SCIENCE CHINA
Information Sciences

• MOOP •

August 2025, Vol. 68, Iss. 8, 184101:1–184101:3
https://doi.org/10.1007/s11432-023-4512-7

PD-NeRF: a general pseudo-depth supervision method for neural radiance fields

Jiaming GU[1,2], Minchao JIANG[1], Xiaoyuan LU[3], Cong HUA[1], Hongsheng LI[1], Guangming ZHU[1] & Liang ZHANG[1*]

[1]*School of Computer Science and Technology, Xidian University, Xi'an 710071, China*
[2]*QINGYI (Shanghai) ITS, Shanghai 201306, China*
[3]*Shanghai Lingang SCDC, Shanghai 201306, China*

Received 15 March 2023/Revised 23 September 2024/Accepted 19 November 2024/Published online 10 July 2025

Citation Gu J M, Jiang M C, Lu X Y, et al. PD-NeRF: a general pseudo-depth supervision method for neural radiance fields. Sci China Inf Sci, 2025, 68(8): 184101, https://doi.org/10.1007/s11432-023-4512-7

Synthesizing novel views of a scene and reconstructing a 3D scene from a sparse set of captured images has been a long-standing challenge in computer vision. Neural radiance fields (NeRF) [1] is a seminal work that introduced a breakthrough approach to 3D reconstruction and novel view synthesis. NeRF differs from traditional 3D reconstruction methods, which represent scenes using explicit structures such as point clouds, grids, and voxels. NeRF sampling points along each ray, determine the 3D location $\boldsymbol{x} = (x, y, z)$ of each sampling point and the 2D viewing direction $\boldsymbol{d} = (\theta, \phi)$ of the ray. These 5D vectors are then fed into a neural network to obtain the color $\boldsymbol{c} = (r, g, b)$ and volume density σ of the sampling point. In other words, NeRF constructs a field parameterized by an multilayer perceptron (MLP) neural network $F_\theta : (\boldsymbol{x}, \boldsymbol{d}) \rightarrow (\boldsymbol{c}, \sigma)$ to reconstruct the scene and continuously optimize parameters θ. However, the traditional NeRF method requires nearly a week to train a single scene. In addition, rendering speed is slow, generating a single image takes several minutes, and the resulting scene reconstruction often lacks fine detail [2].

Accelerating convergence and improving the quality of radiance field reconstruction are critical challenges in NeRF. Instant-NGP [3] demonstrates that by using multi-resolution hash coding and optimizing the sampling structure, training time can be reduced, and volume rendering can enable real-time scene visualization. However, artifacts, floating objects, and time-consuming processes persist in scenes reconstructed via Instant-NGP. Additionally, Instant-NGP incurs a high computational cost for rendering each ray. To address these issues, we propose using pseudo-depth as supervisory information for NeRF to accelerate convergence and enhance reconstruction quality. Pseudo-depth supervision increases the volume density (σ) of sampling points on the object surface, significantly reducing rendering time.

Taking depth into account, we define the loss as follows:

$$\mathcal{L}(\theta) = \mathcal{L}_{rgb}(\theta) + \mathcal{L}_D(\theta). \tag{1}$$

The total loss comprises two components as follows: color loss and depth loss. The color loss resulting from the L2 photometric reconstruction was consistent with the original NeRF,

$$\mathcal{L}_{rgb}(\theta) = \sum_{r\in\mathcal{R}} \left\| \hat{\boldsymbol{C}}(r) - \boldsymbol{C}(r) \right\|_2^2, \tag{2}$$

where $\mathcal{R}$ is the set of rays in each batch. $\hat{\boldsymbol{C}}(r)$ is the color obtained via volume rendering. $\boldsymbol{C}(r)$ is the ground truth RGB colors for ray r.

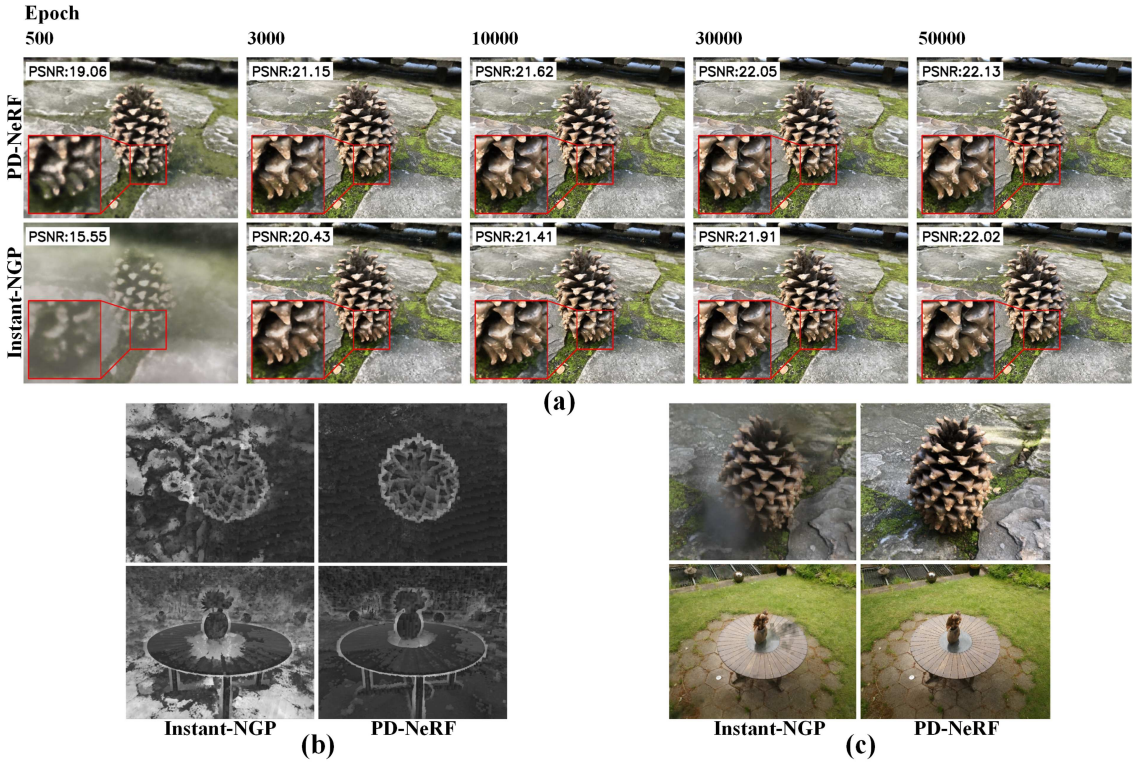
$$\begin{aligned} \hat{\boldsymbol{C}}(r) &= \sum_{i=1}^{N} T_i \left(1 - \exp(-\sigma_i \delta_i)\right) \boldsymbol{c}_i, \\ T_i &= \exp\left(- \sum_{j=1}^{i-1} \sigma_j \delta_j \right), \quad \delta_j = t_{j+1} - t_j. \end{aligned} \tag{3}$$

A ray emitted from the camera to an image pixel is defined as $\boldsymbol{r}(t) = \boldsymbol{o} + t\boldsymbol{d}$ and δ_j represents the distance between two consecutive sampling points along the ray. The pixel color is computed using volume rendering in (3), which allows the color loss to be calculated.

Using an RGB-D camera provides dense depth values; however, it entails high acquisition costs and alignment challenges. Rotation and translation differences between the depth and RGB cameras caused inconsistencies in the pixel coordinate positions for surface points across the two cameras. In addition, the cameras may have different resolutions, with depth cameras typically offering lower resolution. In outdoor environments, ambient light can lead to inaccurate depth measurement, and the RGB-D camera measurement range is often too limited. Considering these challenges, unlike other methods that rely on depth supervision from RGB-D cameras or other sensors, the proposed method derives depth information during the camera pose estimation process.

General NeRF relies on structure-from-motion solvers such as COLMAP [4] to estimate camera poses. During

* Corresponding author (email: liangzhang@xidian.edu.cn)



Method	PSNR↑		SSIM↑		LPIPS↓		FPS	Time to converge (s)
	Dataset1	Dataset2	Dataset1	Dataset2	Dataset1	Dataset2		
Instant-NGP	22.000	24.451	0.631	0.641	0.650	0.757	∼11	∼468
PD-NeRF	**22.285**	**24.708**	**0.653**	**0.678**	**0.617**	**0.720**	**∼ 26**	**∼ 386**

(d)

Figure 1 (Color online) Comparison of performance across different methods. (a) Rendering results of various methods at different epochs; (b) per-pixel rendering cost for different methods (brighter regions indicate a higher cost); (c) effectiveness of floating object removal; (d) evaluation metric results. PSNR: peak signal-to-noise ratio; SSIM: structural similarity index measure; LPIPS: learned perceptual image patch similarity. The best results are in bold.

this process, sparse point clouds are generated. By aligning the camera and the point cloud in the same coordinate system, we obtain the pseudo-depth by calculating the distance between the sparse point cloud and the camera position. This computation incurs minimal overhead and is compatible with any NeRF model. Unlike DS-NeRF [5], which may reduce image resolution to ensure sufficient feature points, the proposed method is better suited for high-resolution images and does not require resolution reduction during training. Compared with DS-NeRF, our depth estimation approach is simpler and more effective. Furthermore, the proposed NeRF model significantly reduces the training and inference time.

Feature points are typically characterized by invariance to scaling, translation, rotation, and brightness changes. However, the feature points extracted from an image are highly sparse. Moreover, during camera pose estimation, only a small subset of these points satisfies the multi-view constraint (i.e., having a reprojection error below a certain threshold), which makes the resulting feature point cloud extremely sparse. Our analysis shows that feature points account for approximately 1/3000 of the total pixels in an image.

For deep supervision, a higher number of feature points is beneficial; however, this increases the preprocessing time significantly. To balance computational efficiency and improve loss convergence stability, we address the issue of sparse feature points by sampling pixels near them using a Gaussian distribution. The proposed method enhances the effectiveness of deep supervision without introducing excessive computational overhead. The Gaussian distribution function is defined in (4).

Because the surface of an object is continuous, it can generally be assumed that the distance between a point near a feature point and the camera varies only slightly. This assumption holds more reliably at points closer to the feature point. Therefore, a Gaussian distribution function that follows this trend is used to represent confidence in describing a point that conforms to this assumption.

$$\omega_{ri} = \exp\left(-\frac{(x-x_i)^2}{2f}\right)\exp\left(-\frac{(y-y_i)^2}{2f}\right), \qquad (4)$$

where ω_{ri} represents the confidence of the pixel depth covered by a feature point. The threshold for ω_{ri} is 0.01. When $\omega_{ri} \leqslant 0.01, \omega_{ri} = 0$. The scaling factor f ensures that the pixels in the regions covered by all feature points

make up approximately 0.6% of the entire image. Experiments have shown that selecting sampling points to occupy 0.6% of the image achieves effective supervision with minimal computational overhead. For an image resolution of $1600 \times 1600, f = 1$. The coordinates of a feature point are denoted as (x_i, y_i).

$$\omega_r = \begin{cases} \sum_{i=1}^{n} \omega_{ri}, & \sum_{i=1}^{n} \omega_{ri} \leqslant 1, \\ 1, & \sum_{i=1}^{n} \omega_{ri} > 1, \end{cases} \tag{5}$$

$$D(r) = \frac{\omega_{r1} D_{r1} + \omega_{r2} D_{r2} + \cdots + \omega_{rn} D_{rn}}{\omega_r}, \tag{6}$$

where D_{ri} denotes the pseudo-depth value of a feature point. A single pixel may be covered by multiple feature points; thus, we apply (6) to weight the depth contribution from each feature point. Additionally, Eq. (5) ensures that the total weight assigned to each depth does not exceed 1. The final pseudo-depth value used for supervision is defined as $D(r)$.

We store the pixel coordinates in 1D format and manage the depth map ($D(r)$) and weight map (ω_r) using HashMap. Compared with uncompressed storage (matrix-based methods), using a HashMap reduces memory usage by more than **9**0%. Furthermore, to scale the model for large scenes while minimizing GPU memory consumption, we store the image in host memory.

$$\mathcal{L}_D(\theta) = \sum_{r \in \mathcal{G}} \omega_r \left\| \hat{D}(r) - D(r) \right\|_2^2. \tag{7}$$

Eq. (7) is the depth loss, $\mathcal{G}$ represents the set of rays in each batch for which ω_r is non-zero. The predicted depth value is obtained through volume rendering as $\hat{D}(r) = \sum_{i=1}^{N} T_i (1 - \exp(-\sigma_i \delta_i)) t_i$. t_i is the distance from the camera origin to the sampling point along the ray.

Experiments and analysis. We evaluated the proposed method using the "Pinecone" dataset (4032 × 3024) from NeRF_real_360 as Dataset1 and the "garden" dataset (1297 × 840) from 360_v2 [6] as Dataset2. The upper layer in Figure 1(a) presents the results obtained by incorporating pseudo-depth supervision. At epoch 30000, the PSNR surpasses the PSNR of the original Instant-NGP at epoch 50000. Additionally, at epoch 500, the PSNR of the model with pseudo-depth supervision was considerably higher than that of the model without it, demonstrating that PD-NeRF approached convergence much earlier. This result indicates that although the final radiance field eventually converges, pseudo-depth supervision considerably accelerates the convergence rate. On the left side of Figure 1(b), the radiance field without depth supervision contains a noticeably larger white region (indicating high rendering overhead) compared with the right side, which suggests increased rendering time. Figure 1(c) further demonstrates that pseudo-depth supervision effectively reduces floating artifacts in the radiance field, thereby improving the overall quality of the 3D reconstruction.

In Figure 1(d), at epoch 50000, we compared the proposed model with Instant-NGP using three evaluation metrics as follows: PSNR, SSIM [7], and LPIPS [8]. The results demonstrate that the proposed model performs well on Dataset1 and Dataset2 consistently outperforming the original radiance field without pseudo-depth supervision across all metrics. In terms of PSNR, the proposed model achieved an improvement of 0.285 (22.000 vs. 22.285) on Dataset1 and 0.257 (24.451 vs. 24.708) on Dataset2 compared with Instant-NGP. These improvements indicate that our rendered images align more closely with human perception.

The proposed method achieved a considerably higher frame rate than the original approach, demonstrating a substantial reduction in rendering overhead. Additionally, the proposed model convergence time measured as the time required to reach the same PSNR value was reduced by nearly 20% (468 s vs. 386 s).

Conclusion. The proposed method achieved strong results across all metrics. By leveraging the proposed pseudo-depth supervision, we effectively accelerate the convergence of the radiance field, significantly reduce the computational overhead of volume rendering, and eliminate floating artifacts to enhance the reconstruction quality. Unlike traditional depth estimation methods that rely on external sensors, pseudo-depth is obtained by computing the distance between the camera coordinates and feature points. This approach is highly effective and computationally efficient; thus, it is adaptable to all NeRF models. In future work, we plan to continue optimizing our NeRF to improve its performance in more complex environments while further reducing computational costs.

Acknowledgements This work was supported by Natural Science Foundation of Shaanxi Province (Grant No. 2024JC-JCQN-66), Science and Technology Commission of Shanghai Municipality (Grant No. 24511106900), and National Natural Science Foundation of China (Grant Nos. 62072358, 62072352).

Supporting information Videos and other supplemental documents. The supporting information is available online at info.scichina.com and link.springer.com. The supporting materials are published as submitted, without typesetting or editing. The responsibility for scientific accuracy and content remains entirely with the authors.

References

1 Mildenhall B, Srinivasan P, Tancik M, et al. NeRF: representing scenes as neural radiance fields for view synthesis. In: Proceedings of European Conference on Computer Vision, Glasgow, 2020. 405–421
2 Martin-Brualla R, Radwan N, Sajjadi M S M, et al. NeRF in the wild: neural radiance fields for unconstrained photo collections. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Montreal, 2021. 7210–7219
3 Müller T, Evans A, Schied C, et al. Instant neural graphics primitives with a multiresolution hash encoding. ACM Trans Graph, 2022, 41: 1–15
4 Schönberger J L, Frahm J-M. Structure-from-motion revisited. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, 2016: 4104–4113
5 Deng K, Liu A, Zhu J Y, et al. Depth-supervised NeRF: fewer views and faster training for free. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, 2022. 12882–12891
6 Barron J T, Mildenhall B, Tancik M, et al. Mip-NeRF: a multiscale representation for anti-aliasing neural radiance fields. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, 2021. 5855–5864
7 Wang Z, Bovik A C, Sheikh H R, et al. Image quality assessment: from error visibility to structural similarity. IEEE Trans Image Process, 2004, 13: 600–612
8 Zhang R, Isola P, Efros A A, et al. The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, 2018. 586–595