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

A. Dey[1], A. I. Comport[1]

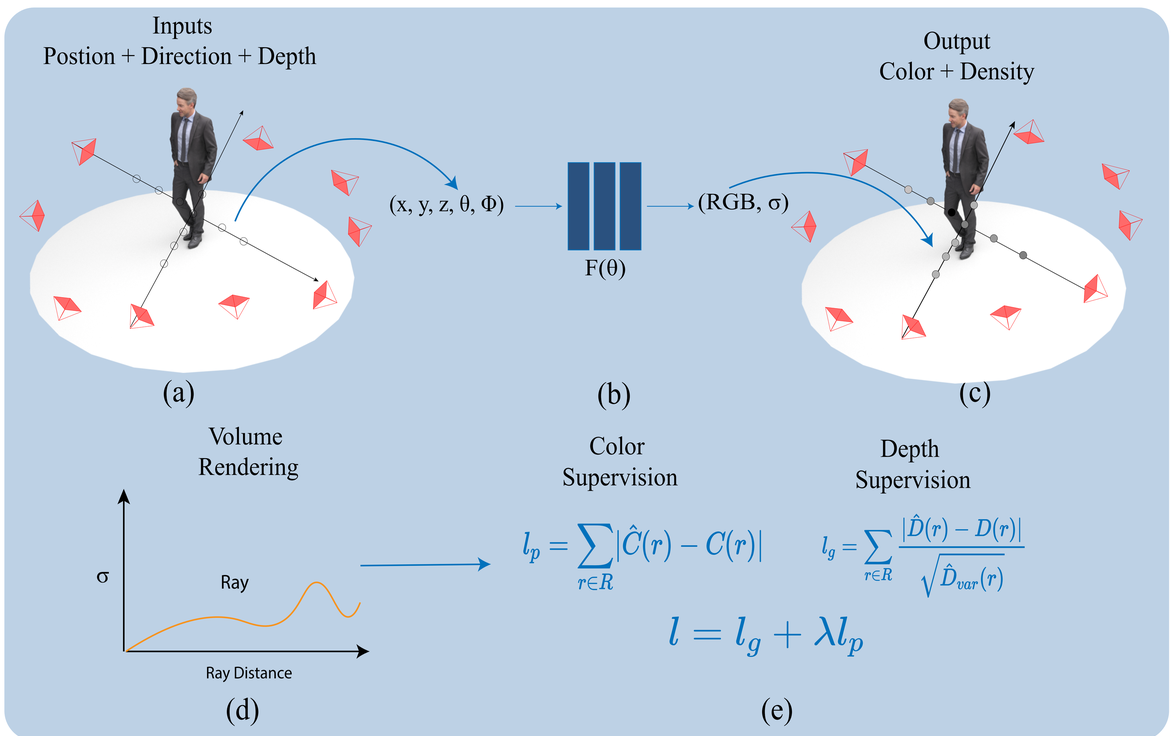
[1] I3S-CNRS/Université Côte d'Azur - Sophia-Antipolis, France

INTRODUCTION

- The aim of this research is to improve neural scene representations (namely NeRF [2]), which provide a continuous scene representation, by incorporating both a color and depth measurements (RGB-D) as opposed to only colour.
- The incorporation of depth information will be shown to improve both the computational efficiency during training (using more efficient sampling) along with more accurate depth representations via a joint optimisation criterion.

Limitation of NeRF →
- Long training time.
- Inaccurate underlying geometry.
- Inefficient sampling.

Proposed method →
- Local sampling to replace coarse network.
- Aditional depth supervision.
- Fewer samples in relevent parts of the scene.
- Depth sensor uncertaintly.

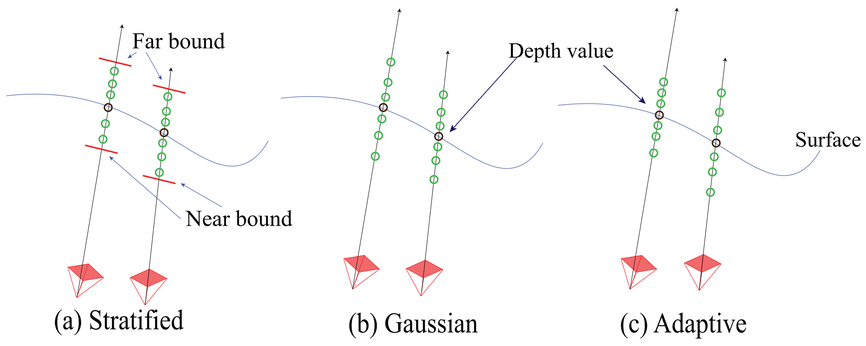


*The RGB-D NeRF training process. **(a)** Depth guided sampling. **(b)** The input to the network is 5D coordinates. **(c)** The network outputs volume density and color for each sample. **(d)** The color and depth of a ray are generated using classic volume rendering. **(e)** The network is optimized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

Related work: [3] proposed local sampling based on an Oracle network but it is **limited** to forward facing views, poses belonging to a view cell. [1] relies on a Structure-from-motion(SfM) algorithm to generate a sparse depth supervision which is **error prone.**

METHOD

The proposed local sampling places fewer samples only on the relevant part of the rays given depth information.



Visualization of proposed sampling strategies. Black lines represent viewing rays. Green circles are samples ponts along the ray.

Stratified Sampling:

Samples are placed between near and far bounds $[t_n, t_f]$ in stratified manner where $t_n = D - \alpha_n$, $t_f = D - \alpha_f$. D is the depth measurement. α_n, α_f are empirically choosen values.

Gaussian Sampling:

Samples are distributed along the ray using a Gaussian distribution, where mean is the depth measurement and standard deviation emprically choosen for each dataset and uncertainty of depth sensor.

Adaptive Sampling:

A multiview depth error map is created using all the depth maps of training set.

$p^{i\to j}, D^{i\to j} = proj(K, T^{i\to j}, D^i)$

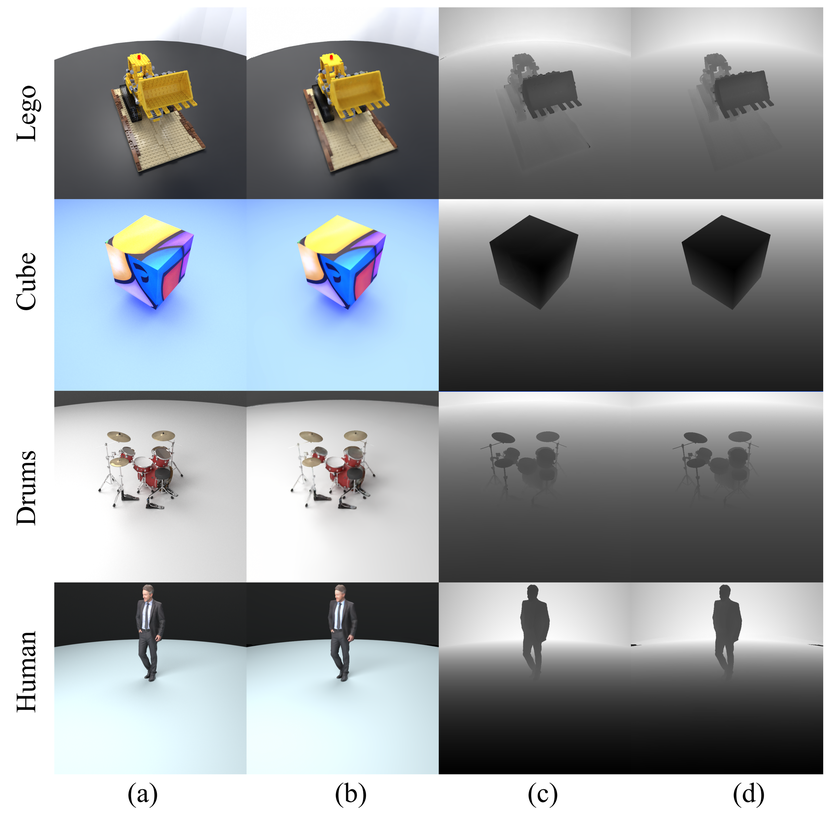
$D^{j'} = D^j(p^{i\to j})$, Where K is intrinsic matrix, $T^{i\to j}$ is relative pose, and $P^{i\to j}$ pixel coordinate $D^{i\to j}$ depth projection.

$e^i = \frac{|D^{j'} - p^{i\to j}, D^{i\to j}|}{p^{i\to j}, D^{i\to j}}$, which is used to calculate stadard deviation of the Gaussian distribution.

$$SD = \begin{cases} \alpha_l \text{ if } e^i < \alpha_l \\ e^i \text{ if } \alpha_l < e^i < \alpha_h \\ \alpha_h \text{ otherwise} \end{cases}$$

, Where α_l and α_h emprically choosen low and high bounds.

RESULTS



Qualitative results. (a) Ground truth image; (b) Predicted image; (c) Ground truth depth; (d) Predicted depth.

	Metrics			
Dataset	PSNR↑	SSIM↑	AbsRel↓	LPIPS↓
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

The results of the proposed method tested on 4 different simulated datasets. The underlying geometry is evaluated with the 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).

	Metrics				
Strategy	PSNR↑	SSIM↑	AbsRel↓	LPIPS↓	Training time↓
Stratified	21.81	0.891	**0.003**	0.002	30m
Gaussian	**24.17**	**0.912**	0.017	0.002	**22m**
Adaptive	23.40	0.910	0.018	0.002	22m
NeRF	22.3	0.84	0.215	0.002	1h 42m

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

The preliminary results show that Gaussian sampling performs best among the proposed local sampling strategies. The local sampling strategies are 4-5 times faster than NeRF for training.

Implementation Details:

The method is implemented using a combination of PyTorch and CUDA. A single MLP network was used for training and prediction.The network has 4 hidden layers and optimized using a RGB-D loss function. 10 frequency bands is used for positional encoding and 4 frequency bands for directional encoding.

CONCLUSIONS

- The preliminary study shows that depth images can be used to perform **local sampling.**
- **Fewer samples** reduce training time without compromising quality.
- Additional information about the surface obtained from depth images can help the network to output a more accurate **underlying geometry.**
- Depth supervision also helps the network to achieve better results from **fewer input views.**

REFERENCES

[1] 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).

[2] 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.

[3] 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.

ACKNOWLEDGEMENTS

This project has received funding from Horizon 2020 COFUND program BoostUrCareer under the Marie Sklodowska-Curie grant agreement no. 847581 supported by European Commission and the conseil Region Sud-Provence-Alpes-Cote d'Azur and IDEX UCA[JEDI].

CONTACT: adey@i3s.unice.fr; Andrew.Comport@cnrs.fr

EG'22 TheEvent

Co-Funded by the European Union