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

A. Dey and A. I. Comport

1 I3S-CNRS/Université Côte d’Azur - Sophia-Antipolis, France

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 relevant 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.

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 \( \{\varepsilon_i\}_{i=1}^N \) 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 \( \{\varepsilon_i\}_{i=1}^N \). It ensures that the spread of the samples are greater when there is more uncertainty in the depth.

3. Preliminary Results

<table>
<thead>
<tr>
<th>Strategy</th>
<th>PSNR↑</th>
<th>SSIM↑</th>
<th>Abs Rel↓</th>
<th>LPIPS↓</th>
<th>Time↓</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stratified</td>
<td>21.81</td>
<td>0.891</td>
<td>0.003</td>
<td>0.002</td>
<td>30m</td>
</tr>
<tr>
<td>Gaussian</td>
<td>24.17</td>
<td>0.912</td>
<td>0.017</td>
<td>0.002</td>
<td>22m</td>
</tr>
<tr>
<td>Adaptive</td>
<td>23.40</td>
<td>0.910</td>
<td>0.018</td>
<td>0.002</td>
<td>22m</td>
</tr>
<tr>
<td>NeRF</td>
<td>22.3</td>
<td>0.84</td>
<td>0.215</td>
<td>0.002</td>
<td>1h 42m</td>
</tr>
</tbody>
</table>

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).

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 Sklodowska-Curie grant agreement no. 847581.

References


© 2022 The Author(s) Eurographics Proceedings © 2022 The Eurographics Association.