Refocus-NeRF: Focus-Distance-Aware Neural Radiance Fields Trained with Focus Bracket Photography

Yuki Yabumoto, Takuhiro Nishida, Takashi Ijiri

Shibaura Institute of Technology

1 Introduction

Focus bracketing is a technique for quickly capturing a sequence of photographs by changing the focus distance. Focus stacking is a technique for synthesizing a single image from a sequence of photographs taken by focus bracketing, such that the image has a greater depth of field (DoF) and all subjects are entirely in focus. They are particularly useful for photographing small objects, such as flowers and insects, which require a macro lens with a very shallow DoF. Researchers have combined focus bracketing and focus stacking with photogrammetry to reconstruct three-dimensional (3D) shapes of small specimens [1, 4]. However, these photogrammetry-based methods have limitations in reconstructing transparent or specular objects.

Neural Radiance Fields (NeRF) [2] is a novel view synthesis method enabling reconstruction and rendering of 3D scenes from sparse 2D photographs. It represents a 3D scene with a multilayer perceptron (MLP) that takes a 3D location $\mathbf{x}$ and viewing direction $\mathbf{d}$ as inputs and outputs density and color at $\mathbf{x}$ viewed from $\mathbf{d}$. NeRF can reconstruct 3D scenes with transparent and specular objects in principle because it trains the MLP to render images with similar appearances to the input photographs. However, the original NeRF assumes input of deep DoF photographs and does not consider defocus blur effects. Wu et al. [5] added defocus blur effects to NeRF framework by blending the colors of neighboring rays. However, the method assumes that the radiance of each sampling ray is concentrated at a specific depth for screen space blending.

Our goal is to reconstruct a 3D scene from sequences of focus bracketing photographs, incorporating the inherent defocus blur effects. We present refocus-NeRF, which extends NeRF representation to receive the focus distance as well as the location and viewing direction as inputs. We train it with sequences of focus bracketing photographs taken from different viewpoints. Because we train the network with focus bracketing photographs, it can render images with defocus blur effects similar to actual photographs, even for scenes containing transparent or specular objects. Our method represents defocus blur as density and color in the 3D scene and dynamically modifies the scene according to the focus distance; it can render a scene using a standard camera ray sampling of the original NeRF without additional screen space blending.

2 Method

The refocus-NeRF model takes three inputs variables; location $\mathbf{x} \in \mathbb{R}^3$, viewing direction $\mathbf{d} = (\theta, \phi) \in \mathbb{R}^2$, and focus distance $f \in [0,1]$. The output of the model is the density $\sigma \in \mathbb{R}$ and RGB color $\mathbf{c} \in \mathbb{R}^3$ at the $\mathbf{x}$ viewed from $\mathbf{d}$. Because our method produces defocus blur effects by adjusting density and color, we input $\mathbf{x}$, $\mathbf{d}$, and f to the density MLP. We also apply multiresolution hash encoding to $\mathbf{x}$ and spherical harmonics encoding to $\mathbf{d}$ similarly to Instant-NGP [3].

We collect datasets to train the network as follows. We first perform focus bracketing from different viewpoints to obtain multiple sequences of photographs. We scale all photographs to align these in each sequence because the angles of view of photographs in a sequence slightly shift along with the change in focal distances (i.e., focus breathing). We define focus distance for a photograph in each sequence such that the photographs with the minimum and maximum focus distances have $f = 0.0$ and $f = 1.0$, respectively, while those in between have f values linearly interpolating 0.0 and 1.0. We next compute the focus stacking image for each bracketing sequence. Finally, we apply structure from motion to all focus stacking images to obtain their 3D poses.

We train the network using sequences of bracketing photographs after the alignment. We minimize the error between the observed photographs and the rendered images of refocus-NeRF from the corresponding viewpoints. Especially, when training with a photograph with focus distance f_i, we input f_i to the network, which enables us to obtain scene dynamically change according to focus distance. We use the Huber loss function for evaluating the error. We implemented our prototype system based on Instant-NGP [3].

Figure 1: We perform focus bracketing from multiple viewpoints and train a network using the obtained photographs (a). We synthesize from unknown cameras with different focus distances (b) and (c).

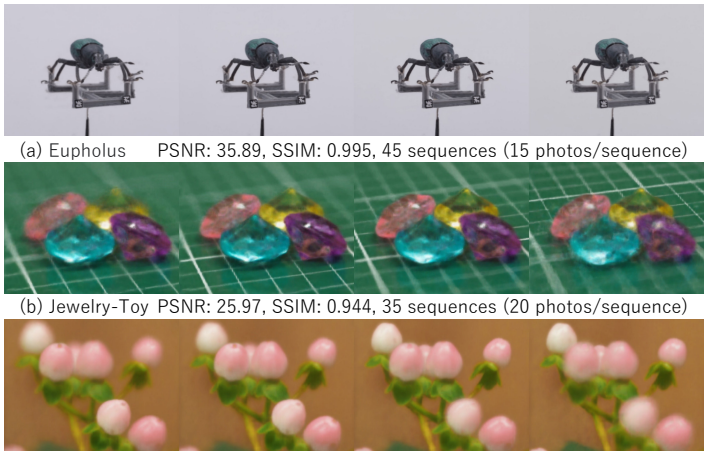

Figure 2: We render three scenes from an unknown camera pose with different f values. Each row shows the accuracy, the number of viewpoints, and the number of photos. We trained the network 40k steps (~15 min).

3 Results and Discussion

We reconstructed four different scenes to demonstrate the feasibility of the refocus-NeRF. Figures 1 and 2 present the rendering results from unknown camera poses with varying focus distances. Figure 2 summarizes the number of focus bracketing sequences (viewpoints) and photographs within each sequence. It also shows the accuracies (PSNR/SSIM) of the rendering results for viewpoints that were not used during training. Our method accurately rendered images containing out-of-focus blur. Our method dynamically modifies the scene using focus distance f; therefore, it can render images with different in-focus positions only with conventional camera ray sampling without screen space blending.

One limitation of our method is its memory cost; it requires more photographs than the traditional NeRF since multiple photographs exist at each viewpoint. In the future, we would like to extend our method to reconstruct similar scenes with fewer photographs. Another future work is to improve our approach to handle scenes beyond forward-facing scenes.

[1] T.-N. Doan and C. V. Nguyen. A low-cost digital 3d insect scanner. *Information Processing in Agriculture*, 2023.

[2] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Commun. ACM*, 2021.

[3] T. Müller, A. Evans, C. Schied, and A. Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Trans. Graph.*, 2022.

[4] Y. Qiu, D. Inagaki, K. Kohiyama, H. Tanaka, and T. Ijiri. Focus stacking by multi-viewpoint focus bracketing. In *SIGGRAPH Asia 2019 Posters*, 2019.

[5] Z. Wu, X. Li, J. Peng, H. Lu, Z. Cao, and W. Zhong. Dof-nerf: Depth-of-field meets neural radiance fields. In *Proceedings of ACM MM '22*, 2022.