View Synthesis In Casually Captured Scenes Using a Cylindrical Neural Radiance Field With Exposure Compensation

Wesley Khademi
California Polytechnic State University
San Luis Obispo, California, USA
wkhademi@calpoly.edu

Jonathan Ventura
California Polytechnic State University
San Luis Obispo, California, USA
jventu09@calpoly.edu

ABSTRACT

We extend Neural Radiance Fields (NeRF) with a cylindrical parameterization that enables rendering photorealistic novel views of 360° outward facing scenes. We further introduce a learned exposure compensation parameter to account for the varying exposure in training images that may occur from casually capturing a scene. We evaluate our method on a variety of 360° casually captured scenes.

ACM Reference Format:
Wesley Khademi and Jonathan Ventura. 2021. View Synthesis In Casually Captured Scenes Using a Cylindrical Neural Radiance Field With Exposure Compensation. In *Special Interest Group on Computer Graphics and Interactive Techniques Conference Posters (SIGGRAPH '21 Posters), August 09-13, 2021.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3450618.3469147

1 INTRODUCTION

Neural Radiance Fields (NeRF) [Mildenhall et al. 2020] have emerged as a promising new approach to view synthesis, producing high-quality photorealistic novel views on bounded inward facing scenes as well as forward facing scenes. However, NeRF struggles to learn to model appearance and geometry when dealing with unbounded 360° scenes or images with varying exposure, which are common issues faced in casual capture, where a person tries to capture imagery of an entire scene by spinning a smartphone in a circle. To this end, we work towards achieving 6-DOF view synthesis of 360° casually captured scenes using NeRF. We first introduce a cylindrical parameterization, which resolves NeRF's inability to learn to represent the scene geometry of large 360° unbounded scenes. We then introduce an exposure compensation technique that aids in reducing artifacts and maintaining consistent exposure across views when training on images of varying exposure.

2 NEURAL RADIANCE FIELDS

We first provide a brief introduction to Neural Radiance Fields. NeRF learns a continuous representation of a scene implicitly as a multilayer perceptron (MLP) that maps a 5D input, 3D position $\mathbf{x} = (x, y, z)$ and 2D viewing direction $\mathbf{d} = (\theta, \phi)$, to a color $\mathbf{c} \in \mathbb{R}^3$ and density $\sigma \in \mathbb{R}$:

$$\sigma(\mathbf{x}), \mathbf{c}(\mathbf{x}, \mathbf{d}) = MLP_\theta(\mathbf{x}, \mathbf{d}) \tag{1}$$

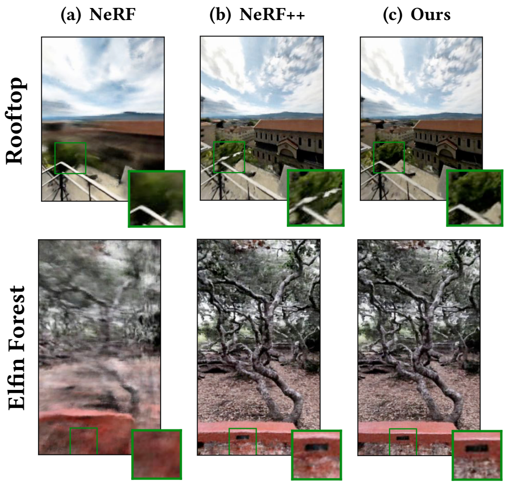

Figure 1: Visual comparison of NeRF, NeRF++, and our Cylindrical NeRF with Exposure Compensation on the casually captured Rooftop and Elfin Forest datasets.

Note that the density σ is only a function of 3D position, which enforces a coherent scene structure across multiple views. On the other hand, the emitted color $\mathbf{c}$ is a function of 3D position and 2D viewing direction, allowing for view-dependent color.

To render a pixel's color, NeRF shoots a ray $\mathbf{r}(t) = \mathbf{o} + t\mathbf{d}$ from the camera origin $\mathbf{o}$ through the center of a pixel out into the scene, queries the MLP at points along $\mathbf{r}$, and then uses numerical quadrature to approximate the volume rendering integral [Max 1995]. The expected pixel color $\widehat{\mathbf{C}}(\mathbf{r})$ is defined as:

$$\widehat{\mathbf{C}}(\mathbf{r}) = \sum_{i=1}^{N} T_i(1 - exp(-\sigma_i\delta_i))\mathbf{c}_i, \tag{2}$$

$$T_i = exp(-\sum_{j=1}^{i-1} \sigma_j\delta_j), \tag{3}$$

where $\mathbf{c}_i$ and σ_i are the color and density at point $\mathbf{r}(t_i)$ and $\delta_i = t_{i+1} - t_i$ is the distance between adjacent samples.

NeRF jointly trains two MLPs with different sampling strategies to improve sampling efficiency. First, the "coarse" model uses stratified sampling to sample points between the near plane t_n and far plane t_f. For a more informed sampling, the "fine" model then uses the output of the "coarse" network to produce a larger number of samples to occur in regions of visible content.

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
SIGGRAPH '21 Posters, August 09-13, 2021, Virtual Event, USA
© 2021 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-8371-4/21/08.
https://doi.org/10.1145/3450618.3469147

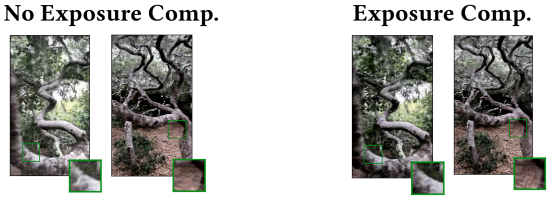

Figure 2: Our exposure compensation method produces sharper images and maintains a more consistent exposure across views.

NeRF optimizes the MLPs by minimizing the mean squared error between a set of ground truth pixels from observed images and the predicted pixels output from the rendering described in Equation 2.

3 OUR APPROACH

While NeRF has shown impressive ability to encode appearance and scene geometry, it is limited to small bounded scenes due to its choice to represent points in Euclidean space. When dealing with 360° outward facing scenes, the range of scene coordinates become too large for NeRF's MLPs to encode, making it difficult for the network to learn to properly reconstruct the scene.

To overcome this, we introduce a new cylindrical parameterization that resembles the inverted sphere parameterization presented in [Zhang et al. 2020] but is more suitable to casual capture where the user spins in a circle and thus the top and bottom of the scene are unlikely to be observed. Instead of performing stratified sampling for points t_i along a ray, we sample cylinders of varying radii:

$$\frac{1}{r_i} \sim \mathcal{U}\left[\frac{1}{t_f} + \frac{i-1}{N}\left(\frac{1}{t_n} - \frac{1}{t_f}\right), \frac{1}{t_f} + \frac{i}{N}\left(\frac{1}{t_n} - \frac{1}{t_f}\right)\right], \quad (4)$$

where t_n is the near radius, t_f is the far radius, and r_i is the radius of a cylinder centered at world origin. Sampling in inverse radius bounds samples such that $1/r \in [0, 1]$ for all $r > 1$, irrespective of scene depth range.

Given a ray $\mathbf{r}(t) = \mathbf{o} + t\mathbf{d}$ with origin $\mathbf{o}$ and direction $\mathbf{d}$, we solve for t_i by finding where along the ray it intersects with a cylinder of radius r_i centered at the world origin. This gives a constraint

$$(o_x + t_i d_x)^2 + (o_z + t_i d_z)^2 = r_i^2 \quad (5)$$

which is easily solved. The solution provides two values for t_i, but we are only concerned with the positive t_i value that corresponds to a point along the ray that lies in front of the camera.

To obtain our 3D point along the ray that intersects a cylinder of radius r_i, we compute $\mathbf{x} = (x, y, z) = \mathbf{r}(t_i) = \mathbf{o} + t_i\mathbf{d}$. We follow the work of [Zhang et al. 2020] and reparameterize the 3D point $\mathbf{x}$ as a 4D point $(x', y', z', 1/r_i)$ where (x', y', z') is the point $\mathbf{x}$ projected onto the unit cylinder. While our method supports view-direction dependence, we opted in our experiments to leave it out for all NeRF models since our scenes are mostly diffuse.

Images taken from a casually captured scene may suffer from slight variations in appearance, such as in exposure, which can lead NeRF to produce renderings that contain severe artifacts or have inconsistent exposure across views. To handle these appearance differences, we further propose a way for NeRF to learn to account for varying exposure across the training images $\{\mathcal{I}_i\}_{i=0}^N$. We introduce a learned brightness vector $\mathbf{b} \in \mathbb{R}^N$ in which each brightness parameter b_i corresponds to an exposure adjustment in training image $\mathcal{I}_i$. We select image $\mathcal{I}_0$ to be the desired exposure we want each training image to match and thus fix $b_0 = 0$. We then learn the appropriate exposure adjustments for the other $n - 1$ images by simultaneously optimizing the brightness parameters $\{b_i\}_{i=1}^N$ along with the weights of NeRF's MLPs. To do so, we minimize a modified version of NeRF's loss:

$$\sum_{ij} ||(\mathbf{C}(\mathbf{r}_{ij}) + b_j) - \widehat{\mathbf{C}}^c(\mathbf{r}_{ij})||_2^2 + ||(\mathbf{C}(\mathbf{r}_{ij}) + b_j) - \widehat{\mathbf{C}}^f(\mathbf{r}_{ij})||_2^2 \quad (6)$$

where $\mathbf{r}_{ij}$ represents the ray that intersects pixel i in image $\mathcal{I}_j$ and b_j is the brightness adjustment for training image $\mathcal{I}_j$.

The advantage to our method is that the learned brightness vector is only required during training, avoiding the need to find an optimal exposure adjustment during test time.

4 DISCUSSION

For a fair comparison, we train all models for 500k iterations, use 384 samples per ray (128 coarse + 256 fine), and downscale images by $4x$. Since all the camera poses lie near the unit cylinder, we don't observe the scene inside the unit cylinder, and thus we also opt to remove NeRF++'s inner volume to avoid artifacts arising from it.

Figure 1 compares novel views of casually captured scenes using the regular NeRF parameterization and our cylindrical parameterization. While NeRF struggles to learn both scenes, our method is able to faithfully reconstruct the scene geometry for both near and far objects. By bounding 3D points to the surface of the unit cylinder and $1/r \in [0, 1]$, our cylindrical parameterization makes it easier for NeRF to learn the radiance field and density of scenes compared to NeRF's original unbounded parameterization of points.

Compared to NeRF++, our cylindrical parameterization leads to results that are of similar quality, but our exposure compensation method helps reduce floating artifacts in the scene. Figure 1 shows that NeRF++ suffers from artifacting near the railing in the Rooftop scene and the bench in the Elfin Forest scene, while our method is able to resolve these artifacts.

Our exposure compensation technique also helps reduce the variation in exposure across novel views. Figure 2 shows that our Cylindrical NeRF produces more consistent exposures between views when using our exposure compensation than without it.

While our method provides a way to extend NeRF to handle casually captured 360° scenes, there are still some limitations to overcome before NeRF is able to achieve high-quality 6-DoF view synthesis from hand-held video. Mainly, our method still suffers from artifacts and degradation in scene geometry when trying to extrapolate unseen parts of a scene or when rendering views that are significantly far away from the training image viewpoints.

REFERENCES

N. Max. 1995. Optical models for direct volume rendering. *IEEE Transactions on Visualization and Computer Graphics* 1, 2 (1995), 99–108. https://doi.org/10.1109/2945.468400

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. In *European Conference on Computer Vision*. Springer, 405–421.

Kai Zhang, Gernot Riegler, Noah Snavely, and Vladlen Koltun. 2020. NeRF++: Analyzing and Improving Neural Radiance Fields. arXiv:2010.07492 [cs.CV]