DBARF: Deep Bundle-Adjusting Generalizable Neural Radiance Fields

Yu Chen Gim Hee Lee
National University of Singapore

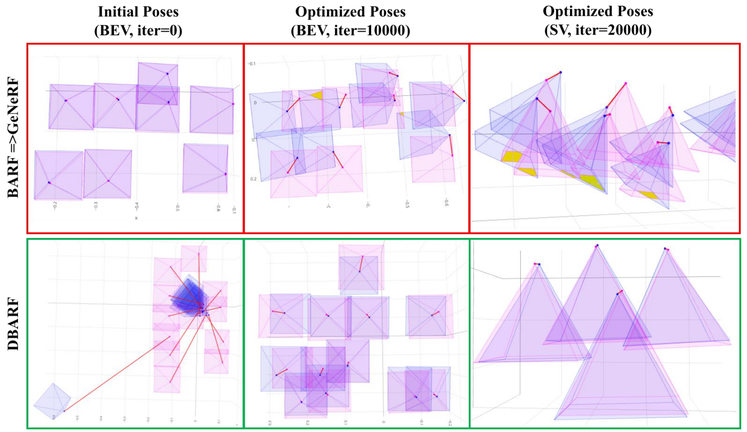
JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

Introduction

BARF is very popular in jointly optimizing camera poses and per-scene NeRFs. However, bundle-adjusting Generalizable NeRF is more challenging.

Question: Can we bundle-adjust Generalizable NeRF as is done in BARF?

Answer: Yes. But not as simple as BARF...

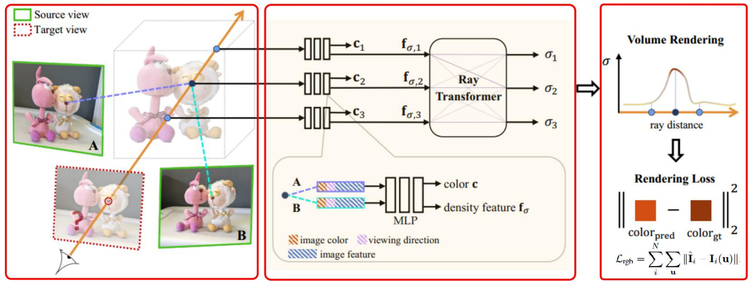


A Typical Pipeline of GeNeRF

1. Projecting a point onto the feature maps of the nearby images $\mathbf{f} = \chi(\Pi(\mathbf{P}_j, \omega(\mathbf{X}_i^k, \mathbf{P}_i)), \mathbf{F}_j)$

2. Aggregating pixel-level features for emitted ray $g_k = f_a(\mathbf{f}_1^k, \mathbf{f}_2^k, \cdots, \mathbf{f}_M^k)$

3. Rendering the target image $\hat{\mathbf{I}}_{\text{target}} := \hat{\mathbf{I}}_i = h(g_1, \cdots, g_K; \mathbf{\Phi})$



Approach

What prevents BA camera poses with GeNeRFs?

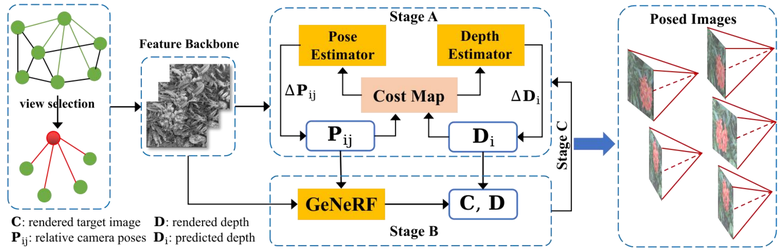
1. $\mathbf{f} = \chi(\Pi(\mathbf{P}_j, \omega(\mathbf{X}_i^k, \mathbf{P}_i)), \mathbf{F}_j)$ 2. $g_k = f_a(\mathbf{f}_1^k, \mathbf{f}_2^k, \cdots, \mathbf{f}_M^k)$ 3. $\hat{\mathbf{I}}_{\text{target}} := \hat{\mathbf{I}}_i = h(g_1, \cdots, g_K; \mathbf{\Phi})$

$$\mathcal{L}_{\text{rgb}} = \sum_i^N \sum_{\mathbf{u}} \|\hat{\mathbf{I}}_i - \mathbf{I}_i(\mathbf{u})\|$$

$$\frac{\partial \mathcal{L}_{\text{rgb}}}{\partial \mathbf{P}_j} = \underbrace{\sum_{i \neq j}^{N} \sum_{\mathbf{u}} \sum_{k}^{K} \frac{\partial h}{\partial g_k} \cdot \frac{\partial g_k}{\partial \mathbf{f}_i^k} \cdot \frac{\partial \mathbf{f}_i^k}{\partial \mathbf{P}_j}}_{\text{image } j \text{ is one of the nearby views of image } i} + \underbrace{\sum_{m}^{M} \sum_{\mathbf{u}} \sum_{k}^{K} \frac{\partial h}{\partial g_k} \cdot \frac{\partial g_k}{\partial \mathbf{f}_m^k} \cdot \frac{\partial \mathbf{f}_m^k}{\partial \mathbf{P}_j}}_{\text{image } j \text{ is the target image}}$$

- **Difficulties**
  - Occulusion contributes feature outliers
  - Image feature space is highly non-smooth

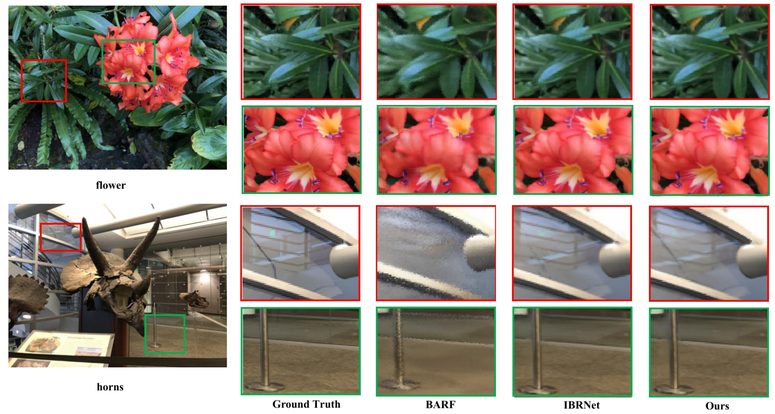
Network Architecture



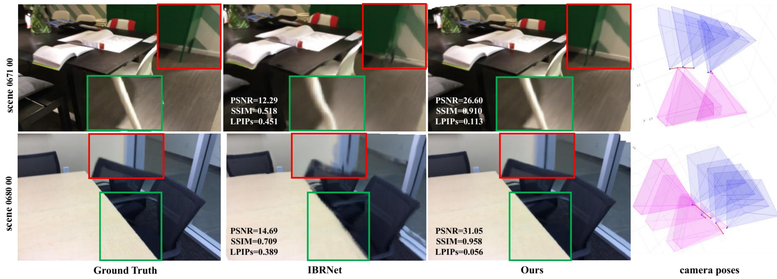
(1) Nearby views are selected from a scene graph since the camera poses are unknown.
(2) Image features are extracted by FPN.
(3) In stage A, the image feature of the target view is warped to each nearby view by the corresponding current camera poses and depth, a cost map is constructed by the image feature difference.
(4) In stage B, we utilize a generalizable NeRF to predict image color and density value, and the final image is rendered by volume rendering.
(5) In stage C, the pose optimizer and the generalizable NeRF are jointly learned.
(6) Finally, our network outputs the posed images.

Results

LLFF Dataset



ScanNet Dataset



Check out our project page for more details and discussions!

https://aibluefisher.github.io/dbarf

[1] Chen-Hsuan Lin, et. at. BARF: Bundle-Adjusting Neural Radiance Fields. ICCV 2021
[2] Qianqian Wang, *et, al.* IBRNet: Learning Multi-View Image-Based Rendering. CVPR 2021