Enhancing Neural Radiance Fields with Depth Supervision: An Evaluation of Techniques for Few-Shot Novel View Synthesis

Tuxun Lu
Department of Computer Science
University of Maryland
College Park, MD, United States, 20742
tuxunlu@umd.edu

Abstract

Neural Radiance Fields (NeRFs) have showcased remarkable capabilities in vision and graphics applications, including novel view synthesis. However, preventing NeRFs from degrading when limited input views are available remains a challenge. A branch of recent work resorted to incorporating depth information as a prior in the training of NeRF to alleviate this issue. This report explores the effectiveness of various techniques of depth supervision.

1 Introduction

Novel view synthesis is the task of generating photorealistic images from arbitrary viewpoints, has been a fundamental challenge in computer vision and graphics. Recently, neural radiance fields (NeRF)[1] and its successors have shown remarkable performance in this task by representing 3D scenes as a continuous volumetric function parameterized by a neural network.

However, the original NeRF heavily relies on the availability of dense input views [2] to produce high-fidelity synthesized views. This drawback limits the application of NeRF in fields where obtaining such datasets is difficult.

To address such issues, some studies exploit depth maps to supervise the few-shot NeRFs. For example, SparseNeRF [3] proposes a local depth ranking method to align the depth ranking of the NeRF with the local depth map patches. FSGS [4] injects geometry coherence from monocular depth to minimize the distribution differences.

Based on these two methods, this work aims to study and evaluate the effectiveness of different depth supervision and different loss functions defined with depth on the popular Local Light Field Fusion (LLFF) dataset.

2 Related Works

Neural Radiance Fields

Neural Radiance Fields [1] have revolutionized the field of novel view synthesis by representing the scene as a continuous volumetric function parameterized by a neural network. The core idea behind NeRF involves training a multi-layer perception (MLP) that takes the positions (x, y, z) and view direction (θ, ϕ) of the camera, and outputs the emitted radiance value and the volume density σ.

Volumetric Rendering

The NeRF represents the scene as the volume density and directional emitted radiance at any point in the space [1]. The novel views are rendered using methods from classical volumetric rendering [5]. Given a camera ray $\mathbf{r}(t) = \mathbf{o} + t\mathbf{d}$, the expected color $C(r)$ retrieved by this ray with near and far bounds t_n and t_f is given by:

$$C(\mathbf{r}) = \int_{t_n}^{t_f} T(t)\sigma(\mathbf{r}(t))\mathbf{c}(\mathbf{r}(t), \mathbf{d})\mathrm{d}t \tag{1}$$

where $T(t) = \exp(-\int_{t_n}^{t} \sigma(\mathbf{r}(s))ds)$ is the accumulated transmittance from point t_n to t along the ray. $\sigma(\mathbf{r}(t))$ is interpreted as the degree to which the substance absorbs or scatters the radiance at a given point along the ray. $\mathbf{c}(\mathbf{r}(\mathbf{t}), \mathbf{d})$ is the emitted radiance at a point towards direction $\mathbf{d}$.

Depth Recovery

The methods to recover the depth map of a scene can be grouped into three categories [6]: (1) Depth completion. The goal is to recover a dense depth map from a sparse one, such as from the point cloud acquired by LiDAR. (2) Binocular depth estimation. A classic way is to perform stereo camera depth estimation. It is used to measure the disparity between the image pairs captured by a stereo of cameras with known transformations. The disparity is then used to determine the depth map [7]. Learning-based methods such as [8] and [9] have been proposed to generate depth maps with scene understanding. (3) Monocular depth estimation. This group of methods aims to estimate the depth map from a single image. Most of the methods in this category are learning-based, including [10], [11] and [12]. Recently, by adopting Vision Transformer (ViT) [13] as the encoder, remarkable results have been produced by [14], [15].

3 Depth Supervision Methods

Following the definition of color in volumetric rendering in Equation 1, the depth map of NeRF can be computed as:

$$D(\mathbf{r}) = \int_{t_n}^{t_f} T(t)\sigma(\mathbf{r}(t))t\mathrm{d}t \tag{2}$$

From this depth map generated by sampling N rays, there are two categories for performing depth supervision. Denote the depth map predicted by the model as $\hat{D}(\mathbf{r})$ and the ground truth as $D(\mathbf{r})$.

Direct Supervision

Direct Supervision compares $\hat{D}(\mathbf{r})$ and $D(\mathbf{r})$ directly, usually by computing a norm of the difference between them. From this, it is possible to define the loss functions. For example, the L1 loss and mean squared error (MSE) loss are as follows:

$$\mathcal{L}_{L1} = |D(\mathbf{r}) - \hat{D}(\mathbf{r})| \tag{3}$$

$$\mathcal{L}_{\mathrm{MSE}} = \|D(\mathbf{r}) - \hat{D}(\mathbf{r})\|_2^2 \tag{4}$$

Oftentimes, the ground truth depth map is not in the same range as the predicted depth map. Here I use the following formula to normalize the ground truth depth map $D(\mathbf{r})$ as:

$$D_{\mathrm{normalized}}(\mathbf{r})[i] = \alpha + (\beta - \alpha)\frac{D(\mathbf{r})[i] - \min(D(\mathbf{r}))}{\max(D(\mathbf{r})) - \min(D(\mathbf{r}))} \tag{5}$$

,

where α and β are the lower and upper bounds of the normalized range respectively. After the normalization, apply direct supervision using Equation 3 and Equation 4.

Indirect Supervision

Indirection supervision often exploits the distribution relationships between $\hat{D}(\mathbf{r})$ and $D(\mathbf{r})$. In SparseNeRF [3], the authors introduce the local depth ranking distillation criteria. They recognize that it is hard to achieve accurate depth prediction due to dataset bias, coarse depth annotations, and imperfect neural models. Given a local patch of the RGB image I, $\hat{D}(\mathbf{r})$ is predicted. On the other hand, they use a pre-trained depth DPT [16] to produce the ground truth depth $D_{\mathrm{dpt}}(\mathbf{r})$. Crop a local patch from each of $\hat{D}(\mathbf{r})$ and $D_{\mathrm{dpt}}(\mathbf{r})$ at the same spatial location and denote them as $\hat{d}(\mathbf{r})$ and $d_{\mathrm{dpt}}(\mathbf{r})$. Let k_1 and k_2 be two randomly sampled 2D pixel coordinates of $\hat{d}(\mathbf{r})$ and $d_{\mathrm{dpt}}(\mathbf{r})$. The depth ranking loss is defined as:

$$\mathcal{L} = \sum_{d^{k_1}_{\mathrm{dpt}} \leq d^{k_2}_{\mathrm{dpt}}} \max(d^{k_1}_{\mathbf{r}} - d^{k_2}_{\mathbf{r}} + m, 0) \tag{6}$$

,

where m is a small margin that allows for some ranking errors.

FSGS [4] introduces geometry coherence criteria from monocular depth estimates. In order to mitigate the scale ambiguity between the true scene scale and the estimated depth, they propose a relaxed relative loss implemented by Pearson correlation. Pearson correlation coefficient is a correlation coefficient ranging from -1 to 1 that measures the linear correlation between two sets of data. A coefficient close to -1 means negative linear correlation and close to 1 means positive linear correlation.

Recall that the predicted depth is $\hat{D}(\mathbf{r})$ and the ground truth is $D(\mathbf{r})$. The Pearson correlation between $\hat{D}(\mathbf{r})$ and $D(\mathbf{r})$ is defined as:

$$\mathrm{Corr}(\hat{D}(\mathbf{r}), D(\mathbf{r})) = \frac{\mathrm{Cov}(\hat{D}(\mathbf{r}), D(\mathbf{r}))}{\sqrt{\mathrm{Var}(\hat{D}(\mathbf{r}))\mathrm{Var}(D(\mathbf{r}))}} \tag{7}$$

This constraint allows the structure of depth in the predicted depth to be aligned with ground truth depth, circumventing the inconsistencies in absolute depth values.

4 Experiments and Results

Dataset

Local Light Field Fusion (LLFF) [17] dataset is commonly used in 3D reconstruction, novel view synthesis and neural rendering. This dataset consists of image data and camera pose information of eight indoor and outdoor objects (*fern, flower, fortress, horns, leaves, orchid, room, and trex*). I choose LLFF because both FSGS and SparseNeRF support the training and testing pipeline of this dataset.

Evaluation Metric

I evaluate the synthesized views at test viewpoints against the ground truth using commonly adopted metrics in the novel view synthesis literature: Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index Measure (SSIM) [18] and Learned Perceptual Image Patch Similarity (LPIPS) [19].

Results and Findings

Due to the slow training speed of SparseNeRF (about 2 hours per scene), I limit the experiment to the performances of different depth loss functions on the FSGS implementation only.

Method	Depth Supervision	Metrics	Scenes							
			fern	flower	fortress	horns	leaves	orchid	room	trex
FSGS	None	SSIM↑	0.73	0.626	0.737	0.682	0.63	0.493	**0.844**	0.809
		PSNR↑	22.168	20.339	23.549	19.38	17.622	15.96	**21.929**	**21.827**
		LPIPS↓	0.209	0.245	0.175	0.264	**0.205**	0.275	0.172	0.155
	L1	SSIM↑	**0.732**	0.638	0.728	0.692	**0.632**	**0.511**	0.829	0.781
		PSNR↑	**22.251**	**20.618**	23.3	19.705	17.66	**16.349**	21.06	20.912
		LPIPS↓	0.216	0.248	0.191	0.263	0.217	0.269	0.188	0.184
	MSE	SSIM↑	0.715	0.613	0.718	0.696	0.62	0.499	0.839	0.807
		PSNR↑	21.822	20.241	23.17	19.747	17.465	16.099	21.553	21.638
		LPIPS↓	0.205	0.251	0.181	0.233	0.215	0.261	0.17	0.158
	Rank	SSIM↑	0.715	**0.641**	**0.747**	0.657	0.34	0.502	0.839	0.788
		PSNR↑	21.422	20.184	**24.055**	18.434	12.629	15.974	21.029	20.831
		LPIPS↓	**0.201**	**0.232**	**0.156**	0.255	0.515	**0.259**	**0.159**	0.164
	Pearson	SSIM↑	0.718	0.633	0.732	**0.729**	0.63	0.51	0.843	**0.811**
		PSNR↑	21.79	20.517	23.759	**20.645**	**17.719**	16.24	21.351	21.815
		LPIPS↓	0.209	0.241	0.171	**0.224**	0.206	0.263	0.17	**0.153**

Table 1: Evaluation and comparison of FSGS trained with different depth supervision loss functions on the LLFF [17] dataset. The best and the second-best results are shown in **bold** and underlined, respectively.

Table 1 shows the scores of each depth loss in different scenes. The Pearson correlation loss has the greatest number of best and the second-best results and achieves consistent and the best performance across all scenes, demonstrating its robustness. The rest of the losses have comparable performances.

Interestingly, the L1 loss consistently outperforms the MSE loss. MSE loss penalizes large errors quadratically, making it sensitive to outliers in the ground truth or predicted depth map. As for intuition on NeRF, I think MSE tends to penalize small variations heavily even when they are visually insignificant, which leads to overfitting. L1, by contrast, does not penalize these variations as heavily as MSE does.

In addition, the ranking supervision works surprisingly well when evaluated against LPIPS compared to other methods, and especially well in the *fortress* scene. This may be explained by simple ground truth depth shown in Figure 1. Ranking loss performs well when the structure of the scene is simple or the transition depth is smooth. This may as well explain why it struggles in the *leaves* scene because the depth of background is inaccurate and not well-defined.

Figure 1 is a grid of depth maps generated by FSGS [4] with different depth losses. Without depth supervision (None), the depth maps exhibit blending between the foreground object and the background. The depth maps from L1 loss and MSE loss are moderately sharp but contain a lot of noise. Pearson correlation loss produces sharp boundaries and is smoother than L1 loss and MSE loss. The depth maps generated by the ranking loss are the most similar to the ground truth depth map.

The findings above demonstrate that the performances of advanced depth supervisions (ranking loss, Pearson correlation loss) exhibit noticeable improvement upon simple or even no supervisions (None, L1 loss, MSE loss), which is the exact opposite of what Wang et al. [6] suggests.

5 Conclusions and Future Work

This study evaluated the impact of various depth supervision techniques, including L1 loss, MSE loss, ranking less and Pearson correlation loss, using the FSGS method on the LLFF dataset for novel view synthesis. Among all techniques, the Pearson correlation loss demonstrated consistent and robust performance across different scenes. The ranking loss also showed strong results in simple scenes with well-defined depth structures but struggled with more complex scenes. These findings suggest

that, compared to direct depth supervisions, using appropriate indirect depth supervisions to train NeRFs can achieve higher performance in few-shot novel view synthesis tasks.

Several directions remain open for future exploration: 1) Combine depth supervision methods. This hybrid approach may leverage the strengths of both direct and indirect supervision methods to achieve better improvements. 2) Experiment with other models. Due to the time scope and limited computing resources of this project, I did not conduct experiments on other models like SparseNeRF [3]. 3) Investigate the impact of the quality of the ground truth depth map. The abrupt increase in performance of the ranking loss in a simple scene with an accurate ground truth depth map may imply that the quality of ground truth can impact the performance of depth supervision significantly.

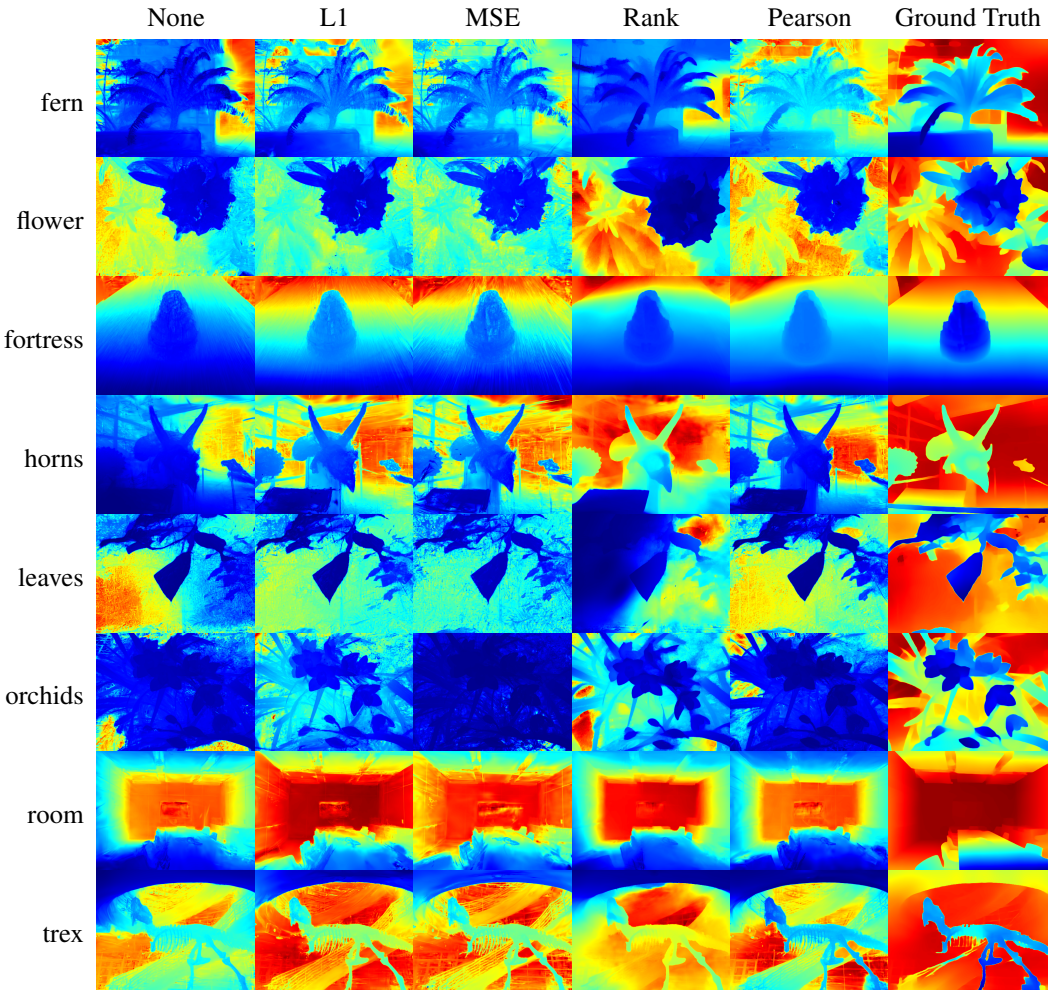

Figure 1: Depth map generated by the FSGS [4]. Depth supervision methods are on the horizontal axis, and scene type is on the verticle axis. Enlarge the figure for better visibility.

References

[1] 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: *Communications of the ACM* 65.1 (2021), pp. 99–106.

[2] Prune Truong, Marie-Julie Rakotosaona, Fabian Manhardt, and Federico Tombari. "Sparf: Neural radiance fields from sparse and noisy poses". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, pp. 4190–4200.

[3] Guangcong Wang, Zhaoxi Chen, Chen Change Loy, and Ziwei Liu. "Sparsenerf: Distilling depth ranking for few-shot novel view synthesis". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023, pp. 9065–9076.

[4] Zehao Zhu, Zhiwen Fan, Yifan Jiang, and Zhangyang Wang. "Fsgs: Real-time few-shot view synthesis using gaussian splatting". In: *European Conference on Computer Vision*. Springer. 2025, pp. 145–163.

[5] James T. Kajiya and Brian P Von Herzen. "Ray tracing volume densities". In: *SIGGRAPH Comput. Graph.* 18.3 (Jan. 1984), pp. 165–174. ISSN: 0097-8930. DOI: 10.1145/964965.808594. URL: https://doi.org/10.1145/964965.808594.

[6] Chen Wang, Jiadai Sun, Lina Liu, Chenming Wu, Zhelun Shen, Dayan Wu, Yuchao Dai, and Liangjun Zhang. "Digging into depth priors for outdoor neural radiance fields". In: *Proceedings of the 31st ACM International Conference on Multimedia*. 2023, pp. 1221–1230.

[7] Daniel Scharstein and Richard Szeliski. "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms". In: *International journal of computer vision* 47 (2002), pp. 7–42.

[8] Alex Kendall, Hayk Martirosyan, Saumitro Dasgupta, Peter Henry, Ryan Kennedy, Abraham Bachrach, and Adam Bry. "End-to-end learning of geometry and context for deep stereo regression". In: *Proceedings of the IEEE international conference on computer vision*. 2017, pp. 66–75.

[9] Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. "Dust3r: Geometric 3d vision made easy". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024, pp. 20697–20709.

[10] David Eigen, Christian Puhrsch, and Rob Fergus. "Depth map prediction from a single image using a multi-scale deep network". In: *Advances in neural information processing systems* 27 (2014).

[11] David Eigen and Rob Fergus. "Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture". In: *Proceedings of the IEEE international conference on computer vision*. 2015, pp. 2650–2658.

[12] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. "Depth anything: Unleashing the power of large-scale unlabeled data". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024, pp. 10371–10381.

[13] Alexey Dosovitskiy. "An image is worth 16x16 words: Transformers for image recognition at scale". In: *arXiv preprint arXiv:2010.11929* (2020).

[14] Chaoqiang Zhao, Youmin Zhang, Matteo Poggi, Fabio Tosi, Xianda Guo, Zheng Zhu, Guan Huang, Yang Tang, and Stefano Mattoccia. "Monovit: Self-supervised monocular depth estimation with a vision transformer". In: *2022 international conference on 3D vision (3DV)*. IEEE. 2022, pp. 668–678.

[15] Ruicheng Wang, Sicheng Xu, Cassie Dai, Jianfeng Xiang, Yu Deng, Xin Tong, and Jiaolong Yang. "Moge: Unlocking accurate monocular geometry estimation for open-domain images with optimal training supervision". In: *arXiv preprint arXiv:2410.19115* (2024).

[16] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. "Vision transformers for dense prediction". In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2021, pp. 12179–12188.

[17] Ben Mildenhall, Pratul P Srinivasan, Rodrigo Ortiz-Cayon, Nima Khademi Kalantari, Ravi Ramamoorthi, Ren Ng, and Abhishek Kar. "Local light field fusion: Practical view synthesis with prescriptive sampling guidelines". In: *ACM Transactions on Graphics (ToG)* 38.4 (2019), pp. 1–14.

[18] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. “Image quality assessment: from error visibility to structural similarity”. In: *IEEE transactions on image processing* 13.4 (2004), pp. 600–612.

[19] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. “The unreasonable effectiveness of deep features as a perceptual metric”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 586–595.