AI 3D Selfie: Real-Time Single-Image 3D Face Reconstruction for Light-Field Displays

Jonghyun Kim, Michael Stengel, Matthew Chan, Koki Nagano, Shalini De Mello, and David Luebke

NVIDIA, Santa Clara, CA

Abstract

We present AI 3D Selfie, a system that enables users to capture their facial images using a single 2D camera and visualize them in 3D in real time. Our method performs real-time single-shot 3D reconstruction by employing a triplane-based NeRF encoder and a fast volumetric rendering algorithm to display the results on a light field display.

Author Keywords

3D uplifting, Neural Radiance Fields, Light Field Displays, Volume Rendering.

1. Introduction

Selfie, a self-taken facial image with a front-facing camera, has become a big industry over the last decade. Social network services provide various tools to support the capturing, editing, and uploading of selfies. In addition, they offer AI-based filters or augmentations using face recognition algorithms. However, these efforts are still confined to 2D, limited by display capabilities and computational complexity.

On the other hand, 3D face reconstruction from a single unposed or multiple 2D images has become feasible with advancements in generative AI and neural 3D representation techniques [1]. Methods such as Nerfies [2], Nersemble, [3] and CAFCA [4] generate a 3D representation using NeRF (Neural Radiance Fields) [5] from multi-view captures or few-shot images. However, these approaches often require significant computation time due to the volume rendering process and iterative optimization algorithms. In contrast, methods such as IMavatar [6] reconstruct a 3D face from a 2D video but suffers from temporal instability, where frames lack smooth transitions over time. Both approaches depend on multiple captures across space or time, limiting their practicality for self-portrait scenarios, where a single-shot, real-time solution is desired.

Recently, we introduced LP3D [7], a single-image-based 3D face reconstruction algorithm that leverages the strengths of EG3D [8], a triplane-based generative 3D representation. While traditional methods like Nerfies and CAFCA rely on multi-view inputs and suffer from high computational costs, LP3D efficiently uplifts a single 2D image to 3D in real time by encoding the facial structure into a lightweight triplane representation. This allows the generated NeRF to be rendered rapidly and displayed on a light field display, enabling applications such as 3D video conferencing [9].

In this work, we present an AI 3D Selfie system that enables users to capture their face in 2D and view it in 3D in real time. Building on the strengths of LP3D, our system incorporates the WYSIWYG [10] model, which enhances training efficiency and eliminates the need for upsampling, resulting in a natively high-resolution triplane NeRF. The reconstructed light field is then rapidly rendered and displayed on a light field display, offering an efficient and high-quality solution for real-time 3D self-portraits and virtual communication.

2. AI 3D selfie

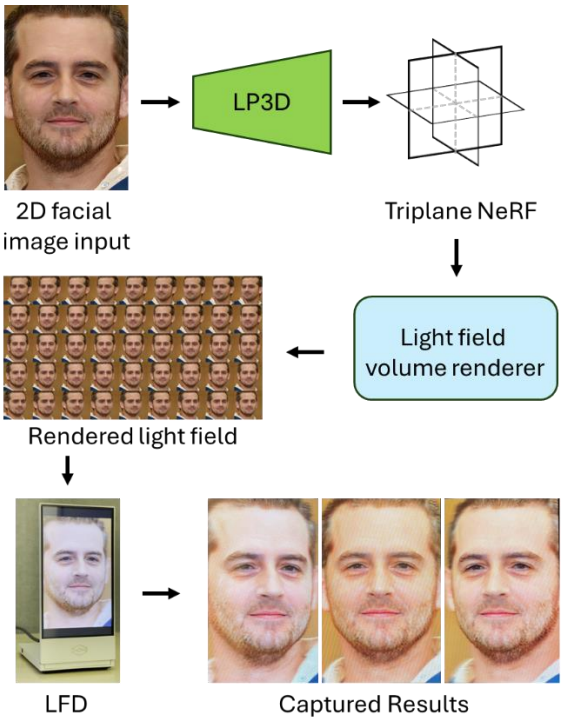

Figure 1. Overview of the AI 3D Selfie system pipeline. A 2D facial image is captured and processed using LP3D to generate a triplane-based NeRF. The NeRF is rendered into a light field and displayed on a light field display, enabling real-time 3D viewing from multiple perspectives.

Figure 1 shows the AI 3D Selfie system pipeline. An unposed 2D facial image is captured using a front-facing camera and processed with the LP3D model, which uplifts the 2D input into a triplane-based NeRF representation that conforms to the latent space of the WYSIWYG model [10]. The resulting NeRF is then rendered into a high-resolution light field using a fast volume rendering algorithm. Finally, the reconstructed 3D face is displayed on a light field display, enabling realistic 3D viewing from multiple perspectives in real time.

3. Light Field Volume Rendering

Figure 2 illustrates our fast light field volume rendering algorithm for NeRF. The neural representation $F_\theta: (x, d) \rightarrow (c, \sigma)$ with a bounding box aims to render a light field quilt from the z+ direction. The evenly spaced light field camera array is oriented towards the $(z = 0)$ plane, as shown on the left of Fig. 2. The cameras are positioned far enough from the bounding box for orthographic projection.

ISSN 0097-996X/25/5602-1133-$1.00 © 2025

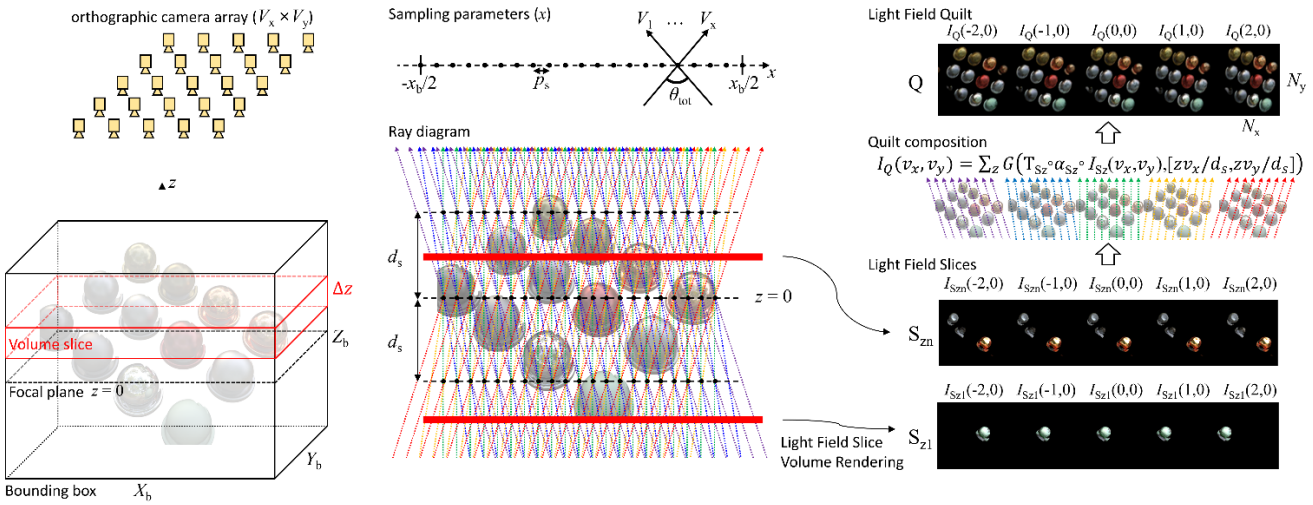

Figure 2. *Left*: Diagram for Light field rendering through a NeRF with a 3D bounding box. An orthographic camera array located near the z^+ axis is focusing on the $z = 0$ plane. *Center*: Sampling parameters and ray diagram of light field rendering. The orthographic projection creates repeated sampling planes with (d_s) interval. *Right*: Composition of a light field quilt from light field slices, where each slice is defined as a light field quilt sampled within a unit volume (a single depth slice). Each slice is composited using the image shift function *G* based on its viewing direction (v_x, v_y).

The light field quilt consists of stitched view images in a $V_x \times V_y$ format, each with resolution $N_x \times N_y$. $V_x \times V_y$ are the numbers of cameras along the x and y dimensions. The upper center of Fig. 2 shows that the sampling parameters of the light field volume rendering. These quilt parameters $[N_x, N_y, V_x, V_y, \theta_{tot,x}, \theta_{tot,y}]$ are adjustable, defining a unique light field quilt for the 3D scene. A light field quilt Q can be expressed as:

$$Q = Concat_{\forall v_x, v_y}\{I_Q(v_x, v_y)\} \quad (1)$$

where v_x and v_y range from $-(V_x - 1)/2$ to $(V_x - 1)/2$ and $-(V_y - 1)/2$ to $(V_y - 1)/2$, respectively. These indices represent the view numbers within the light field quilt, and $I_Q(v_x, v_y)$ denotes the rendered orthographic view image from the direction specified by $[v_x, v_y]$.

The lower center of Fig. 2 illustrates the ray diagram for light field rendering and its repeatability. Unlike conventional perspective camera projections used in novel-view synthesis for NeRF, light field quilt rendering employs an array of parallel light rays from each orthographic camera. This arrangement implies that each set of $V_x \times V_y$ light rays diverges identically from each sampling point. The orthographic projection from the uniformly spaced sampling points results in repeated sampling planes where *all* light rays converge at one of the sampling points, as shown by solid black dots in the diagram. The repeated sampling planes can be expressed as $z = n \times d_s$, where *n* is an integer. Here, d_s is defined as:

$$d_s = p_s \times \frac{V-1}{2\tan(\theta_{tot}/2)}. \quad (2)$$

This definition is generalized for both the *x* and *y* directions, applicable to either $(V_x, \theta_{tot,x})$ or $(V_y, \theta_{tot,y})$. d_s represents the thickness of these repeated sampling planes, which are parallel to the *xy*-plane when $(d_{sx} = d_{sy} = d_s)$. At these planes, rays always converge at one of the sampling points. The lateral shift of each ray, increasing with its off-axis angle, is determined by its direction, affecting its position at the repeated sampling planes.

We define a light field slice S_z, which is a segment of a light field quilt sampled within a unit volume (from \(z\) to \(z + \Delta z\)). The light field slice S_z can be expressed as follows:

$$S_z = Concat_{\forall v_x, v_y}\{I_{Sz}(v_x, v_y)\}. \quad (3)$$

$I_{Sz}(v_x, v_y)$ denotes the orthographic view image from the direction specified by $[v_x, v_y]$, rendered within the unit volume.

By stacking N_z light field slices, which cover the range from $z = -Z_b/2$ to $z = Z_b/2$, and accounting for the lateral shift between slices, we render $I_Q(v_x, v_y)$ as follows:

$$I_Q(v_x, v_y) = \sum_{z=-N_z/2}^{N_z/2} G(T_z \circ \alpha_z \circ I_{Sz}(v_x, v_y), [v_x z/d_s, v_y z/d_s]), \quad (4)$$

where G is the 2D shift function that accounts for pixel shift over slices, $T_z(v_x, v_y)$ represents the cumulative transmittance, $\alpha_z(v_x, v_y)$ denotes the opacity at each slice, and ∘ denotes element-wise multiplication. Note that N_z is also identical to the number of sampling points per ray. This formulation allows for the acquisition of view images in all directions within the light field quilt through a single-pass plane sweeping process.

4. Reconstruction on Light Field Display

A light field quilt can serve as a base image for various types of 3D displays, including multi-view displays, integral imaging, computational light field displays, and holographic displays [11]. This is particularly advantageous for lens-based light field displays such as lenticular displays or integral imaging, where the symmetry between rendering and optics allows for real-time generation of the elemental images [12,13].

Figure 3 illustrates how the elemental images for a lens array-based light field display is generated from a light field quilt. The light rays emitted from each pixel is refracted and redirected by each lens in the lens array. When the lens array plane is aligned with the light field sampling plane ($z = 0$), the rays from each pixel can be directly correlated with a corresponding pixel on the

light field quilt. This process is typically achieved through interlacing using manufacturer calibration data, allowing for real-time rendering.

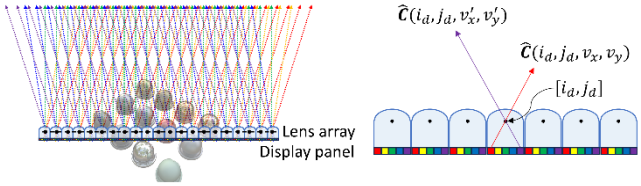

Figure 3. *Left*: Ray diagram for 3D reconstruction using a light field display. *Right*: When the lens array plane is aligned with the sampling plane ($z = 0$), it becomes possible to map each subpixel to the nearest corresponding pixel on the light field quilt for accurate supervision.

5. References

1. M. Askari and J.-H. Park. Pre-compensation of an image blur in holographic projection display using light emitting diode light source. Optics Express, 28(1):146–159, 2020.
2. Park, Keunhong, et al. "Nerfies: Deformable neural radiance fields." *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021.
3. Kirschstein, Tobias, et al. "Nersemble: Multi-view radiance field reconstruction of human heads." *ACM Transactions on Graphics (TOG)* 42.4 (2023): 1-14.
4. Buehler, Marcel C., et al. "Cafca: High-quality Novel View Synthesis of Expressive Faces from Casual Few-shot Captures." *SIGGRAPH Asia 2024 Conference Papers*. 2024.
5. Mildenhall, Ben, et al. "Nerf: Representing scenes as neural radiance fields for view synthesis." *Communications of the ACM* 65.1 (2021): 99-106.
6. Zheng, Yufeng, et al. "Im avatar: Implicit morphable head avatars from videos." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022.
7. Trevithick, Alex, et al. "Real-time radiance fields for single-image portrait view synthesis." *ACM Transactions on Graphics (TOG)* 42.4 (2023): 1-15.
8. Chan, Eric R., et al. "Efficient geometry-aware 3d generative adversarial networks." *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022.
9. Stengel, Michael, et al. "AI-mediated 3D video conferencing." *ACM SIGGRAPH 2023 Emerging Technologies*. 2023. 1-2.
10. Trevithick, Alex, et al. "What You See is What You GAN: Rendering Every Pixel for High-Fidelity Geometry in 3D GANs." *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024.
11. Hong, Jisoo, et al. "Three-dimensional display technologies of recent interest: principles, status, and issues." *Applied optics* 50.34 (2011): H87-H115.
12. Jung, Jae-Hyun, Jonghyun Kim, and Byoungho Lee. "Solution of pseudoscopic problem in integral imaging for real-time processing." *Optics Letters* 38.1 (2012): 76-78.
13. Kim, Jonghyun, et al. "Real-time capturing and 3D visualization method based on integral imaging." *Optics Express* 21.16 (2013): 18742-18753.