Object Reconstruction and Novel View Synthesis Based on NeRF

Xiang Zheng
21307110169@m.fudan.edu.cn

YunNam Chan
21340980053@m.fudan.edu.cn

Abstract

This study investigates the implementation and performance of Neural Radiance Fields (NeRF) for 3D object reconstruction and novel view synthesis from 2D images. Using high-resolution datasets of a tiger model and a vase deck, camera parameters and poses were estimated through COLMAP and Local Light Field Fusion (LLFF). The NeRF model was trained, achieving a training loss of 0.001897 and a PSNR of 30.883136 for the tiger model, and a training loss of 0.005910 and a PSNR of 26.750474 for the vase deck model. Despite the higher PSNR for the tiger model, the vase deck model demonstrated superior qualitative performance in validation images and reconstructed videos due to higher image quality and consistency. This study underscores the importance of high-quality datasets in achieving effective 3D reconstructions with NeRF, and NeRF's ability to render high-quality 3D videos in a short time.

Note: You can find the project files in the GitHub repo. For instructions on how to train and test the neural network, please refer to the readme file.

1 Introduction

Neural Radiance Fields (NeRF) have emerged as a significant advancement in the realm of high-fidelity 3D object reconstruction and novel view synthesis from 2D images. By leveraging deep learning, NeRF models are capable of learning a continuous volumetric scene function, facilitating the synthesis of new views of complex scenes with remarkable detail and accuracy. This report documents our experimental study on reconstructing objects and generating novel views using NeRF.

Our process commenced with the acquisition of multi-angle images of a tiger model, ensuring comprehensive coverage from various perspectives. Given that the self-captured images were of low quality, we incorporated additional images from the official NeRF dataset for comparison. Camera parameters were estimated using COLMAP, an open-source Structure-from-Motion (SfM) and Multi-View Stereo (MVS) software. These estimated parameters were subsequently utilized to train the NeRF model.

The trained NeRF model was employed to create videos simulating a camera orbiting the objects, thereby providing a dynamic view of the reconstructed scene. The results were evaluated using the Peak Signal-to-Noise Ratio (PSNR) metric to measure reconstruction quality against test images.

The report is organized as follows: Section 2 details the dataset, preprocessing steps, and train-validation split. Section 3 delineates the model architecture. Section 4 presents the experimental setup and results. Finally, Section 5 discusses the findings and concludes the study, offering insights into potential future work.

2 Dataset

The dataset utilized in this study comprises 43 high-resolution images, each with dimensions of 1706×1279 pixels, captured around a tiger model. To address the low quality of these images, we selected 40 high-resolution images from the vasedeck dataset (4032×3024) available in the official NeRF dataset provided by the authors of the NeRF paper. This combination provides a comprehensive set of images suitable for evaluating the algorithms employed in this research. Figure 1 offers an overview of the two objects.

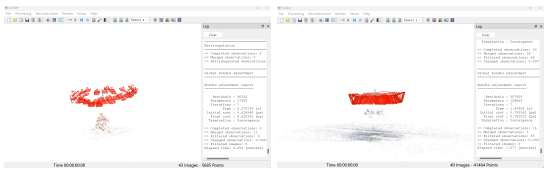
2.1 Preprocessing

The preprocessing phase encompasses several critical steps to prepare the dataset for further analysis. Initially, the images are processed using COLMAP, a robust photogrammetry software. COLMAP is employed to determine camera parameters, perform feature extraction, feature matching, and re-

(a) Tiger Model (b) Vase Deck

Figure 1: An overview of the original objects

construct the 3D structure from the images. The output of COLMAP following the reconstruction process is illustrated in Figure 2. Subsequently, the Local Light Field Fusion (LLFF) algorithm is applied to estimate the poses of the image set and generate the `poses_bounds.npy` file. This file encapsulates the essential pose information necessary for subsequent stages of the analysis.



(a) Tiger Model (b) Vase Deck

Figure 2: A snapshot of the COLMAP interface post-reconstruction

2.2 Training-Validation Split

To evaluate the performance of the model, a hold-out validation method is employed. Specifically, one image out of every eight is selected as a test (validation) image. This systematic sampling results in the following indices being designated as test (validation) images: [0, 8, 16, 24, 32, 40(only in tiger model)]. This approach ensures a representative distribution of the dataset for both training and validation purposes.

3 Model Architecture

NeRF (Neural Radiance Fields) utilizes a fully-connected deep neural network to represent and render a 3D scene. The architecture and its components are detailed below, with the pipeline of NeRF visualized in Figure 3.

3.1 Network Structure

The core of NeRF is a multi-layer perceptron (MLP) that takes a 5D coordinate as input and outputs the volume density and RGB color. Specifically:

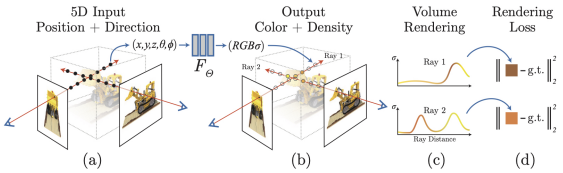


Figure 3: Pipeline of NeRF

- **Input**: A 3D position $\mathbf{x} = (x, y, z)$ and a 2D viewing direction $\mathbf{d} = (\theta, \phi)$.
- **Hidden Layers**: The network has 8 fully-connected layers with 256 channels each, using the ReLU activation function.
- **Output**: The network outputs the volume density σ and the RGB color $\mathbf{c} = (r, g, b)$.

3.2 Positional Encoding

To capture high-frequency details, NeRF applies positional encoding to the input coordinates. This encoding maps each input coordinate into a higher-dimensional space using sinusoidal functions:

$$\gamma(p) = (\sin(2^0\pi p), \cos(2^0\pi p), \ldots, \sin(2^{L-1}\pi p), \cos(2^{L-1}\pi p)),$$

where L is typically set to 10 for spatial coordinates and 4 for viewing direction components.

3.3 Hierarchical Volume Sampling

NeRF employs a hierarchical sampling strategy to efficiently sample points along camera rays, involving two networks:

- **Coarse Network**: Samples points along the rays to provide a rough approximation.
- **Fine Network**: Samples additional points based on the coarse network's output for a more detailed representation.

3.4 Network Details

The MLP processes the positional encoding of the input location through 8 layers. A skip connection is used to concatenate the input to the fifth layer's activation. The architecture details are as follows:

- **Layer Configuration**: 8 fully-connected ReLU layers with 256 channels.
- **Skip Connection**: Input is concatenated with the output of the fifth layer.

- **Output Processing**: A final layer outputs the volume density σ and a 256-dimensional feature vector. This feature vector, combined with the positional encoding of the viewing direction, produces the RGB color $\mathbf{c}$.

3.5 Training and Rendering

For training, the scene is normalized to fit within a unit cube. During rendering, NeRF samples 256 points per ray (64 for the coarse network and 192 for the fine network). Random Gaussian noise is added during optimization to improve visual performance.

4 Train & Result

4.1 Configuration

The configuration settings for the Neural Radiance Fields (NeRF) model utilized in our experiments are detailed in Table 1. These settings closely adhere to the default configurations proposed by the original authors of the NeRF paper, with slight modifications to suit our specific experimental setup.

4.2 Training Loss and PSNR

The performance of the NeRF model during the training process is evaluated using two key metrics: training loss and Peak Signal-to-Noise Ratio (PSNR). Figures 4 and 5 illustrate the training loss and the corresponding PSNR values over the course of the training.

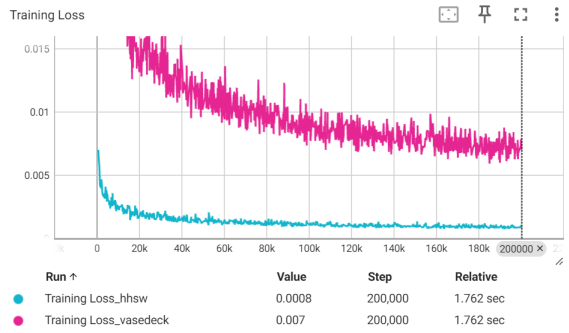


Figure 4: Training loss during the training of NeRF.

As depicted in Figure 4, the training loss exhibits a consistent decreasing trend, indicating effective learning and minimization of error over time. The training loss of the tiger model reaches a minimum value of 0.001897, whereas the training loss of the vase deck model stabilizes around 0.07. The trend in the loss curve suggests that the learning rate for the tiger model may be too high, causing it to converge prematurely, while the learning rate for the vase deck model appears to be appropriate. The continued decrease in the vase deck's training loss until the end of the training indicates that additional iterations could further improve model convergence.

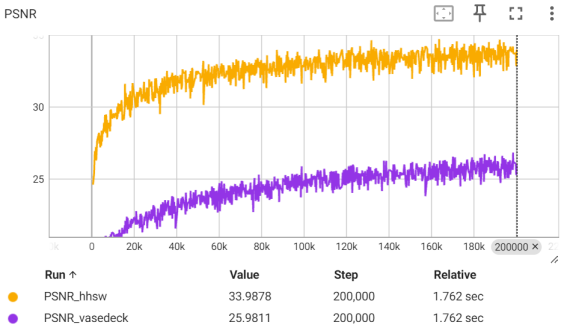


Figure 5: PSNR during the training of NeRF.

Figure 5 shows the PSNR values calculated at various stages of the training process. Generally, a higher PSNR value signifies better image reconstruction quality. The tiger model achieves its highest PSNR value of 30.883136, while the vase deck model reaches a PSNR value of 26.750474, reflecting high-quality image reconstruction for both models.

However, despite the quantitative PSNR values, the qualitative assessment of the reconstruction videos suggests that the vase deck model performs better than the tiger model. This discrepancy between the PSNR values and the visual quality of the videos will be further discussed in the subsequent comparison subsection.

4.3 Validation Images

Figures 6 and 7 display the validation images at different stages of the training process, specifically at iterations 50,000, 100,000, 150,000, and 200,000, for the tiger model and vase deck, respectively.

These images reveal a noticeable improvement in the quality of the rendered outputs over time. Initially, the images are somewhat blurred and lack detail. As training progresses, the images become progressively clearer, demonstrating enhanced detail and reduced blurriness. This improvement indicates effective learning and accurate scene reconstruction as the training advances.

Variable Name	Meaning	Value
dataset_type	Dataset type	llff
factor	Downsample factor for LLFF images	8
llffhold	Holdout test set	8
N_rand	Batch size, number of random rays per gradient step	1024
N_samples	Number of coarse samples per ray	64
N_importance	Number of additional fine samples per ray	64
raw_noise_std	Standard deviation of noise added to regularize sigma output	1
lrate	Learning rate	0.0005
optimizer	Optimizer used	Adam
loss_function	Loss function used	MSE
evaluation_metric	Evaluation metric	PSNR

Table 1: Configuration settings for the NeRF model used in our experiments.

(a) Iteration 50000 (b) Iteration 100000

(c) Iteration 150000 (d) Iteration 200000

Figure 6: Rendered validation images at different training iterations for the tiger model.

(a) Iteration 50000 (b) Iteration 100000

(c) Iteration 150000 (d) Iteration 200000

Figure 7: Rendered validation images at different training iterations for the vase deck model.

4.4 Comparison

Although the training PSNR is higher for the tiger model compared to the vase deck model, the quality of the validation images and the reconstructed videos for the tiger model is noticeably inferior. This discrepancy can be attributed to several factors:

- **Image Quality**: The tiger model images have a lower resolution, which can result in unclear rays and inadequate training data. The lower resolution of the tiger images, approximately one quarter of that of the vase deck images, significantly impacts the model's ability to learn fine details and accurate reconstructions.
- **Camera Angles**: The angles at which the tiger model images were captured may be suboptimal, leading to insufficient coverage and missing perspectives that are crucial for the NeRF model to accurately reconstruct the 3D scene. Poor camera positioning can result in incomplete data for the model to learn from.
- **Lighting Variability**: The tiger model images exhibit significant variations in lighting, which can introduce inconsistencies in the training data. Such variations can make it challenging for the NeRF model to learn a consistent volumetric scene function, leading to poorer reconstruction quality.
- **Camera Quality**: The camera used to capture the tiger model images may be of lower quality compared to the one used for the vase deck images. Poor camera quality can introduce noise and artifacts, further degrading the im-

age quality and negatively affecting the training process.

These factors collectively explain why, despite the higher PSNR values during training, the tiger model exhibits lower visual quality in validation images and reconstructed videos. The vase deck images, with their higher resolution and better camera quality, provide a more robust dataset for NeRF, resulting in superior qualitative performance despite lower PSNR values during training.

This analysis underscores the importance of high-quality, high-resolution images and comprehensive coverage in achieving effective 3D reconstructions with NeRF. Additionally, it highlights the limitations of the NeRF model when dealing with lower-quality inputs and complex textures, emphasizing the need for improved preprocessing and data acquisition techniques to enhance model performance.

5 Discussion & Conclusion

In this study, we implemented and evaluated the Neural Radiance Fields (NeRF) model using datasets comprising high-resolution images of a tiger model and a vase deck. The configuration settings closely followed the default parameters from the original NeRF paper, with minor adjustments for our specific datasets.

The preprocessing phase, involving COLMAP and Local Light Field Fusion (LLFF), was crucial for accurately determining the camera parameters and poses. COLMAP facilitated feature extraction, matching, and 3D reconstruction, while LLFF provided refined pose estimations, resulting in well-prepared datasets for NeRF training.

During training, the NeRF model showed significant improvements in accuracy over time. For the tiger model, the training loss decreased to a minimum value of 0.001897, and the PSNR peaked at 30.883136. For the vase deck model, the training loss stabilized around 0.07, and the PSNR reached 26.750474. Despite the higher PSNR for the tiger model, the vase deck model produced better quality validation images and reconstructed videos. This discrepancy is likely due to the lower resolution, suboptimal shooting angles, and significant lighting variations in the tiger model images.

Rendered validation images at different training iterations demonstrated that the model effectively improved the clarity and detail of reconstructions over time, especially for the vase deck model. The higher quality and consistency of the vase deck images contributed to better performance.

In conclusion, this study demonstrates the NeRF model's capability to produce high-quality 3D reconstructions from 2D images, given high-quality input data. The results highlight the importance of meticulous dataset preparation and configuration. Future work could explore diverse datasets, alternative preprocessing techniques, and optimized training processes to further enhance model performance and reduce computational costs.