FieldFusion: Harmonious Radiance Fields Composition

Ziyang Xie
ziyang8@illinois.edu

Baoyu Li
baoyul2@illinois.edu

1. Motivation and Impact

In recent years, significant progress has been made in 3D object and scene reconstruction, notably with the introduction of radiance fields reconstruction methods like NeRF [1, 2, 6, 7] (Neural Radiance Fields) and 3D Gaussian splatting [5]. These methods have paved the way for the creation of intricate and detailed reconstructions, laying a robust foundation for the development of realistic simulation environments. Such photorealistic simulation environments are of great importance, especially in domains like robotics and autonomous driving, where they play a significant role in testing, training, and validating algorithms in a controlled yet realistic setting.

However, a key challenge persists in the seamless integration of reconstructed foreground objects with background scenes for a fully functional simulation environment. To address this, our project introduces FieldsFusion, a systematic pipeline designed to bridge this gap. FieldsFusion facilitates the seamless blending of foreground and background elements, rendering photorealistic images that maintain geometric and lighting consistency. This approach ensures both visual fidelity and functional utility, enhancing the capabilities of advanced simulation applications. Our code can be found at https://github.com/ZiYang-xie/FieldFusion

2. Approach

2.1. Radiance Fields Preliminary

NeRF [6] and 3D Gaussian-Splatting [5] stand out as two popular methods in the field of 3D reconstruction. Neural Radiance Fields (NeRF) utilizes volume rendering to derive the color C for every position within the scene. In contrast, Gaussian-Splatting takes a slightly different approach, creating 3D Gaussian splats and then projecting and blending them using Gaussian-based α-blending techniques.

Both approaches, despite their differences, are based on a similar foundational image model, where the color C along a given camera ray is determined by the equation:

$$C = \sum_{i=1}^{N} T_i \alpha_i c_i, \quad \textit{where } T_i = \sum_{j=1}^{i-1} (1 - \alpha_j) \qquad (1)$$

The T_i represent the accumulative transmittance which are derived from α_i. In the NeRF framework, the term α_i is calculated as $1 - exp(\sigma_i \delta_i)$ by sample the density σ_i on the ray with intervals δ_i

Conversely, Gaussian-Splatting [5] involves projecting 3D Gaussian onto the 2D image plane. The pixel color C is obtained by α-blending N sequentially layered 2D Gaussians from front to back. Here, the value of α is derived by multiplying the opacity o with the contribution of the 2D covariance relative to the pixel's coordinates.

2.2. 3D Reconstruction

In this project, we collected a diverse dataset featuring various object types and lighting conditions across the UIUC campus. We then separately reconstructed the foreground objects and background scenes for further composition. We provide qualitative result in Fig. 4 and quantitative result in Table. 1

Background Reconstruction We employ 3D Gaussian-Splatting [5] for reconstructing the background scene. This method outperforms the NeRF approach by offering faster reconstruction speeds and superior quality in the background reconstruction.

Foreground Reconstruction In reconstructing the foreground objects, we experimented with both the NeRF method (NeRFacto [9]) and 3D Gaussian-Splatting [5].

While Gaussian-Splatting [5] offers better reconstruction quality compared to NeRF [9] (Table. 1), it faces difficulties in precisely cropping foreground objects from their background, especially in the contact region. Chunks of 3D Gaussians often leads to noise in the composition stage. As a result, NeRF that densely reconstruct the density for each point within the scene proves to be a more suitable choice. Also, the quality of NeRF's reconstruction is sufficiently good for our purposes, leading us to select NeRF for our foreground reconstruction needs.

2.3. 3D Consistent Composition

To effectively integrate the foreground objects with the background scene, we need to address two primary challenges: **1. Geometry Alignment**: For a realistic composition, it's crucial that the foreground objects are positioned accurately within the background scenes. This means avoiding geometry errors such as object penetration

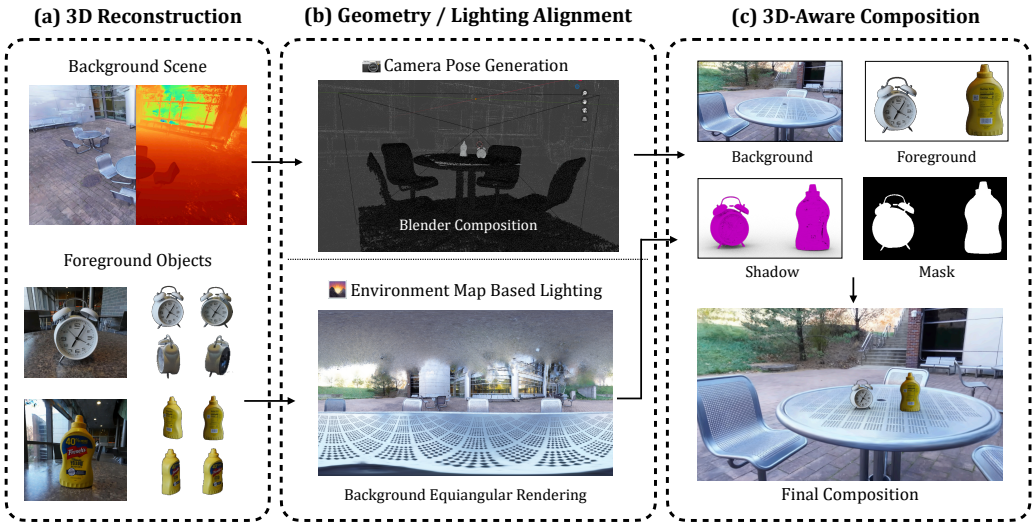

Figure 1. **FieldsFusion Pipeline:** Our method can be spilt into three parts. (a) 3D Reconstruction: We first reconstruct the foreground and background scene separately with 3D Gaussian-Splatting [5] and NeRFacto [9], (b) Then we use Blender [3] to render shadow and generate 3D consistent camera pose for both the foreground objects and background scenes. After that, (c) we render the 3D consistent view from the generated pose and compose them together with 2D blending to get the final composition.

or unrealistic placements. **2. Lighting Consistency**: Since the foreground objects and background scenes are often captured under different lighting conditions, harmonizing the lighting is essential.

This involves ensuring that the foreground object matches the background's lighting and casts appropriate shadows according to the background's light source, which is vital for achieving realistic composition.

To overcome these challenges, we propose the use of Blender [3] as an intermediary tool. Blender facilitates the generation of 3D-consistent camera poses, enabling aligned 2D rendering for both the foreground and background. Additionally, for lighting consistency, Blender's advanced environment allows for effective rendering of shadows using environment maps extracted from our background scene reconstructions.

Camera Pose Generation In order to resolve the issue of geometry alignment, we utilize Blender for generating 3D camera poses that are consistent across different views. This process involves importing the geometric priors, such as point clouds or mesh data, of both the foreground objects and background scene into Blender. These priors are obtained from the reconstruction by exporting pointclouds through density query. Once imported, we accurately position the foreground object within the background scene to ensure seamless integration and alignment.

Additionally, we design a camera trajectory manually for rendering purposes. he camera poses provided by Blender are then converted into relative camera poses for both the foreground and background elements. These poses are subsequently used to render the final RGB images for each element, ensuring that they are accurately aligned and integrated within the scene.

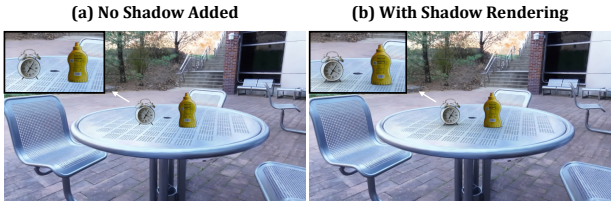

Figure 2. **Shadow Comparison:** (a) Direct composition without shadow. (b) Our Composition with our shadow rendering. Our shadow rendering technique significantly enhances the quality of the composition.

Shadow Rendering Shadows play a crucial role in enabling humans to discern whether an image is fake or not (Fig. 2). In image composition, shadows are essential for creating a sense of depth and spatial coherence. Without properly rendered shadows, these compositions often appear flat and artificial.

In our pipeline, we utilized geometric priors we exported in Sec. 2.3 and employed Blender to render the shadow through ray tracing. We further add an extra plane aligned with the background surface to enable accurate shadow casting from the foreground mesh.

The lighting source in our scene is automatically generated using the environment map, which is rendered with the background scene reconstruction. (Fig. 1 (b) bottom) This method ensures that the lighting dynamically matches the reconstructed background, enhancing the realism of the image. Ultimately, our shadow rendering technique allows the foreground to integrate seamlessly with the background,

creating realistic shadows that add authenticity to the scene.

Inverse Rendering with Relighting To keep lighting consistency between the foreground objects and background scene, we implemented Tensorial Inverse Rendering [4] for relighting. However, since this approach requires object images with different light conditions for training, which is hard to obtain for self-collected data, we only performed relighting for the objects from the TensoIR-Synthetic dataset, such as the hot dog, and combine it with our self-collected UIUC background.

Our environment maps are rendered from background NeRF reconstructions as shown in Fig. 1, we further use an off-the-shelf pretrained model [8] to reconstruct the High Dynamic Range (HDR) environment map from its Low Dynamic Range (LDR) rendering.

With the TensoIR [4], we can decompose the foreground object's albedo and specular components as well as the surface normal. We can further relight the foreground objects with BRDF-based rendering using the reconstructed HDR environment map to ensure seamless composition.

3D-Aware Composition Given the generated camera poses, we can separately render the background scene and the foreground object in Blender. As we have already established their relative 3D positions, the 2D images can be seamlessly composited to achieve 3D-consistency. For the foreground part, we manually assign a 3D bounding box to crop out the original background. We further using ray accumulation to render mask $M = \sum_{i=1}^{N} T_i \alpha_i$ for the foreground object.

Given the foreground image I_f, background image I_b, foreground mask M and shadow image I_s. Final composed image can be represent as a α-blending composition:

$$I_c = I_f \times M + (I_b + I_s) \times (1 - M) \tag{2}$$

3. Results

In this section, we show the qualitative and quantitative result of our FieldsFusion composition.

3.1. Reconstruction Results

In Table 1, we compared the Gaussian-Splatting [5] with the NeRFacto [9] methods on foreground reconstruction quality. We selected 10% of the captured images as the evaluation set while the remaining were used for training.

We segment out the foreground object and only evaluate on the foreground region instead of comparing the entire image with ground-truth with its background. This approach ensures a more precise evaluation of the foreground reconstruction.

For qualitative evaluation, We provide both the background and foreground rendering results in Fig. 4.

Method	PSNR ↑	SSIM ↑	LPIPS ↓
NeRFacto [9]	31.12	0.97	0.028
Gaussian-Splatting [5]	**39.04**	**0.99**	**0.012**

Table 1. **Foreground Reconstruction Results:** We assessed the quality of foreground reconstruction using two real-world objects we captured: a clock and a mustard bottle. In this assessment, Gaussian-Splatting demonstrated superior performance compared to the NeRFacto method. Nevertheless, the NeRF-based method is also good enough for our downstream composition ($>$ 30 PSNR)

3.2. Composition Results

Fig. 3 shows the composition results generated by FieldsFusion. It demonstrates that our pipeline can naturally compose the foreground with background fields together to generate realistic merging results. We also show our relighting results to illustrate how can we handling complex lighting conditions through inverse rendering and BRDF relighting.

3.3. Object Relighting

Relighting results are shown in Fig. 3(d). We use TensoIR [4] to relight the foreground objects with our reconstructed environment map and then implement the same pipeline described in Sec. 2 for composition.

Since we did not collect the direct light source (e.g. sky or sun) above the scene, it is hard to use our environment map to capture the real lighting conditions, resulting in dark relighting results. We claim that this issue could be further resolved by extensively collecting the whole scene, including all the lighting sources, using HDR images.

4. Challenge and Innovation:

4.1. Challenges Faced:

Technical Complexity: The integration of foreground objects with background scenes reconstructed from radiance fields posed a complex challenge. Ensuring a seamless and harmonious composition between these elements was not straightforward, given the inherent technical intricacies of radiance fields methods.

Custom Data Collection and Evaluation: A significant challenge and contribution of our project was collecting our own dataset around UIUC. This complex task required selecting scenes, setting up equipment, and capturing multi-view images in various environmental conditions, often leading to inadequate reconstructions and necessitating repeated captures and adjustments.

Implementation and Adaptation: Adapting existing methods to our specific problem statement was a challenge. We had to interpret unclear steps from existing papers and modify them to fit our project requirements, which involved significant research and experimentation.

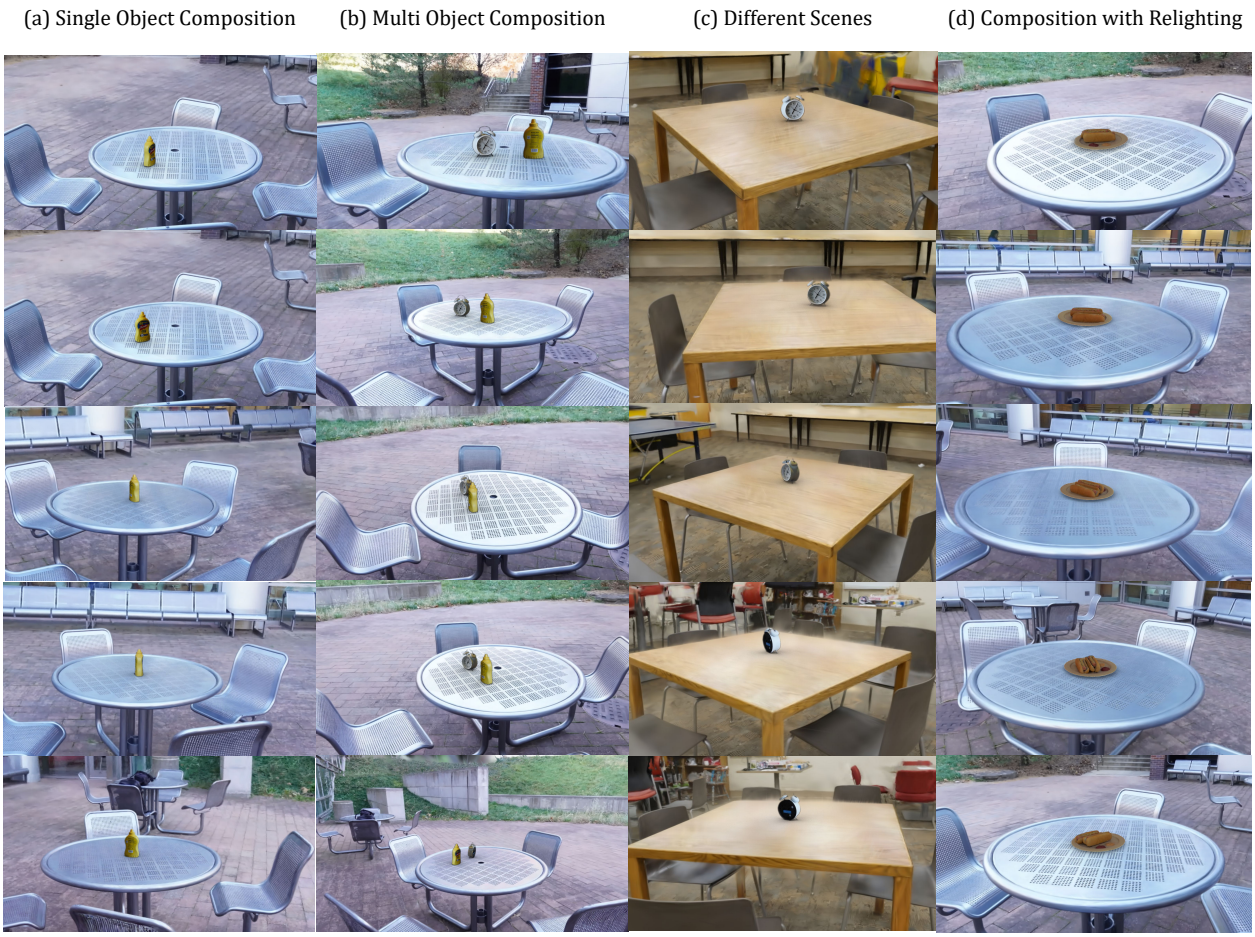

Figure 3. **FieldsFusion Composition Results:** (We provide video link in Sec. 5 for *high resolution results* and *animations*). We showcase several types of composition outcomes: (a) Single Object Composition, where a single object is integrated into a scene, and (b) Multi Objects Composition, which features multiple objects, including scenarios with inter-occlusions among them. Additionally, in (c), we demonstrate the adaptability of our method by compositing our foreground object into a different indoor environment. Finally, in (d), we display our relighting results to illustrate how to effectively handle varying lighting conditions.

4.2. Innovative Aspects:

Our project stands out in its innovative approach to harmoniously blend foreground and background elements in a 3D space. We propose to use Blender [3] as an intermediary tool. This method facilitates the generation of 3D-consistent camera poses, enabling aligned 2D rendering for both the foreground and background. This innovation is particularly relevant for simulation systems in robotics and autonomous driving.

4.3. Justification for Points:

Given the complexity and novelty of our project, we believe a high score in the challenge/innovation component is justified. Our project not only involved technical challenges but also introduced innovative approaches in data collection, processing, and the harmonious composition of radiance fields. This work goes beyond typical projects in its scope and contribution to the field, justifying a higher score for its challenge and innovation aspects.

5. Conclusion

In conclusion, the FieldsFusion is a step forward in the domain of realistic scene composition, offering a robust framework for integrating 3D reconstructed foreground objects with background scenes.

Through the use of advanced techniques such as 3D Gaussian-Splatting, NeRF-based reconstructions, and Blender for precise camera pose generation and lighting harmonization, we have developed a method that enhances the realism of composited images.

The video demo of our composition results can be found at https://youtu.be/Tfl3xf3OHJw?si=vYdQ01zu_zOo5Ep0.

Figure 4. **Reconstruction Results:** Self-collected foreground objects and background scenes.

References

[1] Jonathan T. Barron, Ben Mildenhall, Matthew Tancik, Peter Hedman, Ricardo Martin-Brualla, and Pratul P. Srinivasan. Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. *ICCV*, 2021. 1

[2] Jonathan T. Barron, Ben Mildenhall, Dor Verbin, Pratul P. Srinivasan, and Peter Hedman. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. *CVPR*, 2022. 1

[3] Blender Online Community. *Blender - a 3D modelling and rendering package*. Blender Foundation, 2018. 2, 4

[4] Haian Jin, Isabella Liu, Peijia Xu, Xiaoshuai Zhang, Songfang Han, Sai Bi, Xiaowei Zhou, Zexiang Xu, and Hao Su. Tensoir: Tensorial inverse rendering. In *CVPR*, 2023. 3, 5

[5] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 2023. 1, 2, 3

[6] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020. 1

[7] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Trans. Graph.*, 2022. 1

[8] SMA Sharif, Rizwan Ali Naqvi, Mithun Biswas, and Sungjun Kim. A two-stage deep network for high dynamic range image reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 550–559, 2021. 3

[9] Matthew Tancik, Ethan Weber, Evonne Ng, Ruilong Li, Brent Yi, Justin Kerr, Terrance Wang, Alexander Kristoffersen, Jake Austin, Kamyar Salahi, Abhik Ahuja, David McAllister, and Angjoo Kanazawa. Nerfstudio: A modular framework for neural radiance field development. In *ACM SIGGRAPH 2023 Conference Proceedings*, 2023. 1, 2, 3

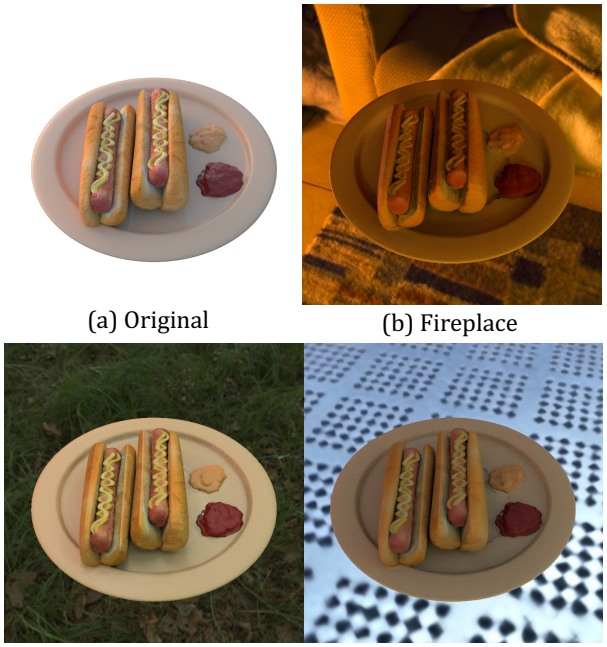

Figure 5. **Relighting Comparison:** (a)-(c) are original and relighting results from TensoIR [4] and (d) is our relighting rendering with reconstructed environment map.

Contribution:

Group members contributed similarly

- **Ziyang Xie**: Data Collection, Pipeline Design, Composition code writing, Paper writing
- **Baoyu Li**: Data Collection, Relighting code writing, Paper writing