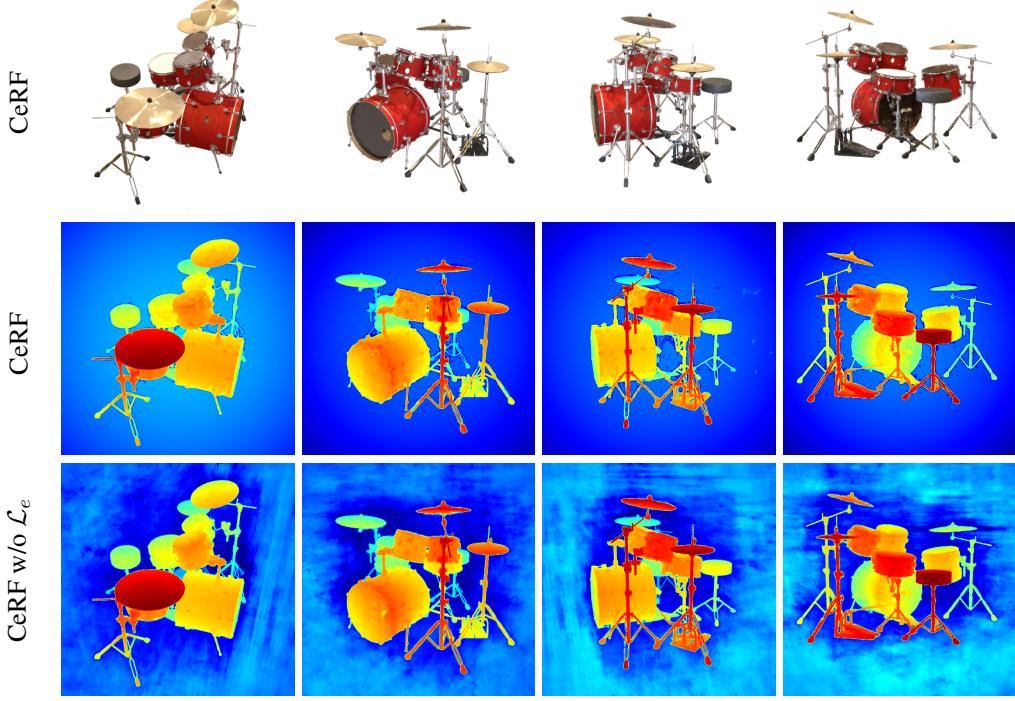


Figure 1: Ablation of Empty Space Regularization. The first line is rendered images. The second line is depth maps generated by full version of CeRF and the third line is depth maps generated by CeRF without Empty Space Regularization.

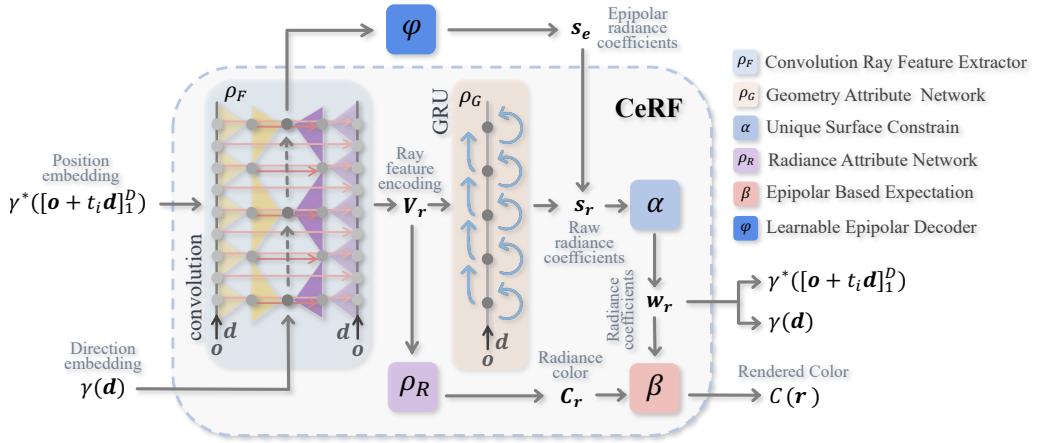


Figure 2: Illustration of CeRF with learnable s_e

39 C Geometry information

40 Our experiments demonstrate that CeRF is able to learn the geometric information of the objects in
 41 the scene, despite the fact that the features extracted in 3D space are incorporated along the ray. We
 42 obtain the distance map by normalizing the radiance coefficients and then multiplying them with the
 43 distances of the sampled points. As shown in Figure 3 and Figure 4, the mesh is extracted from the
 44 model trained from CeRF, and it can be seen that the rich details on the objects are well reconstructed
 45 on the mesh, such as the protrusions on the thin horizontal bar in the Lego scene, and hollow parts

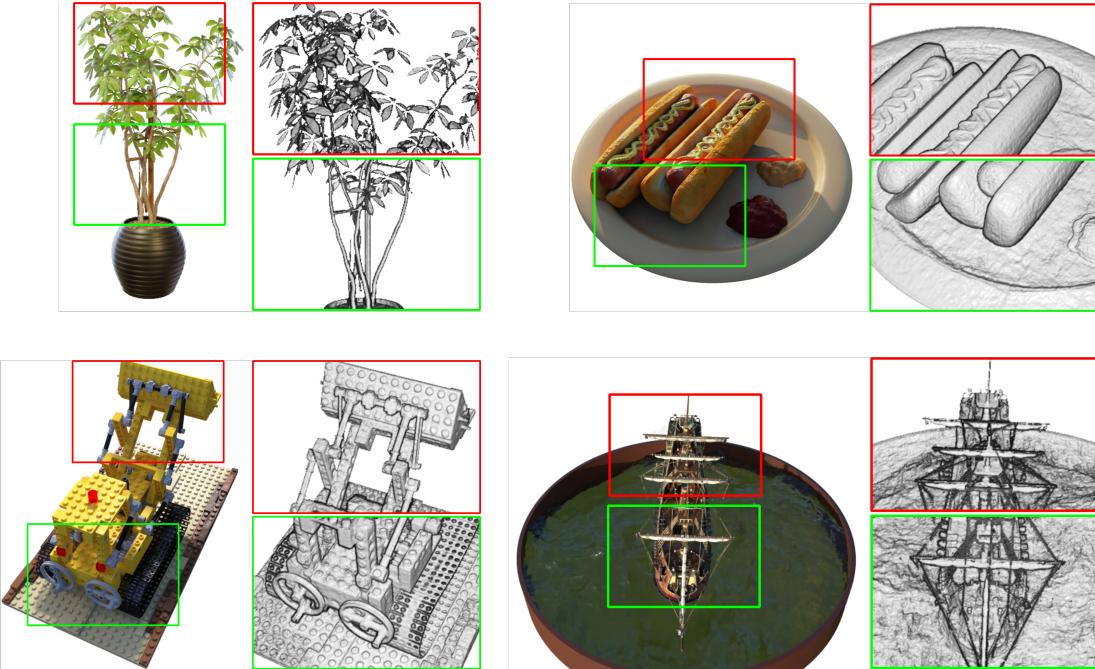


Figure 3: Rendered images and extracted meshes from CeRF. Unlike light field, our method fuses the features along the ray direction but still retains the information of each individual point, so that the mesh can be easily obtained, which retains the advantage of radiance field.

46 of the track. CeRF can learn the geometric information well and extract the geometric information
 47 easily, which retains the advantage of radiance field.

48 **D Additional results**

49 We additionally provide a per-scene comparison between CeRF and other methods, where Table 1, 2
 50 and 3 show the metrics on the Blender dataset, and Table 4, 5 and 6 reflect the quantitative results on
 51 the Shiny Blender Dataset. We also included comparison videos with additional and other methods in
 52 our supplementary material, as well as ablative videos of our module.

Table 1: Per-scene test set PSNRs on the Blender dataset

	chair	drums	ficus	hotdog	lego	materials	mic	ship
Plenoxel	33.98	25.35	31.83	36.43	34.10	29.14	33.26	29.62
R2L	36.71	26.03	28.63	38.07	32.53	30.20	32.80	29.98
Mip-NeRF	35.12	25.36	33.19	37.34	35.92	30.64	36.76	30.52
Ref-NeRF	35.83	25.79	33.91	37.72	36.25	35.41	36.76	30.28
CeRF	35.99	26.41	37.32	38.24	37.79	33.22	36.35	31.23

Table 2: Per-scene test set SSIMs on the Blender dataset

	chair	drums	ficus	hotdog	lego	materials	mic	ship
Plenoxel	0.977	0.933	0.976	0.980	0.975	0.949	0.985	0.890
R2L	0.999	0.988	0.996	0.999	0.994	0.992	0.997	0.986
Mip-NeRF	0.981	0.933	0.980	0.982	0.980	0.959	0.992	0.885
Ref-NeRF	0.984	0.937	0.983	0.984	0.981	0.983	0.992	0.880
CeRF	0.984	0.945	0.990	0.985	0.985	0.975	0.991	0.897

Table 3: Per-scene test set LPIPS on the Blender dataset

	chair	drums	ficus	hotdog	lego	materials	mic	ship
Plenoxel	0.031	0.067	0.026	0.037	0.028	0.057	0.015	0.134
Mip-NeRF	0.020	0.064	0.021	0.026	0.018	0.040	0.008	0.135
Ref-NeRF	0.017	0.059	0.019	0.022	0.018	0.022	0.007	0.139
CeRF	0.017	0.054	0.012	0.021	0.014	0.028	0.008	0.122

Table 4: Per-scene test set PSNRs on the Shiny Blender dataset

	teapot	toaster	coffee	helmet	car	ball
Plenoxel	44.25	19.51	31.55	26.94	26.11	24.52
DVGO	44.79	22.18	31.48	27.75	26.90	26.13
PhySG	35.83	18.59	23.71	27.51	24.40	27.24
Mip-NeRF	46.00	22.37	30.36	27.39	26.50	25.94
Ref-NeRF	47.90	25.70	34.21	29.68	30.82	47.46
CeRF	47.17	26.82	32.44	29.53	27.99	37.66

Table 5: Per-scene test set SSIMs on the Shiny Blender dataset

	teapot	toaster	coffee	helmet	car	ball
Plenoxel	0.996	0.772	0.963	0.913	0.905	0.832
DVGO	0.996	0.848	0.962	0.932	0.916	0.901
PhySG	0.990	0.805	0.922	0.953	0.910	0.947
Mip-NeRF	0.997	0.891	0.966	0.939	0.922	0.935
Ref-NeRF	0.998	0.922	0.974	0.958	0.955	0.995
CeRF	0.997	0.937	0.972	0.958	0.936	0.987

Table 6: Per-scene test set LPIPS on the Shiny Blender dataset

	teapot	toaster	coffee	helmet	car	ball
Plenoxel	0.016	0.244	0.144	0.170	0.086	0.274
DVGO	0.019	0.220	0.141	0.154	0.081	0.253
PhySG	0.022	0.194	0.15	0.089	0.091	0.700
Mip-NeRF	0.008	0.123	0.086	0.108	0.059	0.168
Ref-NeRF	0.004	0.095	0.078	0.075	0.041	0.059
CeRF	0.005	0.077	0.081	0.069	0.052	0.073

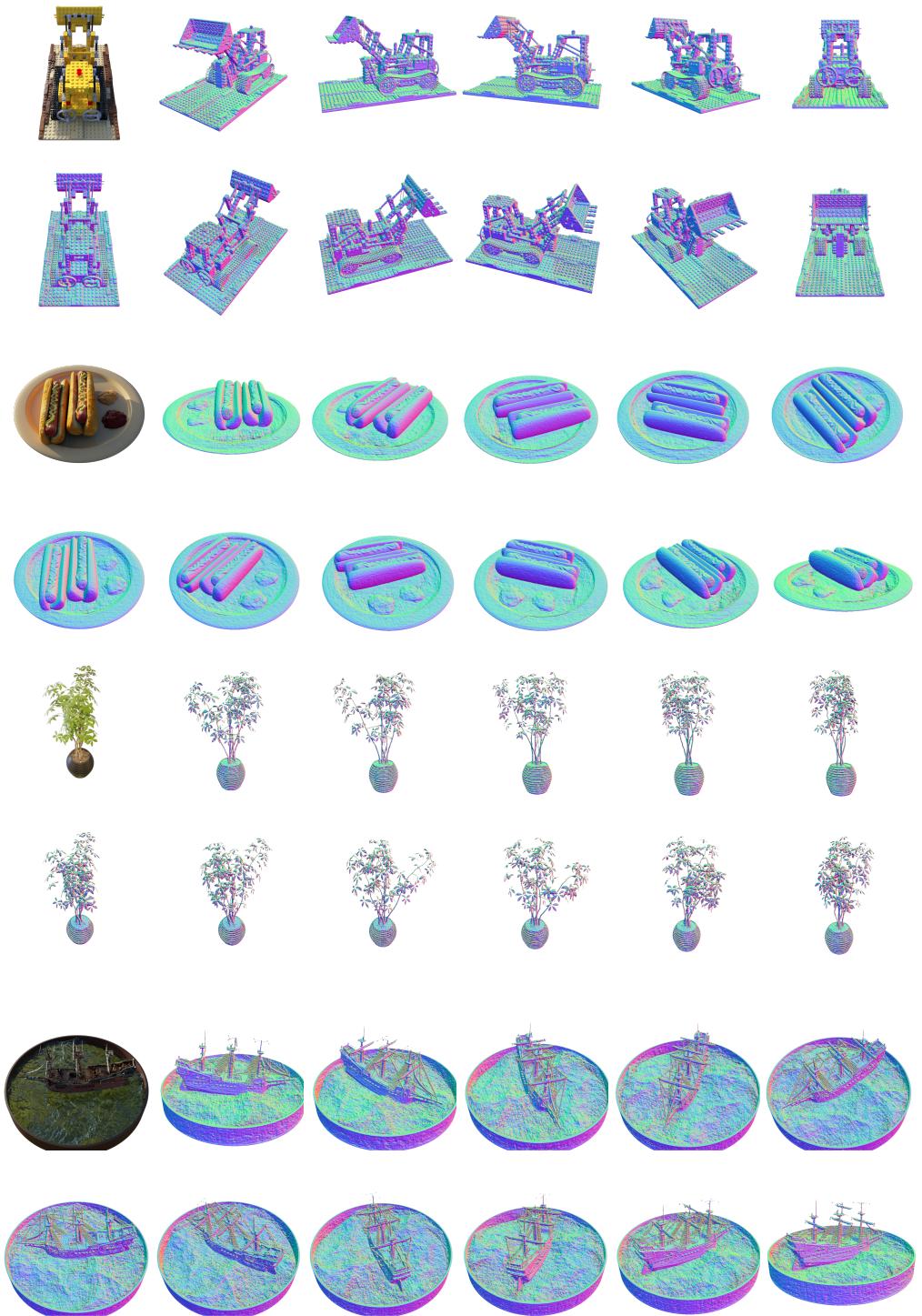


Figure 4: Visualization of extracted mesh results. The surface reconstructed by our method is colored with normal vectors. The color image on the left is the ground truth.