Stable NeRF

Emre Arslan
emre_arslan@brown.edu

Chia-Hong Hsu
chia_hong_hsu@brown.edu

Daniel Cho
daniel_s_cho@brown.edu

Abstract

Generating consistent and realistic novel views from a single image remains a challenging task, especially when relying solely on 2D generative models like Stable Diffusion. Although these models excel in synthesizing high-quality images, they often struggle to maintain consistency between viewpoints. To address this limitation, we propose a novel framework that combines Stable Diffusion with Neural Radiance Fields (NeRF) to achieve stable and coherent novel view generation. Our approach leverages NeRF as a 3D backbone to generate latents, which are used to condition the U-Net in Stable Diffusion for synthesizing novel views. This integration ensures geometric consistency while preserving the high fidelity of 2D generative models. By incorporating the 3D structure provided by NeRF, our framework significantly reduces inconsistencies observed in models relying solely on Stable Diffusion. We demonstrate through qualitative and quantitative experiments that this hybrid approach outperforms existing methods in generating stable, realistic novel views, paving the way for improved applications in 3D-aware image synthesis and view-consistent content generation.

1. Introduction

Synthesizing consistent novel views from a small set of images remains a challenge. While recent advances in 2D generative models demonstrate impressive capabilities in generating high-quality images, these methods struggle with multi-view consistency because they lack an inherent understanding of 3D geometry. In contrast, Neural Radiance Fields (NeRF) provide a continuous 3D representation that ensures geometric consistency for novel view generation. However, NeRF alone lacks the generative capabilities necessary to produce accurate novel views when only provided with a limited number of sampled views.

In this paper, we propose a framework that combines the strengths of Stable Diffusion and NeRF, whose combination addressing the limitations of each. Specifically, we leverage the latent space of Stable Diffusion as the foundation for generative modeling, building on prior work that has demonstrated the benefits of operating in the latent space of diffusion models. With this approach, we inherit the rich 2D priors learned by Stable Diffusion from large-scale datasets, while simultaneously integrating these priors into a continuous 3D representation. This enables us to generate coherent novel views, overcoming the inconsistencies of having few samples of a scene.

Moreover, empirical results have shown that operating in the latent space introduces additional benefits. Unlike methods constrained to single-object scenes, our approach can encode and generate images of multiple objects. This flexibility allows us to further utilize extensive training of Stable Diffusion to generate wider variety of scenes.

Our contributions can be summarized as follows:

1. We propose a method that integrates the 2D priors of Stable Diffusion into a continuous 3D representation using NeRF.

2. By leveraging the latent space of Stable Diffusion, we show that our framework can encode multiple objects and generate diverse range of scenes.

2. Related Work

Zero-1-to-3. This paper introduced a framework for generating novel views using Stable Diffusion conditioned on a single input image. While this method produces promising results, its lack of an underlying 3D representation leads to issues such as view inconsistency and visual noise. This limitation highlights the need to integrate geometric structures into generative pipelines [1].

Diffusion with Forward Models. Another closely related work proposes a diffusion-based framework to model complex stochastic inverse problems. While conceptually similar to our approach, a key difference lies in how priors are utilized. Their method requires training diffusion models from scratch, often requiring a dedicated model for each object. In contrast, we leverage the priors of a pre-trained Stable Diffusion, which not only reduces computational overhead but also enables a more robust framework capable of handling multiple objects. However, our method requires the additional input of a reference view and pose,

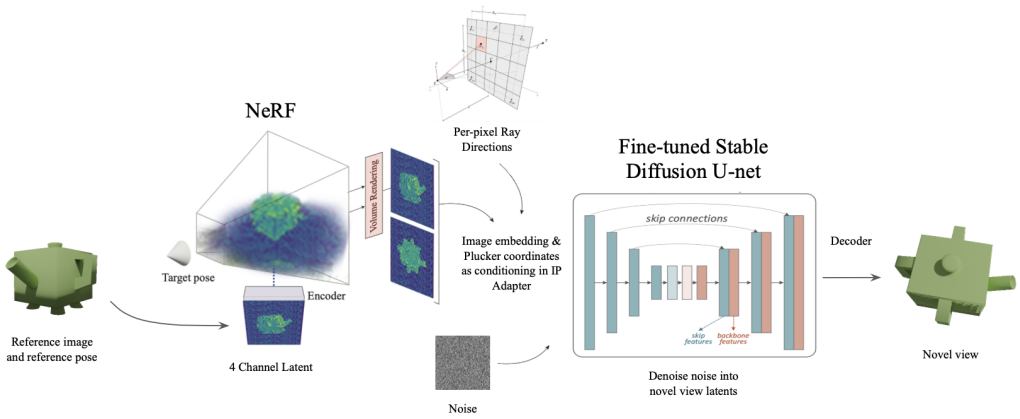

Figure 1. Inference pipeline.

which provides additional conditioning for consistent generation [4].

NeRF. Our work also draws on advancements in Neural Radiance Fields, including Pixel NeRF and Instant Neural Graphics Primitives (NGP). These methods demonstrate the power of NeRF for learning continuous 3D scene representations. However, these approaches alone are not generative and require dense sampling to produce high-quality results [2].

Stable Diffusion. Finally, we build upon recent innovations in Stable Diffusion, including techniques such as the IP-Adapter, which highlight the adaptability of diffusion models for diverse conditioning tasks [3].

3. Datasets

Our method is trained and evaluated on two datasets: the Objaverse dataset and the NeRF dataset, which provide complementary testing grounds for assessing the versatility and robustness of our approach.

The Objaverse dataset consists of a diverse collection of 3D models, with corresponding images and poses generated using Blender. The dataset leverages the EEVEE rendering engine, a real-time renderer that does not use ray tracing. This dataset is valuable for testing our model's ability to generate novel views across a wide variety of relatively simple objects with different shapes and textures.

The NeRF dataset, in contrast, is a measured dataset that includes a set of images capturing a single, but more complex, object—a Lego truck. The dataset is designed for evaluating models on real-world 3D geometry and texture.

4. Method

Our method generates consistent novel views by integrating NeRF into the latent space of Stable Diffusion. This approach combines NeRF's 3D geometric representation with Stable Diffusion's generative power.

4.1. Inference Pipeline

The inference pipeline of our model takes the following inputs: a reference image, reference pose, target image, and target pose.

Latent Generation via NeRF The target pose is fed into a NeRF, which generates a reference latent representation. NeRF provides a 3D-structured latent encoding that captures geometric consistency for the scene. The inference pipeline proceeds as follows:

1. Encoding via Stable Diffusion Encoder. The reference image is passed through the Stable Diffusion encoder to produce a latent representation.

2. Target Latent Generation. The NeRF, trained on Stable Diffusion latents, uses the camera rays produced from the target pose to generate a target latent via volume rendering.

3. Conditioning in the IP Adapter. The reference latent, reference pose, target latent, and target pose are fed into the IP Adapter of the Stable Diffusion U-Net. However, to ensure that sufficient data is present for the IP Adapter to understand the 3D context, the camera pose information is passed in as Plucker Coordinates denoting the per-pixel ray direction. The IP Adapter acts as a bridge, conditioning the U-Net denoising process to utilize the 3D reconstruction from the NeRF.

Furthermore, to make the dimensions of the concatenated latents and Plucker Coordinates compatible with the IP Adapter without loss of information, we introduce a lightweight Convolutional Neural Network (CNN) to do downsampling. The IP Adapter architecture, originally expecting a dimension of 768 or 1024 as opposed to our flattened size of 28672, would have struggled with direct inte-

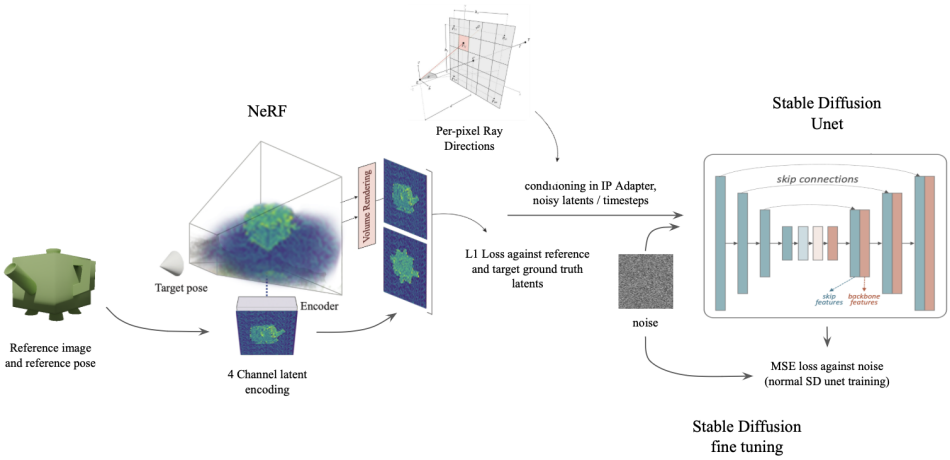

Figure 2. Training pipeline.

gration of such high-dimensional data. By employing this CNN, we achieved gradual downsampling while preserving contextual features, as it is jointly trained to encode and protect essential information, rather than immediately reducing the dimensionality drastically in the subsequent layer.

It is important here to note that the target latent is not fed directly into the Unet. This is because these generated latents often have artifacts that greatly affect the output image–either generating strange features or colors not in the orignal scene.

4. Denoising via U-Net. Pure noise is then denoised by the U-Net, conditioned on the provided inputs. This process produces a novel view latent.

5. Novel View Decoding. Finally, the novel view latent is decoded into a full-resolution image using the Stable Diffusion decoder, producing the output image.

4.2. Training

For training, we use the classic NeRF and Stable diffusion training pipelines.

For our NeRF, we train on preprocessed latents and poses for both the reference and target images. The predicted and true images are the compared using L1 loss over each channel,

$$\text{L1 Loss} = \frac{1}{C \cdot N} \sum_{c=1}^{C} \sum_{n=1}^{N} \left| L_p^{(c)}[n] - L_t^{(c)}[n] \right|, \quad (1)$$

where $L_{p,t}$ are the predicted and true latents, respectively, C the number of channels, and N the indices of the pixels.

For Stable diffusion, rather than fine-tuning the whole Unet from scratch, we opt to leave these weights intact and use an IP Adapter. Now given a reference latent, reference pose, predicted target latent, true target latent, and target pose, all but the true target latent is passed into adapter. Then, we follow classic Stable Diffusion training and calculate the MSE loss between pure noise and the true target latent with incremental noise,

$$\text{MSE Loss} = \frac{1}{C \cdot N} \sum_{c=1}^{C} \sum_{n=1}^{N} \left(L_p^{(c)}[n] - L_t^{(c)}[n] \right)^2, \quad (2)$$

where $L_{p,t}$ are the predicted and true latents, respectively, C the number of channels, and N the indices of the pixels.

5. Results

Despite the challenges associated with generalizing to unseen objects, our framework demonstrated several promising outcomes. While our experiments with the Objaverse dataset did not show significant few-shot generalization of novel view synthesis, the results on the NeRF dataset indicate that the integration of NeRF in latent space offers advantages over classical NeRF.

1. Generalization to Multiple Objects. Classical NeRF struggles to generalize across multiple objects. By leveraging the latent space of Stable Diffusion, our approach successfully encoded and rendered scenes containing multiple objects.

2. Validation on the Lego truck. To evaluate performance in a controlled scenario, we trained and tested our model on the Lego truck. The generated novel views displayed qualitative consistency, confirming the validity of our approach in best-case scenarios.

3. Limitations. While the model performed well on a single object, its generalization capability was limited on the Objaverse dataset. We were not able to produce novel views consistent with the ground truth objects.

Figure 3. Ground truth view.

Figure 4. Predicted view at 400 epochs.

6. Ablations

6.1. Encoding

When NeRF is trained without the Stable Diffusion encoder and decoder, it is unable to train on multiple objects. This behavior is well-documented, as the model solely takes the pose as input [2]. There are other papers that address this issue, but we have noted our experiences here.

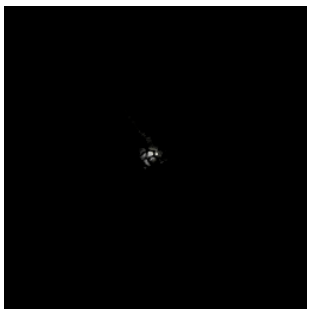

Figure 5. Image from NeRF trained on multiple objects *without* Stable Diffusion Encoding/Decoding.

6.2. IP Adapter

When training NeRF on multiple objects within the latent space of Stable Diffusion, the predicted latents often exhibit artifacts.

Figure 6. Image from NeRF trained on multiple objects *with* Stable Diffusion Encoding/Decoding.

When directly denoised or decoded, the artifacts present in this latent, often create artifacts in the final images.

However, this problem is resolved when working with the IP Adapter. Because latents with artifacts are not directly input into the Unet, these artifacts are not present in the final images.

7. Discussion

Our results underscore the potential and limitations of integrating NeRF with Stable Diffusion in the latent space. This approach addresses some key limitations of classical NeRF and Stable Diffusion for generalizable novel view synthesis, such as the inability to handle multiple objects at high fidelity and the lack of 3D understanding, respectively. However, several areas for improvement were identified.

Our method highly depends on the IP Adapters to learn our conditioning patterns, as well as the NeRF generalizing to distinct classes of data. To enhance generalization, additional training epochs may be required, given the complexity of conditioning a large generative model like Stable Diffusion. In addition, we only use pose information for rendering. Incorporating per-pixel latent features could provide the NeRF with richer contextual information, enabling it to distinguish between objects even when poses are similar.

In conclusion, while our framework represents a step forward in consistent novel view synthesis, there is ample scope for refinement and optimization. These advancements could pave the way for more robust 3D-aware novel view synthesis systems, possibly extended to accommodate even real-world datasets to broaden its applicability.

References

[1] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object, 2023. 1

[2] 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. *CoRR*, abs/2003.08934, 2020. 2, 4

[3] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. *CoRR*, abs/2112.10752, 2021. 2

[4] Ayush Tewari, Tianwei Yin, George Cazenavette, Semon Rezchikov, Joshua B. Tenenbaum, Frédo Durand, William T. Freeman, and Vincent Sitzmann. Diffusion with forward models: Solving stochastic inverse problems without direct supervision. In *arXiv*, 2023. 2