

---

# NERF++: ANALYZING AND IMPROVING NEURAL RADIANCE FIELDS

**Kai Zhang**  
Cornell Tech

**Gernot Riegler**  
Intel Labs

**Noah Snavely**  
Cornell Tech

**Vladlen Koltun**  
Intel Labs

## ABSTRACT

Neural Radiance Fields (NeRF) achieve impressive view synthesis results for a variety of capture settings, including 360° capture of bounded scenes and forward-facing capture of bounded and unbounded scenes. NeRF fits multi-layer perceptrons (MLPs) representing view-invariant opacity and view-dependent color volumes to a set of training images, and samples novel views based on volume rendering techniques. In this technical report, we first remark on radiance fields and their potential ambiguities, namely the *shape-radiance ambiguity*, and analyze NeRF’s success in avoiding such ambiguities. Second, we address a parametrization issue involved in applying NeRF to 360° captures of objects within large-scale, unbounded 3D scenes. Our method improves view synthesis fidelity in this challenging scenario. Code is available at <https://github.com/Kai-46/nerfplusplus>.

## 1 INTRODUCTION

Recall your last vacation where you captured a few photos of your favorite place. Now at home you wish to walk around in this special place again, if only virtually. This requires you to render the same scene from different, freely placed viewpoints in a possibly unbounded scene. This *novel view synthesis* task is a long-standing problem in computer vision and graphics (Chen & Williams, 1993;Debevec et al., 1996; Levoy & Hanrahan, 1996; Gortler et al., 1996; Shum & Kang, 2000).

Recently, learning-based methods have led to significant progress towards photo-realistic novel view synthesis. The method of Neural Radiance Fields (NeRF), in particular, has attracted significant attention (Mildenhall et al., 2020). NeRF is an implicit MLP-based model that maps 5D vectors—3D coordinates plus 2D viewing directions—to opacity and color values, computed by fitting the model to a set of training views. The resulting 5D function can then be used to generate novel views with conventional volume rendering techniques.

In this technical report, we first present an analysis of potential failure modes in NeRF, and an analysis of why NeRF avoids these failure modes in practice. Second, we present a novel spatial parameterization scheme that we call *inverted sphere parameterization* that allows NeRF to work on a new class of captures of unbounded scenes.

In particular, we find that in theory, optimizing the 5D function from a set of training images can encounter critical degenerate solutions that fail to generalize to novel test views, in the absence of any regularization. Such phenomena are encapsulated in the *shape-radiance ambiguity* (Figure 1, left), wherein one can fit a set of training images perfectly for an arbitrary incorrect geometry by a suitable choice of outgoing 2D radiance at each surface point. We empirically show that the specific MLP structure used in NeRF plays an important role in avoiding such ambiguities, yielding an impressive ability to synthesize novel views. Our analysis offers a new view into NeRF’s impressive success.

We also address a spatial parameterization issue that arises in challenging scenarios involving 360° captures around objects within unbounded environments (Figure 1, right). For 360° captures, NeRF assumes that the entire scene can be packed into a bounded volume, which is problematic for large-scale scenes: either we fit a small part of the scene into the volume and sample it in detail, but completely fail to capture background elements; or, we fit the full scene into the volume and lack detail everywhere due to limited sampling resolution. We propose a simple yet effective solution that separately models foreground and background, addressing the challenge of modeling unbounded 3D

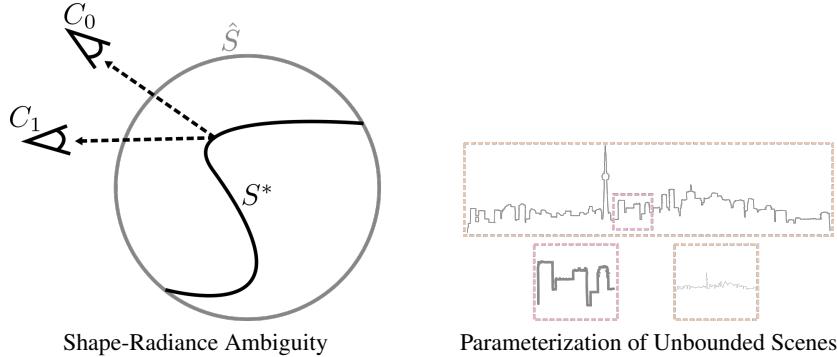


Figure 1: *Shape-radiance ambiguity* (left) and *parameterization of unbounded scenes* (right). **Shape-radiance ambiguity:** our theoretical analysis shows that, in the absence of explicit or implicit regularization, a set of training images can be fit independently of the recovered geometry (e.g., for incorrect scene geometry  $\hat{S}$  rather than correct geometry  $S^*$ ) by exploiting view-dependent radiance to simulate the effect of the correct geometry. **Parameterization of unbounded scenes:** with standard parameterization schemes, either (1) only a portion of the scene is modeled (red outline), leading to significant artifacts in background elements, or (2) the full scene is modeled (orange outline), which leads to an overall loss of details due to finite sampling resolution.

background content with an inverted sphere scene parameterization. We show improved quantitative and qualitative results on real-world captures from the Tanks and Temples dataset (Knapitsch et al., 2017) and from the light field dataset of Yücer et al. (2016).

In summary, we present an analysis on how NeRF manages to resolve the shape-radiance ambiguity, as well as a remedy for the parameterization of unbounded scenes in the case of 360° captures.

## 2 PRELIMINARIES

Given posed multi-view images of a static scene, NeRF reconstructs an opacity field  $\sigma$  representing a soft shape, along with a radiance field  $\mathbf{c}$  representing view-dependent surface texture. Both  $\sigma$  and  $\mathbf{c}$  are represented implicitly as multi-layer perceptrons (MLPs); the opacity field is computed as a function of 3D position,  $\mathbf{x} \in \mathbb{R}^3$ , and the radiance field is parameterized by both 3D position and viewing direction,  $\mathbf{d} \in \mathbb{S}^2$  (i.e., the set of unit 3-vectors). Hence, we use  $\sigma(\mathbf{x})$  to refer to opacity as a function of position, and  $\mathbf{c}(\mathbf{x}, \mathbf{d})$  to refer to radiance as a function of position and viewing direction.

Ideally,  $\sigma$  should peak at the ground-truth surface location for opaque materials, in which case  $\mathbf{c}$  reduces to the surface light field (Wood et al., 2000). Given  $n$  training images, NeRF uses stochastic gradient descent to optimize  $\sigma$  and  $\mathbf{c}$  by minimizing the discrepancy between the ground truth observed images  $I_i$ , and the predicted images  $\hat{I}_i(\sigma, \mathbf{c})$  rendered from  $\sigma$  and  $\mathbf{c}$  at the same viewpoints:

$$\min_{\sigma, \mathbf{c}} \frac{1}{n} \sum_{i=1}^n \|I_i - \hat{I}_i(\sigma, \mathbf{c})\|_2^2. \quad (1)$$

The implicit volumes  $\sigma$  and  $\mathbf{c}$  are ray-traced to render each pixel of  $\hat{I}(\sigma, \mathbf{c})$  (Kajiya & Von Herzen, 1984). For a given ray  $\mathbf{r} = \mathbf{o} + t\mathbf{d}$ ,  $\mathbf{o} \in \mathbb{R}^3$ ,  $\mathbf{d} \in \mathbb{S}^2$ ,  $t \in \mathbb{R}^+$ , its color is determined by the integral

$$\mathbf{C}(\mathbf{r}) = \int_{t=0}^{\infty} \sigma(\mathbf{o} + t\mathbf{d}) \cdot \mathbf{c}(\mathbf{o} + t\mathbf{d}, \mathbf{d}) \cdot e^{-\int_{s=0}^t \sigma(\mathbf{o} + s\mathbf{d}) ds} dt. \quad (2)$$

To compensate for the network’s spectral bias and to synthesize sharper images, NeRF uses a positional encoding  $\gamma$  that maps  $\mathbf{x}$  and  $\mathbf{d}$  to their Fourier features (Tancik et al., 2020):

$$\gamma^k : \mathbf{p} \rightarrow (\sin(2^0 \mathbf{p}), \cos(2^0 \mathbf{p}), \sin(2^1 \mathbf{p}), \cos(2^1 \mathbf{p}), \dots, \sin(2^k \mathbf{p}), \cos(2^k \mathbf{p})), \quad (3)$$

where  $k$  is a hyper-parameter specifying the dimensionality of the Fourier feature vector.

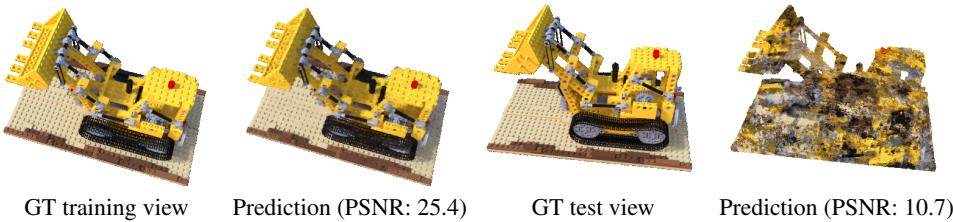


Figure 2: To demonstrate the shape-radiance ambiguity, we pretrain NeRF on a synthetic dataset where the opacity field  $\sigma$  is optimized to model an incorrect 3D shape (a unit sphere, instead of a bulldozer shape), while the radiance field  $c$  is optimized to map the training rays’ intersection with the sphere and view directions to their pixel color. In this example, we use 3 MLP layers to model the effects of view-dependence (see the MLP structure in Figure 3), and fit to 50 synthetic training images with viewpoints randomly distributed on a hemisphere. The resulting incorrect solution explains the training images very well (left two images), but fails to generalize to novel test views (right two images).

### 3 SHAPE-RADIANCE AMBIGUITY

The capacity of NeRF to model view-dependent appearance leads to an inherent ambiguity between 3D shape and radiance that can admit degenerate solutions, in the absence of regularization. For an arbitrary, incorrect shape, one can show that there exists a family of radiance fields that perfectly explains the training images, but that generalizes poorly to novel test views.

To illustrate this ambiguity, imagine that for a given scene we represent the geometry as a unit sphere. In other words, let us fix NeRF’s opacity field to be 1 at the surface of the unit sphere, and 0 elsewhere. Then, for each pixel in each training image, we intersect a ray through that pixel with the sphere, and define the radiance value at the intersection point (and along the ray direction) to be the color of that pixel. This artificially constructed solution is a valid NeRF reconstruction that perfectly fits the input images. However, this solution’s ability to synthesize novel views is very limited: accurately generating such a view would involve reconstructing an arbitrarily complex view-dependent function at each surface point. The model is very unlikely to accurately interpolate such a complex function, unless the training views are extremely dense, as in conventional light field rendering works (Buehler et al., 2001; Levoy & Hanrahan, 1996; Gortler et al., 1996). This shape-radiance ambiguity is illustrated in Figure 2.

Why does NeRF avoid such degenerate solutions? We hypothesize that two related factors come to NeRF’s rescue: 1) incorrect geometry forces the radiance field to have higher intrinsic complexity (i.e., much higher frequencies) while in contrast 2) NeRF’s specific MLP structure implicitly encodes a smooth BRDF prior on surface reflectance.

**Factor 1:** As  $\sigma$  deviates from the correct shape,  $c$  must in general become a high-frequency function with respect to  $d$  to reconstruct the input images. For the correct shape, the surface light field will generally be much smoother (in fact, constant for Lambertian materials). The higher complexity required for incorrect shapes is more difficult to represent with a limited capacity MLP.

**Factor 2:** In particular, NeRF’s specific MLP structure encodes an implicit prior favoring smooth surface reflectance functions where  $c$  is smooth with respect to  $d$  at any given surface point  $x$ . This MLP structure, shown in Figure 3, treats the scene position  $x$  and the viewing direction  $d$  asymmetrically:  $d$  is injected into the network close to the end of the MLP, meaning that there are fewer MLP parameters, as well as fewer non-linear activations, involved in the creation of view-dependent effects. In addition, the Fourier features used to encode the viewing direction consist only of low-frequency components, i.e.,  $\gamma^4(\cdot)$

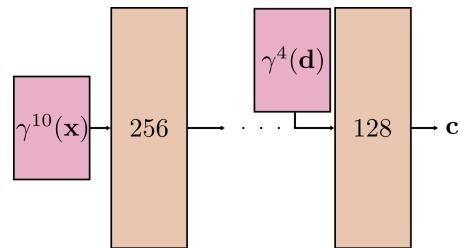


Figure 3: Structure of NeRF’s MLP for modeling radiance  $c$ .

	Scan 65			Scan 106			Scan 118		
	$\downarrow$ LPIPS	$\uparrow$ SSIM	$\uparrow$ PSNR	$\downarrow$ LPIPS	$\uparrow$ SSIM	$\uparrow$ PSNR	$\downarrow$ LPIPS	$\uparrow$ SSIM	$\uparrow$ PSNR
NeRF MLP	<b>0.0258/0.0825</b>	<b>0.988/0.967</b>	<b>33.47/28.41</b>	<b>0.0641/0.0962</b>	<b>0.978/0.959</b>	<b>35.12/30.65</b>	<b>0.0380/0.0638</b>	<b>0.987/0.969</b>	<b>37.53/31.82</b>
Vanilla MLP	0.0496/0.139	0.975/0.937	28.55/21.65	0.116/0.173	0.948/0.915	30.29/26.38	0.0609/0.094	0.979/0.952	35.31/29.58

Table 1: On DTU scenes (Jensen et al., 2014), replacing NeRF’s MLP with a vanilla MLP significantly reduces generalization to novel views. We use the same data split as Riegler & Koltun (2020). Numbers on the left are for interpolation, numbers on the right are for extrapolation. They are evaluated on the full images with background masked out.

vs.  $\gamma^{10}(\cdot)$  for encoding  $\mathbf{d}$  vs.  $\mathbf{x}$  (see Eq. 3). In other words, for a fixed  $\mathbf{x}$ , the radiance  $\mathbf{c}(\mathbf{x}, \mathbf{d})$  has limited expressivity with respect to  $\mathbf{d}$ .

To validate this hypothesis, we perform an experiment where we instead represent  $\mathbf{c}$  with a vanilla MLP that treats  $\mathbf{x}$  and  $\mathbf{d}$  symmetrically—i.e., accepting both as inputs to the first layer and encoding both with  $\gamma^{10}(\cdot)$ —to eliminate any implicit priors involving viewing direction that arise from the network structure. If we train NeRF from scratch with this alternate model for  $\mathbf{c}$ , we observe reduced test image quality compared with NeRF’s special MLP, as shown in Figure 4 and Table 1. This result is consistent with our hypothesis that implicit regularization of reflectance in NeRF’s MLP model of radiance  $\mathbf{c}$  helps recover correct solutions.

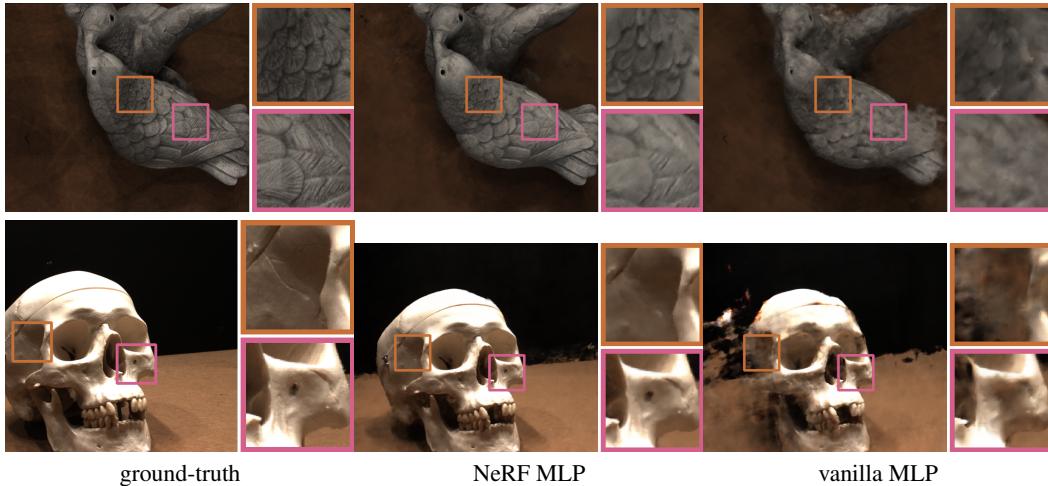
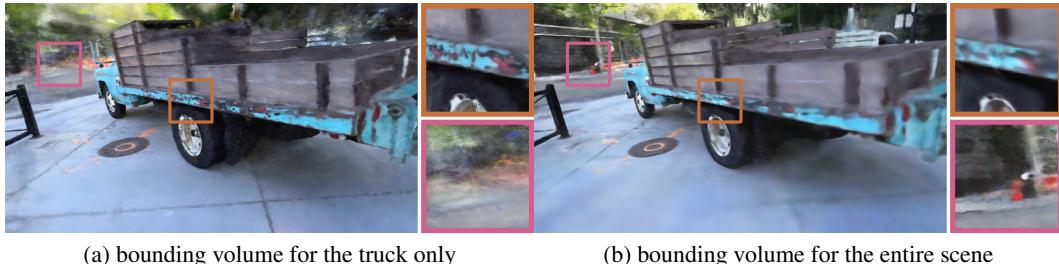


Figure 4: On DTU scenes (Jensen et al., 2014; Riegler & Koltun, 2020), this figure shows the effect of replacing NeRF’s model for the radiance field  $\mathbf{c}$  with a vanilla MLP (while keeping the structure of  $\sigma$  the same and training both fields from scratch). The vanilla MLP compromises NeRF’s ability to generalize to novel views.

#### 4 INVERTED SPHERE PARAMETRIZATION

The volumetric rendering formula in Eq. 2 integrates over Euclidean depth. When the dynamic range of the true scene depth is small, the integral can be numerically well-approximated with a finite number of samples. However, for outdoor, 360° captures centered on nearby objects that also observe the surrounding environment, the dynamic depth range can be extremely large, as the background (buildings, mountains, clouds, etc.) can be arbitrarily far away. Such a high dynamic depth range leads to severe resolution issues in NeRF’s volumetric scene representation, because to synthesize photo-realistic images, the integral in Eq. 2 needs sufficient resolution in both foreground and background areas, which is challenging to achieve by simply sampling points according to a Euclidean parameterization of 3D space. Fig. 5 demonstrates this tradeoff between scene coverage and capturing detail. In a more restricted scenario where all cameras are forward-facing towards a plane separating the cameras from the scene content, NeRF addresses this resolution issue by projectively mapping a subset of the Euclidean space, i.e., a reference camera’s view frustum, to Normalized Device Coordinates (NDC) (McReynolds & Blythe, 2005), and integrating in this NDC



(a) bounding volume for the truck only

(b) bounding volume for the entire scene

Figure 5: For 360° captures of unbounded scenes, NeRF’s parameterization of space either models only a portion of the scene, leading to significant artifacts in background elements (a), or models the full scene and suffers from an overall loss of detail due to finite sampling resolution (b).

space. However, this NDC parameterization also fundamentally limits the possible viewpoints, due to its failure to cover the space outside the reference view frustum.

We address this restriction with an inverted sphere parameterization that facilitates free view synthesis. In our representation, we first partition the scene space into two volumes, an inner unit sphere and an outer volume represented by an inverted sphere covering the complement of the inner volume (see Figure 6 for an illustration and Figure 7 for a real-world example of a scene modeled in this way). The inner volume contains the foreground and all the cameras, while the outer volume contains the remainder of the environment.

These two volumes are modelled with two separate NeRFs. To render the color for a ray, they are raycast individually, followed by a final composition. No re-parameterization is needed for the inner NeRF, as that part of the scene is nicely bounded. For the outer NeRF, we apply an *inverted sphere parametrization*.

Specifically, a 3D point  $(x, y, z)$ ,  $r = \sqrt{x^2 + y^2 + z^2} > 1$  in the outer volume can be re-parameterized by the quadruple  $(x', y', z', 1/r)$ ,  $x'^2 + y'^2 + z'^2 = 1$ , where  $(x', y', z')$  is a unit vector along the same direction as  $(x, y, z)$  representing a direction on the sphere, and  $0 < 1/r < 1$  is the inverse radius along this direction specifying the point  $r \cdot (x', y', z')$  outside the sphere. Unlike Euclidean space where objects can be at unlimited distance from the origin, all the numbers in the re-parametrized quadruple are bounded, i.e.,  $x', y', z' \in [-1, 1]$ ,  $1/r \in [0, 1]$ . This not only improves numeric stability, but also respects the fact that farther objects should get less resolution. We can directly raycast this 4D bounded volume (only 3 degrees of freedom) to render a camera ray’s color.

$$B = \{(x, y, z) : \sqrt{x^2 + y^2 + z^2} = 1\}$$

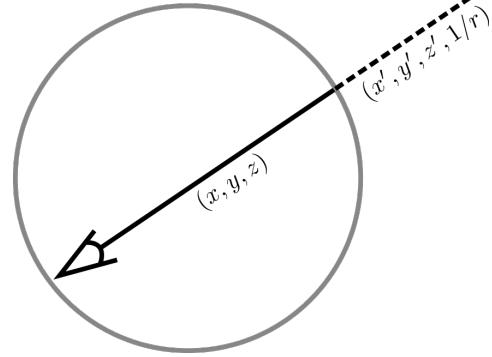


Figure 6: NeRF++ applies different parameterizations for scene contents inside and outside the unit sphere.



(a) NeRF++ prediction

(b) predicted foreground

(c) predicted background

Figure 7: NeRF++ separates the modelling of foreground and background. The image synthesized by NeRF++ (a) is a composite of the predicted foreground (b) and background (c).

Note that the composite of foreground and background is equivalent to breaking the integral in Eq. 2 into two parts, integration inside the inner and outer volumes. In particular, consider that the ray  $\mathbf{r} = \mathbf{o} + t\mathbf{d}$  is partitioned into two segments by the unit sphere: in the first,  $t \in (0, t')$  is inside the sphere; in the second,  $t \in (t', \infty)$  is outside the sphere. We can rewrite the volumetric rendering integral in Eq. 2 as

$$\begin{aligned} \mathbf{C}(\mathbf{r}) &= \underbrace{\int_{t=0}^{t'} \sigma(\mathbf{o} + t\mathbf{d}) \cdot \mathbf{c}(\mathbf{o} + t\mathbf{d}, \mathbf{d}) \cdot e^{-\int_{s=0}^t \sigma(\mathbf{o} + s\mathbf{d}) ds} dt}_{(i)} \\ &\quad + \underbrace{e^{-\int_{s=0}^{t'} \sigma(\mathbf{o} + s\mathbf{d}) ds}}_{(ii)} \cdot \underbrace{\int_{t=t'}^{\infty} \sigma(\mathbf{o} + t\mathbf{d}) \cdot \mathbf{c}(\mathbf{o} + t\mathbf{d}, \mathbf{d}) \cdot e^{-\int_{s=t'}^t \sigma(\mathbf{o} + s\mathbf{d}) ds} dt}_{(iii)}. \end{aligned} \quad (4)$$

Terms (i) and (ii) are computed in Euclidean space, while term (iii) is computed in inverted sphere space with  $\frac{1}{r}$  as the integration variable. In other words, we use  $\sigma_{in}(\mathbf{o} + t\mathbf{d})$ ,  $\mathbf{c}_{in}(\mathbf{o} + t\mathbf{d}, \mathbf{d})$  in terms (i) and (ii), and  $\sigma_{out}(x', y', z', 1/r)$ ,  $\mathbf{c}_{out}(x', y', z', 1/r, \mathbf{d})$  in term (iii).

In order to compute term (iii) for the ray  $\mathbf{r} = \mathbf{o} + t\mathbf{d}$ , we first need to be able to evaluate  $\sigma_{out}, \mathbf{c}_{out}$  at any  $1/r$ ; in other words, we need a way to compute  $(x', y', z')$  corresponding to a given  $1/r$ , so that  $\sigma_{out}, \mathbf{c}_{out}$  can take  $(x', y', z', 1/r)$  as input. This can be achieved as follows.

As shown in Fig. 8, let the ray intersect the unit sphere at point  $\mathbf{a}$ , and the midpoint of the chord aligning with the ray be the point  $\mathbf{b}$ . Point  $\mathbf{a} = \mathbf{o} + t_a \mathbf{d}$  is computed by solving  $|\mathbf{o} + t_a \mathbf{d}| = 1$ , while  $\mathbf{b} = \mathbf{o} + t_b \mathbf{d}$  is acquired by solving  $\mathbf{d}^T (\mathbf{o} + t_b \mathbf{d}) = 0$ . Then to get  $(x', y', z')$  given  $1/r$ , we can rotate the vector  $\mathbf{a}$  along the vector  $\mathbf{b} \times \mathbf{d}$  by the angle  $\omega = \arcsin |\mathbf{b}| - \arcsin(|\mathbf{b}| \cdot \frac{1}{r})$ . Once we can evaluate  $\sigma_{out}, \mathbf{c}_{out}$  at any  $1/r$ , we simply sample a finite number of points from the interval  $[0, 1]$  to compute term (iii).

The inverted sphere parameterization for the outer volume has an intuitive physical explanation. It can be viewed in terms of a virtual camera whose image plane is the unit sphere at the scene origin. Hence, the 3D point  $(x, y, z)$  is projected to the pixel  $(x', y', z')$  on the image plane, while the term  $1/r \in (0, 1)$  serves as the (inverse) depth, or disparity, of the point. From this perspective, the NDC parameterization suitable only for forward-facing capture is related to our representation, as it uses a virtual pinhole camera rather than a spherical projection surface. In this sense, our inverted sphere parameterization is related to the concept of *multi-sphere images* (a scene representation consisting of nested concentric spheres sampled according to inverse depth from the sphere centers) proposed in recent work on view synthesis (Attal et al., 2020; Broxton et al., 2020).

## 5 EXPERIMENTS

We validate NeRF++ and compare it with NeRF on two real-world datasets captured with hand-held cameras: Tanks and Temples (T&T) (Knapitsch et al., 2017) and the Light Field (LF) dataset of Yücer et al. (2016). We report PSNR, SSIM, and LPIPS (Zhang et al., 2018) as our quantitative metrics for measuring the quality of synthesized test images.

**T&T dataset.** We use the training/testing images and SfM poses provided by Riegler & Koltun (2020). This dataset consists of hand-held 360° captures of four large-scale scenes; the camera poses are estimated by COLMAP SfM (Schönberger & Frahm, 2016). For NeRF, we normalize the scene such that all cameras are inside the sphere of radius  $\frac{1}{8}$ . This normalization ensures that the unit sphere covers the majority of the background scene content (although some background geometry still lies outside the bounding unit sphere). To numerically compute the volumetric rendering integral for each camera ray, we uniformly sample points from the ray origin to its intersection with the unit sphere.

	Truck			Train			M60			Playground		
	$\downarrow$ LPIPS	$\uparrow$ SSIM	$\uparrow$ PSNR	$\downarrow$ LPIPS	$\uparrow$ SSIM	$\uparrow$ PSNR	$\downarrow$ LPIPS	$\uparrow$ SSIM	$\uparrow$ PSNR	$\downarrow$ LPIPS	$\uparrow$ SSIM	$\uparrow$ PSNR
NeRF	0.513	0.747	20.85	0.651	0.635	16.64	0.602	0.702	16.86	0.529	0.765	21.55
NeRF++	<b>0.298</b>	<b>0.823</b>	<b>22.77</b>	<b>0.523</b>	<b>0.672</b>	<b>17.17</b>	<b>0.435</b>	<b>0.738</b>	<b>17.88</b>	<b>0.391</b>	<b>0.799</b>	<b>22.37</b>
	Africa			Basket			Torch			Ship		
NeRF	0.217	0.894	26.16	0.377	0.805	20.83	0.347	0.811	22.81	0.372	0.801	23.24
NeRF++	<b>0.163</b>	<b>0.923</b>	<b>27.41</b>	<b>0.254</b>	<b>0.884</b>	<b>21.84</b>	<b>0.226</b>	<b>0.867</b>	<b>24.68</b>	<b>0.241</b>	<b>0.867</b>	<b>25.35</b>

Table 2: We compare NeRF++ with NeRF on four Tanks and Temples scenes: Truck, Train, M60, and Playground, as well as four LF scenes: Africa, Torch, Ship, and Basket. NeRF++ consistently outperforms NeRF in all metrics.

Since NeRF++ ray-casts both inner and outer volumes and hence uses twice the number of samples per camera ray compared to a single volume, we also double the number of samples used by NeRF for the sake of fairness. Specifically, NeRF’s coarse-level MLP uses 128 uniform samples, while the fine-level MLP uses 256 additional importance samples. Under this hyper-parameter setting, NeRF has roughly the same computation cost and GPU memory footprint during training and testing as NeRF++. We randomly sample 2048 camera rays at each training iteration, and train NeRF and NeRF++ for 250k iterations with a learning rate of 5e-4.

**LF dataset.** We use four scenes from the LF dataset: Africa, Basket, Ship, and Torch. Each scene is densely covered by 2k–4k hand-held captured images, and camera parameters are recovered using SfM. We construct a sparse surrounding capture by temporally subsampling the images by a factor of 40. In particular, the training images are frames 0, 40, 80, ..., and the testing images are frames 20, 60, 100, ... . The scene normalization method, number of samples per camera ray, batch size, training iterations, and learning rate are the same as in the T&T dataset.

**Results.** As shown in Table 2, NeRF++ significantly outperforms NeRF in challenging scenarios involving 360° captures of objects within large-scale unbounded scenes. In Figure 9, we can see that images synthesized by NeRF++ have significantly higher fidelity.

## 6 OPEN CHALLENGES

NeRF++ improves the parameterization of unbounded scenes in which both the foreground and the background need to be faithfully represented for photorealism. However, there remain a number of open challenges. First, the training and testing of NeRF and NeRF++ on a single large-scale scene is quite time-consuming and memory-intensive. Training NeRF++ on a node with 4 RTX 2080 Ti GPUs takes ~24 hours. Rendering a single 1280x720 image on one such GPU takes ~30 seconds at test time. Liu et al. (2020) have sped up the inference, but rendering is still far from real-time. Second, small camera calibration errors may impede photorealistic synthesis. Robust loss functions, such as the contextual loss (Mechrez et al., 2018), could be applied. Third, photometric effects such as auto-exposure and vignetting can also be taken into account to increase image fidelity. This line of investigation is related to the lighting changes addressed in the orthogonal work of Martin-Brualla et al. (2020).

## REFERENCES

- Benjamin Attal, Selena Ling, Aaron Gokaslan, Christian Richardt, and James Tompkin. MatryOD-Shka: Real-time 6DoF video view synthesis using multi-sphere images. In *ECCV*, 2020.
- Michael Broxton, John Flynn, Ryan Overbeck, Daniel Erickson, Peter Hedman, Matthew DuVall, Jason Dourgarian, Jay Busch, Matt Whalen, and Paul Debevec. Immersive light field video with a layered mesh representation. *ACM Transactions on Graphics*, 39(4), 2020.
- Chris Buehler, Michael Bosse, Leonard McMillan, Steven Gortler, and Michael Cohen. Unstructured lumigraph rendering. In *SIGGRAPH*, 2001.
- Shenchang Eric Chen and Lance Williams. View Interpolation for Image Synthesis. In *SIGGRAPH*, 1993.



Figure 9: We qualitatively compare NeRF++ with NeRF on two T&T scenes (Truck, Playgroud) and two LF scenes (Africa, Torch). NeRF++ produces sharper images than NeRF and is able to better represent both the foreground and the background.

Paul E. Debevec, Camillo J. Taylor, and Jitendra Malik. Modeling and rendering architecture from photographs: A hybrid geometry-and image-based approach. In *SIGGRAPH*, 1996.

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

Rasmus Jensen, Anders Dahl, George Vogiatzis, Engil Tola, and Henrik Aanæs. Large scale multi-view stereopsis evaluation. In *CVPR*, 2014.

James T Kajiya and Brian P Von Herzen. Ray tracing volume densities. In *SIGGRAPH*, 1984.

Arno Knapitsch, Jaesik Park, Qian-Yi Zhou, and Vladlen Koltun. Tanks and temples: Benchmarking large-scale scene reconstruction. *ACM Transactions on Graphics*, 36(4), 2017.

Marc Levoy and Pat Hanrahan. Light field rendering. In *SIGGRAPH*, 1996.

Lingjie Liu, Jiatao Gu, Kyaw Zaw Lin, Tat-Seng Chua, and Christian Theobalt. Neural sparse voxel fields. *arXiv:2007.11571*, 2020.

Ricardo Martin-Brualla, Noha Radwan, Mehdi SM Sajjadi, Jonathan T Barron, Alexey Dosovitskiy, and Daniel Duckworth. NeRF in the wild: Neural radiance fields for unconstrained photo collections. *arXiv:2008.02268*, 2020.

- 
- Tom McReynolds and David Blythe. *Advanced graphics programming using OpenGL*. Elsevier, 2005.
- Roey Mechrez, Itamar Talmi, and Lihi Zelnik-Manor. The contextual loss for image transformation with non-aligned data. In *ECCV*, 2018.
- 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. In *ECCV*, 2020.
- Gernot Riegler and Vladlen Koltun. Free view synthesis. In *ECCV*, 2020.
- Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *CVPR*, 2016.
- Harry Shum and Sing Bing Kang. Review of image-based rendering techniques. In *Visual Communications and Image Processing*, 2000.
- Matthew Tancik, Pratul P Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan T Barron, and Ren Ng. Fourier features let networks learn high frequency functions in low dimensional domains. *arXiv:2006.10739*, 2020.
- Daniel N Wood, Daniel I Azuma, Ken Aldinger, Brian Curless, Tom Duchamp, David H Salesin, and Werner Stuetzle. Surface light fields for 3D photography. In *SIGGRAPH*, 2000.
- Kaan Yücer, Alexander Sorkine-Hornung, Oliver Wang, and Olga Sorkine-Hornung. Efficient 3D object segmentation from densely sampled light fields with applications to 3D reconstruction. *ACM Transactions on Graphics*, 35(3), 2016.
- Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018.