Study on reducing the impact of irrelevant Objects in 3D Reconstruction Using NeRF

Qihui Fan
Student, MS Data Science
fan.qih@northeastern.edu

Zuyu Guo
Student, MS Computer Science
guo.zuy@northeastern.edu

Yiqian Li
Student, MS Computer Science
li.yiqian@northeastern.edu

Abstract

3D reconstruction from 2D images has long been a central topic in the field of computer vision, with applications spanning industries such as medical research, mapping, and the rapidly growing areas of virtual reality **(VR)** and augmented reality **(AR)**. Among the state-of-the-art methods, **Neural Radiance Fields (NeRF)** has emerged as a powerful framework for generating high-quality 3D reconstructions.[6] However, NeRF's reliance on large amounts of high-quality training data makes it particularly sensitive to irrelevant or dynamic objects in the input images, which can degrade the quality of the reconstructed views.[5] To address this challenge, this project proposes a preprocessing pipeline to enhance NeRF's performance by leveraging advanced segmentation and inpainting techniques. Specifically, models like **MaskFormer**[3] and **Latent Diffusion Model**[7] are employed to remove and restore regions with irrelevant objects in the training data, thereby providing NeRF with cleaner inputs. Comprehensive experiments are conducted to evaluate the robustness of NeRF under real-world scenarios, with a focus on improving its ability to handle complex, dynamic outdoor environments. The results demonstrate the effectiveness of the proposed approach in optimizing scene reconstruction quality.

Keywords: 3D reconstruction, NeRF, MaskFormer, Image Segmentation, Scene Optimization, Latent Diffusion

1 Introduction

3D reconstruction is the process of creating a 3D model from 2D images or data, with Neural Radiance Field (NeRF) standing out as a state-of-the-art algorithm in this domain. Despite its advancements, NeRF often struggles when the training images include distracted or dynamic elements, such as tourists around a landmark, which detract from the target subject.[2] We propose a novel pipeline that enhances NeRF's robustness by integrating advanced computer vision techniques. The pipeline employs image segmentation to accurately identify and mask unwanted objects, followed by inpainting using stable diffusion models to seamlessly fill the masked areas with contextually appropriate content. By preprocessing the input data in this way, our approach minimizes distractions, allowing NeRF to achieve higher-quality reconstructions even in complex scenarios. This method demonstrates improved flexibility and reliability, extending NeRF's applicability to more complex and cluttered real-world datasets.

Preprint. Under review.

2 Related Works

2.1 Neural Radiance Field (NeRF)

Neural Radiance Fields (NeRF) is a groundbreaking neural network method that optimizes radiance fields in 3D space to generate high-quality 3D reconstructions.[6] While NeRF performs exceptionally well in static scenes, its performance significantly deteriorates in dynamic datasets. This is because dynamic scenes, with moving objects or changing environments, introduce inconsistencies in the data, negatively impacting model training and rendering. In outdoor scenes, in particular, dynamic elements such as pedestrians and vehicles often result in incomplete or distorted 3D reconstructions. To address this issue, we propose an innovative preprocessing pipeline that leverages advanced segmentation and inpainting techniques to eliminate the impact of dynamic elements on NeRF, thereby significantly improving its performance in dynamic datasets.

2.2 NeRF-HuGS: Enhancing NeRF for Transient Object Handling in Non-Static Scenes

NeRF-HuGS, as one of the most recent studies related studies attempting to addresses the limitations of traditional Neural Radiance Fields (NeRF) in handling transient distractors in non-static scenes by integrating heuristic-based methods and state-of-the-art segmentation models like the Segment Anything Model (SAM).[2]. It combines Structure-from-Motion (SfM)-based heuristics for identifying high-frequency static details and residual-based heuristics from partially trained NeRF models for low-frequency textures, enabling robust static-transient object separation. While NeRF-HuGS achieves superior performance on datasets like Kubric, Distractor, and Phototourism, excelling in metrics such as PSNR, SSIM, and LPIPS,[2] it still encounters certain limitations. These include sensitivity to heuristic thresholds, scene dependence on SfM and SAM accuracy, and computational overhead, making it less efficient for large-scale or real-time applications. Despite these challenges, NeRF-HuGS provides a significant step forward in improving view synthesis and 3D reconstruction in real-world scenarios with transient distractors.

Similarly, we address similar limitations of NeRF using a different approach by utilizing the advancement of cutting-edge computer vision techniques.

2.3 Masked-attention Mask Transformer (MaskFormer)

MaskFormer, initially proposed by Facebook AI Research (FAIR), is one of the state-of-the-art segmentation models, with its core innovation being the introduction of a Transformer-based attention mechanism.[3] This mechanism captures the global context of an image, making it particularly effective in understanding complex scenes. Compared to traditional RCNN methods, MaskFormer excels at generating accurate segmentation results in scenes with significant fo reground occlusions.[3] Moreover, its capability for Panoptic Segmentation, which combines semantic segmentation and instance segmentation, is crucial for our approach. Outdoor environments typically consist of diverse and complex elements, including both static structures (like buildings and trees) and dynamic objects (like vehicles and pedestrians).[3] Accurately separating and identifying these elements is essential for achieving a clear understanding of the scene. Panoptic Segmentation allows the model to provide a holistic view of the scene, enabling us to distinguish between background elements and individual foreground objects, even when they overlap or occlude one another. This comprehensive understanding ensures that our pipeline can preprocess destructed outdoor scenes effectively, providing NeRF with clean and well-structured input data for accurate 3D reconstruction. We utilized Hugging Face's open-source Mask2Former, based on MaskFormer, to perform semantic image segmentation and loaded the pretrained Cityscape weights to segment out content unrelated to the static scene.

2.4 Latent Diffusion Model (LDM)

Latent Diffusion Model is a highly flexible generative model that excels at producing high-quality inpainting results.[7] By integrating with the segmentation outputs from MaskFormer, Latent Diffusion can intelligently fill in missing or dynamic regions based on provided prompts. Compared to traditional inpainting methods, Latent Diffusion not only delivers superior quality but is also more seamlessly integrated into our processing pipeline. This approach effectively handles dynamic scenes and generates coherent and stable input data for subsequent NeRF training. In this research, we utilized the pretrained Latent Diffusion model from Hugging Face, namely *Stable Diffusion 2 Inpainting*, which is pretrained on LAION-5B and capable of handling comprehensive and complex scenes[1], as its training dataset includes a large number of realistic scenes and photos, making it highly suitable for our study.

In this work, we integrate MaskFormer and Latent Diffusion into a preprocessing pipeline designed to address the challenges of dynamic datasets in NeRF. By combining state-of-the-art segmentation and inpainting techniques, our approach aims to deliver cleaner, more consistent input data, enabling NeRF to achieve superior reconstruction quality even in dynamic outdoor environments.

3 Methods

3.1 Overview

In this section, we present our proposed preprocessing pipeline designed to enhance NeRF's performance on dynamic datasets. The pipeline consists of three stages: (1) segmentation of dynamic elements using MaskFormer, (2) inpainting of segmented regions using Latent Diffusion, and (3) training NeRF on the preprocessed, static-like datasets. This framework ensures the input data's consistency and improves the 3D reconstruction quality.

3.2 Semantic Segmentation Module

To precisely illustrate our experiment, it is necessary to first define and recognize the distracted objects in our experimental setting. In our attempt, we *only* define "random people" as the focused distracted object, and use one of the most state-of-the-art image segmentation models, namely MaskFormer. Hugging Face open-source, MaskFormer-based Mask2Former has been utilized to perform semantic image segmentation, and pretrained *Cityscape* weights have been loaded for its outstanding segmentation performance on our subject matter.

In the pretrained model, we only focus on the **building** class segmentation among all 34 pre-trained classes and its reverse "non-building" class, making it a binary classification and segmentation task. This approach fulfills our experiment on reconstructing building-like objects. To ensure more flexible and accurate segmentation, we set a confidence threshold θ of 0.85 for segmentation proposals.

For the segmentation modules, we consider consistent training *images* I with a width of W and a height of H. An intermediate Numpy mask matrix M is generated with dimensions $W \times H$, using binary coding for each pixel. The matrix will be utilized in future inpainting module and visual inspection. In the matrix, 1 represents "building" components (regions to be preserved), and 0 represents the "non-building" class (e.g., regions with people, which should be masked). Formally, the segmentation process can be expressed as:

$$M(x,y) = \begin{cases} 1, & \text{if } \text{logit}(I(x,y)) > \theta, \quad \text{where } \theta = 0.85. \\ 0, & \text{otherwise} \end{cases}$$

Here, $M(x,y)$ is the binary segmentation mask for each pixel (x,y) in the image $I(x,y)$, where:

$$x \in \{1,2,\dots,W\}, \quad y \in \{1,2,\dots,H\}.$$

3.3 Inpainting Module

After the segmentation module, we have the following training data:1. *Images*, $I(x,y)$, 2. *Segmentation masks*, $M(x,y)$.

In this module, we first load the *Latent Diffusion Model (LDM)* from Hugging Face. We utilize its pretrained model for the image inpainting task—removing objects—using the specific model, namely *Stable Diffusion 2 Inpainting*, capable of handling comprehensive and complex scenes. To improve the quality of the generated images, we employ prompt tuning techniques, adjusting the prompts appropriately for each scene to produce more accurate results.[4] However, the LDM requires square-sized image shapes. If the input image is not square, it will be stretched during the inpainting process, which can distort the results.

This requirement breaks the dependencies of the NeRF, therefore we pad the image to a square size. The square size is determined as:

$$H^* = W^* = \max(H, W)$$

The padded image $I^*(x^*, y^*)$ and padded mask $M^*(x^*, y^*)$ are constructed as follows:

$$I^*(x^*, y^*) = \begin{cases} I(x, y), & \text{if } (x^*, y^*) \in \text{original image bounds,} \\ 0, & \text{otherwise (padded area).} \end{cases}$$

$$M^*(x^*, y^*) = \begin{cases} M(x, y), & \text{if } (x^*, y^*) \in \text{original image bounds,} \\ 1, & \text{otherwise (padded area to avoid inpainting).} \end{cases}$$

This ensures that the padded area is not inpainted, maintaining the integrity of the image.

The padded image I^* and the mask M^* are then passed to the LDM to obtain the inpainted image I_{inpaint}. The inpainting process is defined as:

$$I_{\text{inpaint}}(x^*, y^*) = \begin{cases} \text{LDM}(I^*(x^*, y^*)), & \text{if } M^*(x^*, y^*) = 0, \\ I^*(x^*, y^*), & \text{if } M^*(x^*, y^*) = 1. \end{cases}$$

Finally, the image is cropped back to its original size to ensure integrity, defined as:

$$I_{inpaint}(x, y) = I_{\text{inpaint}} \left[x_{\text{center}} - \frac{W}{2} : x_{\text{center}} + \frac{W}{2},\ y_{\text{center}} - \frac{H}{2} : y_{\text{center}} + \frac{H}{2} \right]$$

where x_{center} and y_{center} are the coordinates of the center of the padded image.
We would mannually select those images with relatively smooth and matching inpainting, reducing possible negative imapcts in the experiment.

3.4 3D Reconstruction Module

After processing the inpainting module, we obtain the inpainted image $I_{\text{inpaint}}(x^*, y^*)$, where (x^*, y^*) represent the valid coordinates of the inpainted region. Using the spatial information provided by the image dataset, we reconstruct the 3D scene via NeRF.

The input to the NeRF model includes the inpainted images I_{inpaint} along with the corresponding camera parameters, including:

- **Camera Intrinsics:** The focal length and principal point coordinates (f_x, f_y, c_x, c_y).
- **Camera Extrinsics:** The rotation matrix R and translation vector t for each image.

NeRF utilizes these parameters to trace rays through each pixel, sampling along the ray to estimate the volumetric radiance field. The reconstructed radiance field is then optimized to minimize the discrepancy between the predicted and ground truth pixel intensities, using the following loss function:

$$\mathcal{L} = \frac{1}{N} \sum_{i=1}^{N} \left\| \hat{C}_i - C_i \right\|^2,$$

where:

- $\hat{C}_i$: The predicted color of ray i,
- C_i: The ground truth color of ray i,
- N: The total number of rays sampled.

The final 3D reconstruction is generated by aggregating the optimized radiance field across all input images. To ensure robustness, the model training excludes images with residual artifacts or mismatched inpainting results. These exclusions help minimize the potential negative impacts on reconstruction quality and improve the overall integrity of the 3D scene.

The reconstructed 3D model is evaluated qualitatively and quantitatively using metrics such as PSNR, providing a comprehensive assessment of the effectiveness of the inpainting module and the NeRF pipeline.

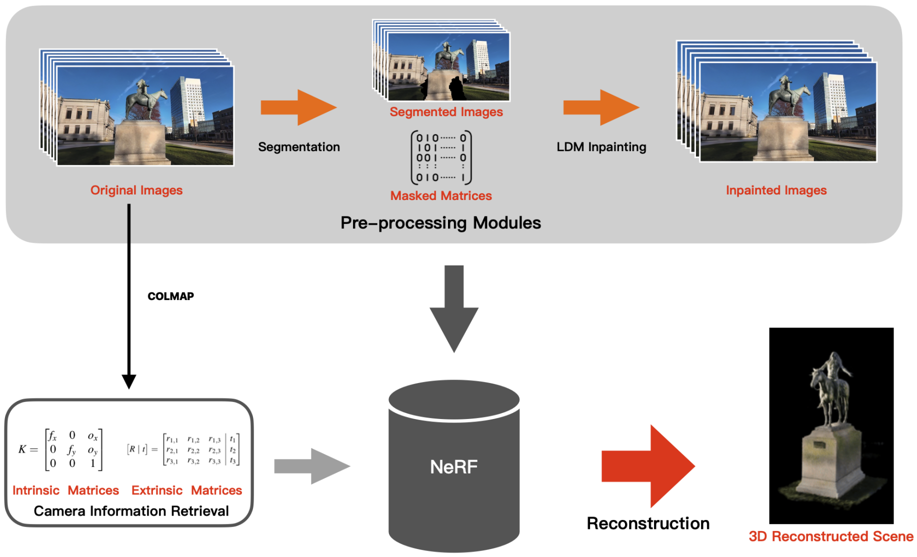

Figure 1: Our method Pipeline

4 Experiments

4.1 Experiment Setup

We use two real-world self-made datasets to evaluate our model. First, we compare the semantic segmentation mask combined with NeRF against the baseline NeRF on two datasets, using quantitative and qualitative evaluations to examine the effectiveness of target individual segmentation. Then, we integrate the results after diffusion model and further perform quantitative and qualitative analyses to evaluate how our complete model extends and improves upon the baseline NeRF reconstruction.
Datasets.

(a) Horse Statue

(b) Fine-arts Statue

Figure 2: Visual Samples from the Horse Statue and Fine Arts Statue Datasets for Reconstruction Evaluation.

Horse Statue. This dataset captures photographs of the ***horse statue*** located in front of the Museum of Fine Arts in Boston, following a spiral shooting path. The statue, with its well-defined structural patterns, offers a controlled setting for evaluating reconstruction performance. Images were taken both with and without intentional distractions, introduced by two authors to simulate variability.
Fine-arts Statue. This dataset consists of images of a ***fine-arts statue*** at Northeastern University, also captured in a spiral path. Unlike the horse statue, this statue features irregular textures, providing a

more challenging case for reconstruction. Similarly, images were taken with and without distracting objects (the two authors standing as distractors), and also in a spiral shooting path, to serve as benchmarks.

4.2 Evaluation Metrics

The experiment focused on and compared the reconstruction performance of NeRF on these datasets under three conditions: (1) clean datasets with no distracting objects as the baseline, (2) datasets where distracting objects (people) were masked black, and (3) datasets where masked regions were inpainted using a diffusion model. Quantitative evaluations, using Peak Signal-to-Noise Ratio (PSNR), and qualitative evaluations, based on visual inspection, were conducted to determine the effectiveness of target object segmentation and inpainting in enhancing reconstruction. The PSNR is defined as:

$$\text{PSNR} = 10 \cdot \log_{10} \left(\frac{\text{MAX}^2}{\text{MSE}} \right)$$

where:

- MAX is the maximum possible pixel value of the image (e.g., 255 for 8-bit images).
- MSE is the Mean Squared Error between the original image and the reconstructed image, defined as:

$$\text{MSE} = \frac{1}{m \cdot n} \sum_{i=1}^{m} \sum_{j=1}^{n} (I(i,j) - K(i,j))^2$$

  where:

  - $I(i,j)$ is the pixel value of the original image at position (i,j),
  - $K(i,j)$ is the pixel value of the reconstructed image at position (i,j),
  - m, n are the dimensions of the image.

Higher PSNR values indicate better image quality, as they represent a smaller error between the original and reconstructed images. This approach allowed us to assess the impact of preprocessing techniques on NeRF reconstruction quality comprehensively.

4.3 NeRF Training

Due to the unique nature of each scene, NeRF was trained individually for each dataset and setting, as pre-trained weights could not be transferred from one scene to another. Each training was conducted on a single NVIDIA RTX 4080 GPU with 16GB of memory, taking approximately 8 hours for each experiment.

4.4 Comparision of Scene Rendering

4.4.1 Quantitative Analysis

Methods	**Fine-arts Statue**	**Horse Sculpture**
	PSNR	**PSNR**
NeRF	24.92	24.06
Mask-Nerf	23.76	24.49
Our Method	**28.08**	**21.66**

Table 1: To evaluate the full scene rendering quality, we quantitatively compare our method with NeRF and Mask-NeRF on the Artifact and Horse Sculpture datasets, using PSNR values obtained after 40,000 iterations.

Baseline As our baseline, we use the original NeRF model, which directly reconstructs the 3D scene using raw input images without any pre-processing or mask guidance. The baseline assumes that the training data is clean and includes all objects within the scene. This model demonstrates robust performance on both the Artifact and Horse Sculpture datasets, achieving PSNR scores of 24.92 and 24.06, respectively. However, the presence of irrelevant objects, such as humans in the images, can interfere with the training process, leading to suboptimal reconstruction in complex scenarios.

Comparisons We analyze the performance of our method quantitatively by comparing the PSNR values achieved after 40,000 iterations with those of NeRF and Mask-NeRF on two distinct datasets, evaluating reconstruction accuracy across consistent training conditions.

Mask-Nerf: A variant of NeRF where human subjects are segmented out from the images using a semantic mask. While this approach improves the training data quality by removing distracting elements, it does not handle the missing regions caused by segmentation. Mask-NeRF achieves slightly lower PSNR than the original NeRF on the Artifact dataset (23.76) but slightly higher on the Horse Sculpture dataset (24.49), indicating its ability to reduce distractions in certain scenes.

Our Method: We extend Mask-NerF by incorporating a latent diffusion-based inpainting model (Latent DDPM) to fill the segmented regions with plausible background content. This approach aims to reconstruct more complete scenes by leveraging generative models to synthesize missing details. However, our method results in a lower PSNR (20.08 on the Artifact dataset and 21.66 on the Horse Sculpture dataset), likely due to inconsistencies in the inpainted regions and their integration with surrounding content. These results highlight the trade-offs introduced by inpainting and the challenges in aligning generative models with the NeRF framework.

4.4.2 Qualitative Analysis

(a) Horse Statue (b) Fine-arts Statue

Figure 3: Qualitative Analysis of Segmentation and Reconstruction Results

To further understand the strengths and weaknesses of each method, we present a qualitative analysis focusing on the reconstructed images' visual fidelity and consistency.

Baseline The original NeRF performs poorly on real-world datasets with significant human occlusions. In our two datasets, individuals obstructing key structures such as buildings or statues are incorrectly modeled as part of the scene. Due to the complex shapes of the human regions and their discontinuity with surrounding architectural features, NeRF fails to accurately fit these areas. This results in severe geometric distortions near the occluded structures, significantly degrading reconstruction quality. Consequently, the 3D reconstruction often contains clusters of randomly distributed points or "noise artifacts" around the occluded regions, further highlighting the limitations of the original NeRF in handling such challenges.

Comparisons

MaskFormer & **NeRF** Through fine-tuning the pre-trained model, our approach achieves highly accurate identification of human subjects in the images. The segmentation masks effectively mark the identified humans as black regions, ensuring their removal from the scene during reconstruction.

As shown in multiple images across the dataset, the segmentation results are consistent and precise, demonstrating the robustness of the segmentation step in handling complex scenarios with varying lighting and background conditions. This success lays a strong foundation for the subsequent inpainting process.

Our Method Building on the segmentation results, we utilize the Latent Diffusion Model(LDM) to fill the masked regions with plausible background content. The reconstruction does benefit from inpainted regions recovering some pixels, though the region is not as dense visually. While the inpainting process successfully synthesizes the missing building regions from the occluded buildings in the case of the horse statue, it produces incoherent results in the case of Fine-arts Statue with highly irregular textures or complex geometric patterns. Specifically, inpainting-based reconstructions show mediocre results when there are irregularities in color or texture alignment near occluded regions, which affects the overall visual coherence in multi-view reconstructions. Despite these limitations, the DDPM-based approach significantly enhances the completeness of the reconstructed scenes, particularly in areas where simple masking alone would leave noticeable gaps.

Testing On Public Available Dataset We also qualitatively analyzed the reconstruction using the publicly available *Brandenburger Tor* dataset, which consists of images of the landmark collected from various sources, all containing random distracting objects such as people. For our analysis, we selected a subset of 45 images from the dataset, focusing on tourists as the primary distractors. We applied our method to this subset and compared the results against the same subset without processing through our pipeline prior to reconstruction. Since obtaining an absolutely “clean” dataset for direct comparison is impractical, we rely on visual inspection to evaluate the effectiveness of our approach.

Our result shows significant improvements on the 3d reconstruction model

5 Conclusion

In summary, we proposed a novel pipeline integrating semantic segmentation and inpainting techniques with Neural Radiance Fields (NeRF) to address the challenge of distracting input data in 3D scene reconstruction. Our study demonstrates the interplay between segmentation masks, inpainting quality, and the resulting NeRF reconstruction, emphasizing the potential of preprocessing techniques in NeRF reconstruction. Furthermore, we show that this pipeline offers a largely automated solution for removing extraneous distractions in static scenes, improving the overall effectiveness of the reconstruction process.

The results underline the impact of preprocessing steps, particularly segmentation and inpainting, on reconstruction quality as measured by Peak Signal-to-Noise Ratio (PSNR). While our approach successfully removes unnecessary elements and reconstructs missing regions, we observed limitations in the latent diffusion model (LDM). Specifically, large inpainting regions can lead to inconsistencies, and non-inpainted areas are sometimes affected, resulting in outlier pixels or reduced resolution in the reconstructed model. In addition, due to the limited computational resources, we were not able to fine-tune the LDM specifically for proposed NeRF tasks, making more comprehensive benchmarks **almost impossible** in the setting of a course project. These findings highlight the importance of balancing inpainting quality with integration into NeRF’s pipeline for achieving optimal results.

Moreover, our pipeline showcases the practicality of combining semantic segmentation and inpainting for destruction removal in static scenes. The automation of the process ensures easier implementation and wider applicability, even in challenging datasets with significant occlusions or irregular textures.

Looking ahead, our findings contribute valuable insights into the integration of preprocessing techniques with NeRF for enhanced denoising and reconstruction fidelity. Future work could explore advanced models like Nerfacto[8] for static scene reconstruction and investigate modifications to NeRF’s architecture to directly bypass segmented points, further optimizing PSNR and visual quality. These advancements open new opportunities for improving 3D reconstruction workflows in applications such as computer graphics, virtual reality, and robotics. In terms of other computer vision techniques, future research could focus on fine tuning a better context-aware diffusion models for inpainting tasks. Lastly, we **strongly** suggest to utilize the mask matrices generated from our pipeline for NeRF to better bypass unwanted regions during training, given enough time and resources.

References

[1] Stability AI. Stable diffusion 2 inpainting. https://huggingface.co/stabilityai/stable-diffusion-2-inpainting, 2022. Accessed: 2024-12-10.

[2] Jiahao Chen, Yipeng Qin, Lingjie Liu, Jiangbo Lu, and Guanbin Li. Nerf-hugs: Improved neural radiance fields in non-static scenes using heuristics-guided segmentation, 2024. URL https://arxiv.org/abs/2403.17537.

[3] Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation, 2022. URL https://arxiv.org/abs/2112.01527.

[4] Hyungjin Chung, Jong Chul Ye, Peyman Milanfar, and Mauricio Delbracio. Prompt-tuning latent diffusion models for inverse problems, 2023. URL https://arxiv.org/abs/2310.01110.

[5] Miriam Jäger, Steven Landgraf, and Boris Jutzi. Density uncertainty quantification with nerf-ensembles: Impact of data and scene constraints, 2023. URL https://arxiv.org/abs/2312.14664.

[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, 2020. URL https://arxiv.org/abs/2003.08934.

[7] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2022. URL https://arxiv.org/abs/2112.10752.

[8] Matthew Tancik, Ethan Weber, Evonne Ng, Ruilong Li, Brent Yi, Terrance Wang, Alexander Kristoffersen, Jake Austin, Kamyar Salahi, Abhik Ahuja, David Mcallister, Justin Kerr, and Angjoo Kanazawa. Nerfstudio: A modular framework for neural radiance field development. In *Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Proceedings*, SIGGRAPH '23, page 1–12. ACM, July 2023. doi: 10.1145/3588432.3591516. URL http://dx.doi.org/10.1145/3588432.3591516.