# Increasing Accuracy in Expanded MPI Inference for Single-Image Novel View Synthesis

Sylvia Yuan
Department of Computer Science
Stanford University
luyuan@stanford.edu

## Abstract

*Novel view synthesis is a challenging and intersting problems that focuses on generating 3D consistent views from input images. The task becomes particularly demanding in the context of single-image novel view synthesis, where the available information about the scene is severely limited. There has been advancements in leveraging methods such as outpainting with diffusion models to gain more information about the scene to facilitate novel view synthesis [4]. This project seeks to explore, compare, and improve upon existing methods on the single-image novel view synthesis task.*

## 1. Introduction

Novel view synthesis is an important problem in the field of computer vision. Novel view synthesis aims to generate consistent 3D views from given input images. This problem has various practical applications, as understanding and reconstructing the 3D structure of a scene from limited viewpoints is crucial.

The challenge of generating accurate and realistic new perspectives of a scene based on one or more input images increases substantially when only a single input image is available. Single-image novel view synthesis is particularly demanding because it involves inferring missing information about the scene's depth, structure, and appearance that is not explicitly present in the single view, from severely limited information. This requires methods capable of leveraging limited knowledge and inferring the unseen parts of the scene.

One promising approach in addressing the limitations of single-image inputs, as introduced and leveraged in SinMPI [4], is the use of outpainting techniques, which extend the boundaries of an image by predicting plausible content beyond its edges. Recent advancements have seen the integration of diffusion models in outpainting tasks. Diffusion models, a class of generative models, have demonstrated remarkable performance in generating high-quality images by iteratively refining noisy data into coherent images. By leveraging these models, researchers can effectively extrapolate and infer additional scene details, thus enhancing the quality and consistency of synthesized views from a single image.

The central challenge in single-image novel view synthesis lies in the inherent lack of information available from a single viewpoint. When only a single image is provided, the task of generating new, accurate views requires knowledge about the scene outside of the image boundaries that are not readily available. The lack of data makes it exceedingly difficult to reconstruct spatial context, leading to inaccuracies and lack of information in the synthesized views. However, recent advancements in image generation models, particularly the development of diffusion models, offer a promising avenue to address this issue. Foundational diffusion models are generative models that have been trained on large datasets comprising millions of images from the internet, endowing them with a deep, nuanced understanding of real-world scenes. These models leverage their extensive training to fill in the gaps, effectively extrapolating missing information by drawing upon patterns and structures they have learned from the wide array of training images. By iteratively refining noisy or incomplete input data, diffusion models can generate plausible and coherent extensions of the scene beyond the provided image. This capability allows them to predict the appearance of occluded or unobserved regions with a high degree of realism, thereby improving the fidelity and consistency of the synthesized views. The rich repository of visual knowledge encapsulated within these models enables them to make educated guesses about unseen parts of a scene, facilitating more accurate and realistic novel view synthesis from a single image.

This project aims to delve into the single-image novel view synthesis problem, exploring, comparing, and improving upon existing methods. Specifically, it seeks to harness

the power of advanced outpainting techniques and diffusion models to overcome the inherent challenges posed by the limited information in single-image inputs. This project achieves the following:

1. Explores the SinMPI method and evaluates its performance on the LLFF dataset.

2. Improves SinMPI performance by applying different Stable-Diffusion-based outpainting techniques.

## 2. Related Work

This section discusses categories of approaches to solve the single-image novel view synthesis problem, as well as the diffusion outpainting pipeline used in SinMPI.

### 2.1. NeRF-Based Methods

Neural radiance field methods often require multiple views to optimize continuous fields, such as the original NeRF method, requiring a large amount of poses as input. With recent improvements, more NeRF-based methods have been developed to enable novel view synthesis and 3D reconstruction with fewer input images. For example, SinNeRF [6] produces novel views from single image inputs. SinNeRF uses semantic and geometric regularizations to construct a semi-supervised learning process. SinNeRF uses geometry and semantic pseudo labels to guide the training. But these NeRF based methods suffers from difficulty to infer light fields in addition to colors from single-view image information and therefore suffer from low image quality and artifacts on generated results.

### 2.2. MPI Methods

Discrete multiplane image (MPI) representations is another category of methods that are effective in novel view synthesis. MPIs are a layer-based 3D representation that is camera-centric. However, MPI-based methods are typically constrained by the original camera frustum, which limits their ability to generalize beyond the initial viewpoint, especially in single-image novel view synthesis tasks. These limitations hinder the practical application of MPI-based methods in generating high-quality novel views. SinMPI [4] improves upon prior MPI-based methods and constructs an expanded multiplane image from a single input view, enabling efficient 3D-consistent novel view synthesis.

### 2.3. Diffusion Outpainting

In SinMPI, one of the key details about their methods is applying diffusion outpainting techniques with Stable Diffusion [5] to gain additional depth and pixel information beyond the boundary of the initial input image. Outpainting, a specific application of inpainting that extends the content beyond the original image borders, plays a crucial role in this process. In traditional inpainting, the goal is to fill in missing or corrupted regions within the existing image boundaries. Outpainting, however, pushes this concept further by generating new content outside the initial frame, thus providing a more complete representation of the scene. In the context of SinMPI, outpainting is applied to the boundaries of the input image to produce the missing portions of the scene that were not visible in the single image input. This process involves the use of diffusion inpainting pipelines, which systematically extend the known image content into the unknown areas, ensuring that the newly generated regions are consistent with the existing parts of the scene. Then, SinMPI is able to optimize their MPI representation towards images and depth information produced by a depth-aware warping and inpainting module.

## 3. Approach

This section introduces the methods used in this project, specifically the SinMPI method, and the outpainting pipeline used to enhance the results of SinMPI. Section 3.3 gives more information on the LLFF dataset used in the experiments and evaluation.

### 3.1. SinMPI

SinMPI [4] focuses on using an expanded multiplane image (MPI) as the scene representation for generating novel views from a multiplane space. SinMPI leverages a diffusion model to generate scene information that is not in the original single-view image, through outpainting techniques. Then SinMPI ptojects these generated scene contents into MPI by using depth estimation techniques. SinMPI uses discrete multiplane representations to efficiently create photorealistic 3D views from a single image. This method reduces computational costs compared to continuous representations, allowing for the construction of large 3D spaces and wide-ranging novel views. The process involves projecting content onto an expanded multiplane image and optimizing it with pseudo-multi-view images for guidance.

### 3.2. Modified Outpainting Method

However, we observe problems with the outpainting module of the SinMPI method, as shown in Figure 2. The figure illustrates the comparison between the outpainting results from the original SinMPI implementation and our automated, iterative outpainting pipeline. The figure highlights a critical issue in the SinMPI outpainting module: rather than generating new and contextually appropriate content, the module sometimes mirrors existing portions of the image, leading to unrealistic and repetitive artifacts. These artifacts underscore the module's problem in its inability to seamlessly extend images while maintaining consistency and coherence with the original scene. Addition-

ally, without automatic prompt generation, users are required to manually input prompts for each scene, or risk results similar to those shown in the figure, where the bottom parts of the images often turn into oceans instead of continuous, contextually accurate scenes.

One way to improve SinMPI is to improve its outpainting module, which contains obvious problems and maybe one reason why the results do not seem to be optimal, as shown in Figure 2. I focus on attempting to improve the novel view synthesis results by improving the outpainting method.

We implement two measures to attempt to mitigate the problems shown in Firgue 2, as discussed in the following sections. Figure 1 demonstrates how thes modifications work.

#### 3.2.1 Automatic Prompt Generation with BLIP

Bootstrapping Language-Image Pre-training (BLIP) [1] is a model for vision-language tasks capable of image captioning. To enable automatic prompt generation, we use BLIP on each image to obtain prompt for each input image.

#### 3.2.2 Iterative Outpainting

I observe that iterative outpainting significantly mitigates the issue of diffusion models producing duplicate objects and artifacts in the masked regions of the canvas. This problem is common in outpainting with generative models, where the large masks for inpainting can lead to the generation of improbable or disjointed elements that do not seamlessly integrate with the original image. By employing an iterative approach, the model is better equipped to understand and extend the context incrementally, thereby enhancing the coherence and realism of the outpainted regions while avoiding to producing duplicate subjects according to the prompts.

Given these observations, we decided to implement iterative outpainting for our modified method. Each iteration extends the boundaries of the input image by one-third of its original height and width. This incremental approach is different from with the grid-based outpainting used in the original SinMPI implementation, which often attempts to fill larger masked areas in a single pass. The grid-based method, while efficient, tended to produce less coherent results because it required the model to infer too much contextual information at once, leading to more noticeable artifacts and unrealistic content.

As shown in Figure 1, with our iterative outpainting method, the process begins with a relatively small expansion of the image boundaries. By extending the image in smaller, more manageable increments, the model can more effectively utilize the contextual information from the existing image. Each iteration builds upon the previous one, allowing the model to generate new content that is more consistent with the scene. This step-by-step process ensures that the outpainted regions blend seamlessly with the original image, resulting in a more natural and realistic extension of the scene.

### 3.3. Dataset

I use a subset of the LLFF dataset due to limitation on compute cost. Local Light Field Fusion [2] introduces a dataset of real world photo samples, each scene containing 30 photos taken from different angles. One of these photos will be sampled as the input single view image. The others will be ground truth images for generated novel views. The LLFF dataset also include camera pose information.

## 4. Experiments and Analysis

This section discusses the experimental setup and the evaluation metrics used, as well as analysis of the results.

### 4.1. Experimental Setup

We are using a subset of the scenes from the LLFF Dataset along with their camera pose information. We are running SinMPI, with their depth-aware inpainting module trained for 10 epochs and MPI optimization running for 10 epochs. Each scene is reshaped to be a square image and outpainted to twice the input height and width, with three passes through the stable diffusion inpainting pipeline. We use prompts produced from BLIP as prompts and ”text, border, grid, clutter, lots of things, copy, solid color, distorted, unrealistic, frame” as the negative prompt.

### 4.2. Qualitative Evaluation

For qualitative evaluation, we render various generated images from a range of camera poses and conduct a comparative analysis. By systematically examining these outputs, we can better understand the improvements brought about by our modified outpainting method. As shown in Figure 3, our method exhibits notable advancements over the original SinMPI implementation. In these renderings, the central objects in each scene maintain a high degree of similarity, as the core SinMPI methodology is still employed for the primary reconstruction. This consistency is crucial for ensuring that the fundamental elements of the scene remain accurate and recognizable. However, it is at the periphery of these scenes where the most significant enhancements can be observed.

The borders of the scenes, which were often previously filled with unrealistic and repetitive artifacts, now display a much higher degree of realism and coherence. Our improved outpainting pipeline effectively mitigates issues such as mirrored textures and abrupt transitions, which were common shortcomings in the original method. The newly generated content at the scene boundaries blends seamlessly

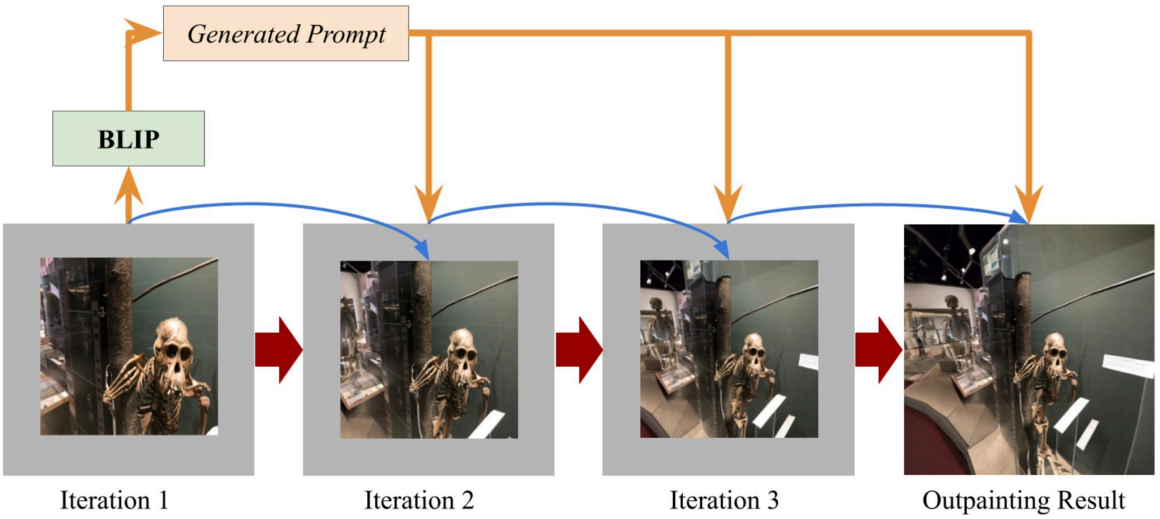


Figure 1. This figure shows the iterative outpainting pipeline used in our outpainting module.

with the existing imagery, creating a more natural and continuous scene.

### 4.3. Quantitative Evaluation

We use the LPIPS [7] metric to measure the quality and accuracy of novel view synthesis on the baseline method. Learned Perceptual Image Patch Similarity (LPIPS) is a metric designed to calculate the perceptual similarity between two images. To compute LPIPS, images are passed through a pre-trained neural network, typically a deep convolutional network such as VGG or AlexNet, which has been trained on a large dataset like ImageNet. LPIPS uses features from these networks. This metric evaluates the perceptual similarity across visual representations and is widely used in novel view synthesis method evaluations to measure the similarity between generated and ground truth results.

| Method | Original SinMPI | New Outpainting |
|---|---|---|
| LPIPS | 0.6985 | 0.6209 |

Table 1. Table showing the LPIPS scores on the subset of LLFF dataset. Our modified outpainting method shows minor improvement in LPIPS score from the original SinMPI output.

Table 1 shows the LPIPS scores produced by the original SinMPI method and SinMPI with our modified outpainting module. Our modified outpainting method shows minor improvement in LPIPS score from the original SinMPI output, likely due to more accurate outpainting producing more realistic scene out of the initial single image.

### 4.4. Analysis

The advancements made in the modified outpainting method have demonstrably improved the outputs of the SinMPI method, particularly in terms of visual quality and realism. Through careful visual inspection, we observe that the enhanced outpainting process generates much more plausible and contextually appropriate content, significantly reducing the occurrence of unrealistic and repetitive artifacts. This improvement underscores the efficacy of our automated, iterative outpainting pipeline in maintaining the continuity and coherence of the extended scene.

However, it is important to note that while visual improvements are evident, the quantitative metrics, specifically the LPIPS score, does not show significant improvement. This discrepancy can be attributed to the intrinsic properties of diffusion models. While diffusion models excel at producing visually appealing and contextually relevant out-of-scene information, they are inherently nondeterministic. This means that the exact scene outside the bounds of the single image input cannot be reproduced with absolute fidelity. The outpainting process, driven by these models, can generate an infinite number of plausible continuations, each varying subtly or significantly from the true scene. This variability, while beneficial for visual plausibility, introduces challenges in achieving improvements in quantitative metrics like LPIPS, which rely on specific measures of similarity to a reference image.


Figure 2. This figure shows comparison of outpainting results from the original SinMPI implementation and our automated, iterative outpainting pipeline. This figure demonstrates the problem in the SinMPI outpainting module. Instead of generating new and contextually appropriate content, the module sometimes mirrors existing portions of the image, resulting in unrealistic and repetitive artifacts. These errors highlight the module's struggle to seamlessly extend images while maintaining consistency and coherence with the original content. And without automatic prompt generation, users need to manually input prompts for each scene or expect results such as in the figure, where the bottom parts of the images are usually turned into oceans, not continuous scenes.

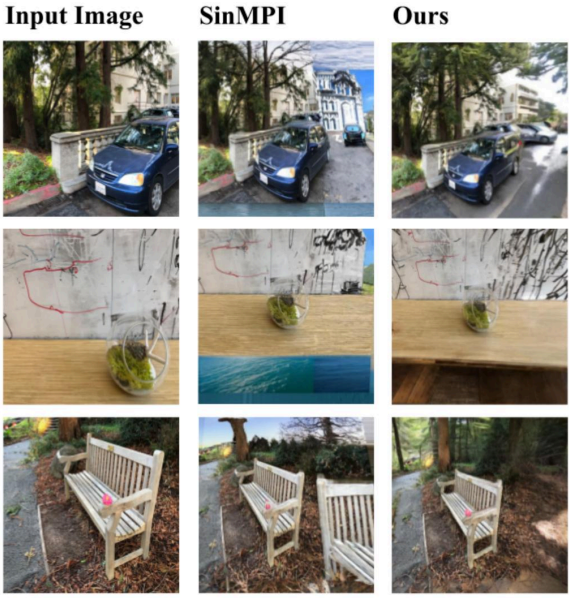


Figure 3. This figure shows comparison of synthesized novel view results from the original SinMPI implementation and SinMPI with our automated, iterative outpainting pipeline. The central objects in each scene remain consistent, maintaining the accuracy of the core SinMPI methodology for primary reconstruction. The most significant enhancements are evident at the periphery of the scenes, where our method effectively addresses previous issues, creating more realistic and coherent extensions.

## 5. Conclusion and Future Works

In conclusion, our modified outpainting approach is able to produce improved results from the original SinMPI method. We achieve qualitative improvements in synthesized novel views while also show minor improvements with quantitative evaluation with the LPIPS metric.

One potential direction for future works is exploring the future works section of SinMPI, which states that the inpainter in SinMPI struggles with realistic textures and inferring light conditions. Diffusion models have also been used to perform relighting tasks [3]. DiFaReli decodes a disentangled light encoding for 3D shape and specifically human facial identity with a DDIM. It would be meaningful to explore if similar relighting techniques can be used to make progress on producing realistic lighting conditions in single-image novel view synthesis.

## 6. Code

Code can be accessed at https://github.com/sylviayuan-sy/CS231A_Project.git. This code is adapted from the official PyTorch implementa-

tion of SinMPI (https://github.com/TrickyGo/SinMPI).

## References


[1] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation, 2022. 3

[2] Ben Mildenhall, Pratul P. Srinivasan, Rodrigo Ortiz-Cayon, Nima Khademi Kalantari, Ravi Ramamoorthi, Ren Ng, and Abhishek Kar. Local light field fusion: Practical view synthesis with prescriptive sampling guidelines, 2019. 3

[3] Puntawat Ponglertnapakorn, Nontawat Tritrong, and Supasorn Suwajanakorn. Difareli: Diffusion face relighting, 2023. 5

[4] Guo Pu, Peng-Shuai Wang, and Zhouhui Lian. Sinmpi: Novel view synthesis from a single image with expanded multiplane images, 2023. 1, 2

[5] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2022. 2

[6] Dejia Xu, Yifan Jiang, Peihao Wang, Zhiwen Fan, Humphrey Shi, and Zhangyang Wang. Sinnerf: Training neural radiance fields on complex scenes from a single image, 2022. 2

[7] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric, 2018. 4