

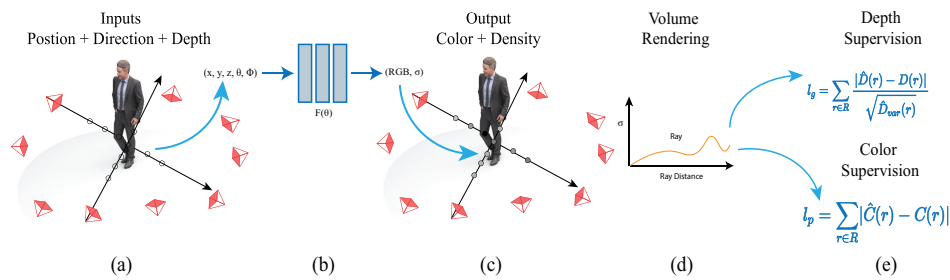


# RGB-D Neural Radiance Fields: Local Sampling for Faster Training

A. Dey<sup>1</sup>  and A. I. Comport<sup>1</sup> 

<sup>1</sup>IS3-CNRS/Université Cote d'Azur - Sophia-Antipolis, France



**Figure 1:** The RGB-D NeRF training process. (a) Depth guided sampling. (b) The input to the network is 5D coordinate. (c) The network outputs volume density and color for each sample. (d) The color and depth of a ray generated using classic volume rendering. (e) The network is optimized using color and depth loss.

## Abstract

Learning a 3D representation of a scene has been a challenging problem for decades in computer vision. Recent advancements in implicit neural representation from images using neural radiance fields (NeRF) have shown promising results. Some of the limitations of previous NeRF based methods include longer training time, and inaccurate underlying geometry. The proposed method takes advantage of RGB-D data to reduce training time by leveraging depth sensing to improve local sampling. This paper proposes a depth-guided local sampling strategy and a smaller neural network architecture to achieve faster training time without compromising quality.

## CCS Concepts

• **Computing methodologies** → **Appearance and texture representations;**

## 1. Introduction and related work

Learning the 3D representation (shape and texture) of a scene is important for novel view synthesis, 3D human modeling, virtual reality, etc. Recent advancement in neural scene representations, more specifically NeRF [MST\*20], showed that neural networks can be used to encode high-quality images of 3D scenes. NeRF based methods use two multilayer perceptrons (coarse and fine) to learn radiance and volume density from RGB images and their corresponding camera poses.

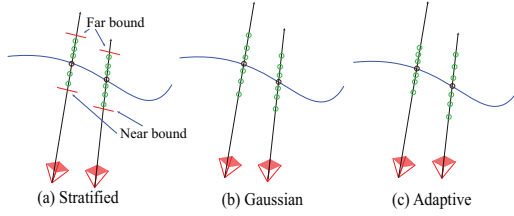
Although NeRF-like methods can produce high quality novel views of scenes, they are too expensive to train. NeRF uses a classic volume rendering technique to compute pixel colors by placing 256 samples along each viewing ray. Each of those samples need a full network evaluation to compute a color. Recently, [NSP\*21] proposed real-time rendering by limiting the number of samples. They use an oracle network and ground truth depth to predict rele-

vant sampling locations on rays. Their method is limited to forward facing scenes and poses belonging to a view cell. Alternatively, [DLZR21] uses sparse depth supervision generated by a Structure-from-motion (SfM) algorithm to optimize the network using both color and depth information together, which allows them to use fewer input views. [SLOD21] achieved real-time SLAM based on NeRF by using a smaller network, lower resolution inputs, and removed the viewing direction. The method proposed here uses local sampling based on a depth sensor to reduce the number of samples and replaces the coarse network of NeRF. This study aims to prove that faster training time can be achieved by local sampling without limiting scene representation quality while using a single network.

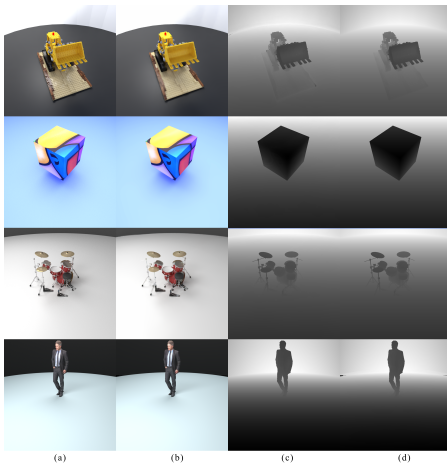
## 2. Local sampling

NeRF-like methods estimate pixel color using classic volume rendering, alpha composition [MST\*20] of sample volume density,

and color to generate the final rendering. Samples with higher volume density have a larger contribution towards the final color. The proposed local sampling places fewer samples only on the relevant part of the rays given depth information.



**Figure 2:** Visualization of three different sampling strategies. Black lines represent rays coming from camera and circles are samples.



**Figure 3:** Qualitative results from simulated data. (a) Ground truth image; (b) Predicted image; (c) Ground truth depth; (d) Predicted depth.

### 2.1. Stratified sampling

This approach is very similar to NeRF [MST\*20] sampling, the only difference is that the near and far bounds of the sampling are set using depth information.

### 2.2. Gaussian sampling

Instead of placing samples in a stratified manner, a Gaussian distribution is used to distribute sample locations around the surface. The mean of the distribution is the depth measurement, which ensures more samples are placed close to the surface.

### 2.3. Adaptive sampling

A multiview depth error map  $\{e^i\}_{i=1}^N$  is generated using all the depth maps of the training set (see the poster for the definition). The standard deviation of the Gaussian distribution is computed from  $\{e^i\}_{i=1}^N$ . It ensures that the spread of the samples are greater when there is more uncertainty in the depth.

## 3. Preliminary Results

dataset	Metrics			
	PSNR $\uparrow$	SSIM $\uparrow$	Abs Rel $\downarrow$	LPIPS $\downarrow$
Lego	27.4	0.933	0.012	0.0009
Cube	37.76	0.95	0.005	0.0001
Drums	29.66	0.91	0.004	0.0008
Human	38.83	0.98	0.003	0.00006

**Table 1:** The results of proposed method tested on 4 different simulated datasets. Underlying geometry is evaluated by Absolute relative distance (AbsRel). Photometric quality evaluated by PSNR (peak signal to noise ratio), SSIM (structural similarity index), and LPIPS (Learned Perceptual Image Patch Similarity).

Strategy	Metrics				
	PSNR $\uparrow$	SSIM $\uparrow$	AbsRel $\downarrow$	LPIPS $\downarrow$	Time $\downarrow$
Stratified	21.81	0.891	<b>0.003</b>	0.002	30m
Gaussian	<b>24.17</b>	<b>0.912</b>	0.017	0.002	<b>22m</b>
Adaptive	23.40	0.910	0.018	0.002	<b>22m</b>
NeRF	22.3	0.84	0.215	0.002	1h 42m

**Table 2:** The proposed local sampling strategies compared with baseline NeRF. The dataset contains 8 training images. Experiments were performed with 16 sample points.

Qualitative results are shown in Table 1 and 2. The Figure 3 shows qualitative results of the proposed local sampling. The best method according to the preliminary result is Gaussian sampling.

## 4. Conclusions

A preliminary study has been presented that shows depth images can be used to perform local sampling and fewer samples can reduce training time without compromising quality. The results suggest that surface information about the scenes can provide additional supervision to achieve better underlying geometry and photometry from fewer input views.

## 5. Acknowledgements

This project funded by EU H2020 COFUND BoostUrCareer, Marie Skłodowska-Curie grant agreement no. 847581.

## References

- [DLZR21] DENG K., LIU A., ZHU J.-Y., RAMANAN D.: Depth-supervised nerf: Fewer views and faster training for free. *arXiv preprint arXiv:2107.02791* (2021). 1
- [MST\*20] MILDENHALL B., SRINIVASAN P. P., TANCIK M., BARRON J. T., RAMAMOORTHI R., NG R.: Nerf: Representing scenes as neural radiance fields for view synthesis. In *European conference on computer vision* (2020), Springer, pp. 405–421. 1, 2
- [NSP\*21] NEFF T., STADLBAUER P., PARGER M., KURZ A., ALLA CHAITANYA C. R., KAPLANYAN A., STEINBERGER M.: Donerf: Towards real-time rendering of neural radiance fields using depth oracle networks. *arXiv e-prints* (2021), arXiv–2103. 1
- [SLOD21] SUCAR E., LIU S., ORTIZ J., DAVISON A. J.: imap: Implicit mapping and positioning in real-time. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2021), pp. 6229–6238. 1