

Abstract of Diffusion-guided Multiview Synthesis for Scene Reconstruction from Single Image, by Zixuan Guo, Brown University, May 2025

This thesis presents a diffusion-guided approach to reconstructing high-quality 3D scenes from a single input image. The proposed method combines generative inpainting techniques and state-of-the-art video diffusion models to synthesize multi-view consistent images, enabling robust 3D reconstructions despite minimal input data. Specifically, our pipeline combines Stable Diffusion inpainting and the TrajectoryCrafter video diffusion model to generate additional scene views based on an initial coarse scene representation from FLAGS. These synthesized views are then reconstructed into a scene with 3D Gaussian Splatting, significantly enhancing geometric accuracy and visual details. Experiments demonstrate that our method consistently produces high-quality, coherent 3D reconstructions, supporting a diverse range of artistic and visual styles. Despite limitations due to diffusion model stability and reliance on COLMAP-based camera estimation, the results highlight the efficacy of integrating generative priors and explicit 3D optimization for high-quality single-image scene reconstruction.

Diffusion-guided Multiview Synthesis for Scene
Reconstruction from Single Image

By

Zixuan Guo

Sc.B., Brown University, 2025

Thesis

Submitted in partial fulfillment of the requirements for Honors in the

Department of Computer Science at Brown University

PROVIDENCE, RHODE ISLAND

May 2025

© Copyright 2025 Zixuan Guo

This thesis by Zixuan Guo is accepted in its present form by the Department of
Computer Science as satisfying the research requirement for the awardment of Honors.

Date _____

James Tompkin, Advisor

Date _____

Srinath Sridhar, Reader

Acknowledgments

I would like to express my profound gratitude to Professor James Tompkin for his unwavering guidance and support throughout the past two years. His mentorship has been invaluable, particularly during moments of self-doubt, and has significantly shaped my personal academic development. I am also deeply thankful to Professor Srinath Sridhar for serving as my second reader.

I am incredibly fortunate to have collaborated with my brilliant lab mates - Runfeng, Mikhail, and Anh. Our countless discussions and shared discoveries over these past two years have made this journey not only intellectually rewarding but also genuinely enjoyable.

I wish to thank my friends, Alicia, Rachel, and Yiyang for their support and accompany throughout this process, especially during the many sleepless nights spent working on this thesis.

Finally, I would like to thank my mom and dad for their care, support, and encouragement. Their steadfast belief in me has been my foundation and source of strength.

CHAPTER 1

INTRODUCTION

Reconstructing detailed, geometrically coherent 3D scenes from a single input image remains fundamentally ill-posed due to the inherent lack of multi-view geometric constraints. While recent advancements, such as NeRF [Mildenhall et al., 2020] and Gaussian Splatting [Kerbl et al., 2023], have significantly enhanced novel-view synthesis from multiple views, the single-image scenario remains a severe challenge. Traditional multi-view stereo (MVS) methods, like COLMAP [Schonberger and Frahm, 2016], inherently depend on overlapping multi-view information for triangulation and are thus inapplicable in single-image contexts. Conversely, monocular depth estimation techniques, such as MiDaS [Ranftl et al., 2020] and DepthAnything [Li and et al., 2024], leverage learned priors from extensive datasets to predict plausible scene geometry. However, these methods often exhibit local inaccuracies and global inconsistencies due to their reliance solely on learned priors without explicit multi-view validation.

Recent progress in generative modeling, especially diffusion models such as Stable Diffusion [Rombach et al., 2022], has shown promising capabilities in synthesizing high-quality image completions through inpainting. However, as these methods operate primarily in the 2D image domain without geometric constraints, they fail to provide the 3D precision required for robust 3D reconstruction. Addressing these limitations, recent video diffusion

models, notably TrajectoryCrafter [Xie, 2024], have emerged, generating temporally and spatially consistent image sequences by conditioning on user-defined camera trajectories. Such diffusion-based view synthesis approaches, including InfiniteNature [Liu et al., 2021] and VideoFusion [Luo, 2023], leverage generative priors to enhance view consistency, substantially improving upon isolated 2D inpainting techniques.

Despite advancements in diffusion-guided view synthesis, directly applying synthesized views to achieve geometrically accurate 3D reconstructions remains challenging. Effective reconstruction necessitates an explicit, optimizable 3D representation seamlessly integrated with diffusion-generated imagery. Recently, the WonderWorld framework [Wu, 2025] introduced a promising solution through the Fast LAYered Gaussian Surfels (FLAGS) representation, enabling rapid scene generation from a single image using depth-guided diffusion. This method significantly accelerates scene initialization, typically achieving results in seconds per scene on a single GPU. However, FLAGS-generated representations remain inherently coarse compared to other 3D representations such as NeRF and 3D Gaussian Splatting, frequently exhibiting incomplete geometry and boundary distortions due to their trade-off of 3D precision for speed.

Building upon these insights, our method introduces a hybrid pipeline specifically targeting high-fidelity reconstruction from single-image inputs. We first utilize Stable Diffusion inpainting in conjunction with FLAGS to generate an initial coarse yet geometrically plausible 3D scene scaffold. Leveraging the temporal and spatial consistency capabilities of the TrajectoryCrafter video diffusion model, we then synthesize multiple views to refine this coarse representation. Integrating both synthesized views and the initial FLAGS scaffold, our pipeline employs iterative optimization through 3D Gaussian Splatting (3DGS) to progressively enhance geometric coherence, detail fidelity, and mitigate occlusions. By combining the strengths of diffusion-based generative priors with robust 3D optimization techniques, this work demonstrates substantial progress in reconstructing high-fidelity 3D scenes from minimal input.

CHAPTER 2

Related Works

2.1 Neural Scene Representations for Reconstruction

Neural scene representations, particularly Neural Radiance Fields [Mildenhall et al., 2020] and 3D Gaussian Splatting (3DGS) [Kerbl et al., 2023], have significantly advanced novel view synthesis, enabling photorealistic rendering from densely captured multi-view data. However, such methods typically rely heavily on extensive, controlled image datasets to accurately capture scene geometry and appearance [Chen et al., 2021, Fan et al., 2024]. Reconstructing scenes from sparse or casually captured data presents significant challenges, including handling occlusions and geometric ambiguity [Guangcong, 2023, Müller et al., 2022]. Recent methods have adapted neural representations to monocular video inputs, exploiting limited baseline motion for depth estimation [Gao et al., 2021]. While effective for visible regions, these approaches struggle with unseen scene parts, limiting their robustness in diverse viewing conditions. In contrast, our method tackles an even more challenging scenario by synthesizing novel views from a single image, significantly extending beyond observable viewpoints through generative diffusion guidance.

2.2 Generative Diffusion-based Novel View Synthesis

Generative diffusion models have recently emerged as powerful methods for synthesizing novel views from limited input data, such as Stable Diffusion [Rombach et al., 2022] and ZeroNVS [Sargent and et al., 2024]. Approaches such as Zero-1-to-3 [Liu and et al., 2023] first proposed camera-conditioned diffusion for objects in simplified scenarios, later generalized by ZeroNVS to broader scene types. Other significant advances include ReConFusion [Wu and et al., 2024] and CAT3D [Gao and et al., 2024], which employ pose supervision and 3D-aware attention mechanisms to enhance spatial coherence. Recent developments in video diffusion models have further improved temporal consistency by integrating camera embeddings [Wang and et al., 2024], depth-warped conditioning [Yu and et al., 2024], Plücker coordinates [Bahmani and et al., 2025], and point cloud supervision [Yu and et al., 2024]. However, these methods face limitations, such as fixed camera trajectories, unsuitable conditioning data for general applications, low-quality outputs, or the necessity of extensive per-scene optimization, restricting broader applicability [Zhang et al., 2024, Sun et al., 2024].

To address these limitations, our approach uniquely synthesizes multiview-consistent novel views from a single input image through a hybrid generative pipeline. Specifically, by combining Stable Diffusion-based inpainting and the pre-trained, trajectory-flexible video diffusion model TrajectoryCrafter [Xie, 2024], we generate spatially and temporally coherent novel views along explicitly defined camera trajectories, significantly improving flexibility, visual quality, and applicability compared to prior works.

2.3 Diffusion-Guided Scene Generation

Previous generative approaches to 3D scene reconstruction have predominantly focused either on generating high-quality static scenes from dense input data or on rapid but coarse approximations [Wu, 2025, Chung and et al., 2023]. Notably, the recent WonderWorld

framework introduces Fast LAYered Gaussian Surfels (FLAGS), a representation enabling rapid scene generation from single image input through depth-guided diffusion [Wu, 2025]. While FLAGS efficiently generates high-fidelity, connected, and diverse 3D scenes, it is nevertheless limited by the flat geometric characteristic of Gaussian surfels and is prone to artifacts and occlusions due to inaccuracies in depth estimation and boundary alignment.

Our method directly builds upon this idea by using FLAGS-generated coarse geometry as an initial scaffold. We then systematically refine this representation using novel views generated by the TrajectoryCrafter video diffusion model [Xie, 2024]. Integrating these multi-view consistent synthetic images into an iterative 3D Gaussian Splatting (3DGS) optimization pipeline significantly improves geometric accuracy, visual detail, and overall consistency. This hybrid pipeline effectively bridges diffusion-generated views and 3D reconstruction, addressing inherent limitations in FLAGS and enabling detailed, high-quality reconstruction from single image input.

CHAPTER 3

Preliminaries

3.1 Fast Layered Gaussian Surfels (FLAGS)

Fast Layered Gaussian Surfels (FLAGS), introduced by [Wu, 2025], represent 3D scenes using layered 2D-aligned Gaussian primitives for efficient rendering and initialization. A FLAGS scene is defined as a union of three surfel layers: foreground $\mathcal{L}_{\text{fg}}$, background $\mathcal{L}_{\text{bg}}$, and sky $\mathcal{L}_{\text{sky}}$, where each layer consists of a set of surfels

$$\mathcal{L} = \{\mathbf{p}_i, \mathbf{q}_i, \mathbf{s}_i, o_i, \mathbf{c}_i\}_{i=1}^N.$$

Each surfel is parameterized by its 3D position $\mathbf{p}_i$, orientation quaternion $\mathbf{q}_i$, anisotropic scale $\mathbf{s}_i = [s_x, s_y]$, opacity o_i , and view-independent color $\mathbf{c}_i$. The spatial extent of each surfel is defined via a Gaussian kernel:

$$G(\mathbf{x}) = \exp\left(-\frac{1}{2}(\mathbf{x} - \mathbf{p})^\top \Sigma^{-1}(\mathbf{x} - \mathbf{p})\right),$$

where the covariance matrix $\Sigma = \mathbf{Q} \text{diag}(s_x^2, s_y^2, \epsilon^2) \mathbf{Q}^\top$ is constructed from the rotation matrix $\mathbf{Q}$ derived from $\mathbf{q}_i$, and a small thickness ϵ to increase expressiveness.

FLAGS are initialized from 2D layer images (e.g., depth-inpainted foregrounds) using pixel-aligned geometry and estimated normals, enabling surfels to be seeded directly in camera space. Unlike full 3D Gaussian Splatting, FLAGS fix the depth axis and optimize in image space, enabling fast initialization and coarse rendering within seconds on a single GPU. While FLAGS sacrifices view-dependent effects, it offers a fast and differentiable scene representation well-suited for single-image initialization.

3.2 Stable Diffusion Inpainting

Stable Diffusion [Rombach et al., 2022] is a latent denoising diffusion probabilistic model (DDPM) that learns to generate high-quality images via iterative denoising in a learned latent space. In the inpainting setting, it is conditioned on a masked image $\mathbf{x}_0 \odot \mathbf{m}$, where $\mathbf{m} \in \{0, 1\}^{H \times W}$ indicates the known regions. The forward process adds Gaussian noise to the latent encoding $\mathbf{z}_0 = \mathcal{E}(\mathbf{x}_0)$, yielding:

$$\mathbf{z}_t = \alpha_t \mathbf{z}_0 + \sigma_t \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I}).$$

The denoising model $\boldsymbol{\epsilon}_\theta$ is trained to predict the noise, minimizing:

$$\min_{\theta} E_{t, \boldsymbol{\epsilon}} [\|\boldsymbol{\epsilon}_\theta(\mathbf{z}_t, t, \mathbf{c}) - \boldsymbol{\epsilon}\|_2^2],$$

where $\mathbf{c}$ encodes both the masked image and optional text prompts. After denoising, the clean image is reconstructed via a VAE decoder $\mathcal{D}$, yielding $\hat{\mathbf{x}}_0 = \mathcal{D}(\mathbf{z}_0)$. This process enables spatially constrained hallucination, ideal for filling occlusions in view synthesis tasks.

3.3 Video Diffusion Models

Video diffusion models extend denoising diffusion probabilistic models to spatiotemporal data [Xie, 2024], operating on a video tensor $\mathbf{x}_0 \in R^{n \times 3 \times H \times W}$ over time t . The forward process progressively adds Gaussian noise:

$$\mathbf{x}_t = \alpha_t \mathbf{x}_0 + \sigma_t \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I}).$$

The reverse process learns a denoiser $\boldsymbol{\epsilon}_\theta$, trained by minimizing:

$$\min_{\theta} E_{t, \boldsymbol{\epsilon}} [\|\boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t) - \boldsymbol{\epsilon}\|_2^2].$$

Modern approaches compress input videos into latent representations $\mathbf{z} = \mathcal{E}(\mathbf{x})$ using a pre-trained video VAE. The latent tensor is patchified and concatenated with conditioning tokens—e.g., text prompts or camera parameters—before being passed to a transformer-based denoiser such as the Diffusion Transformer (DiT). At inference, denoised latents are decoded via $\mathcal{D}(\mathbf{z})$ to produce the final output $\hat{\mathbf{x}}_0$. This framework enables temporally consistent generation and conditioning over structured inputs like camera trajectories, essential for view synthesis tasks.

3.4 3D Gaussian Splatting

3D Gaussian Splatting, proposed by [Kerbl et al., 2023], represents 3D scenes efficiently through scalable Gaussian-shaped primitives and volume rendering techniques. Each Gaussian primitive $\mathbf{G}_i$ is characterized by its opacity factor $o_i \in [0, 1]$, a center position $\mathbf{x}_i \in R^3$, and a covariance matrix $\boldsymbol{\Sigma}_i$. Formally, the Gaussian primitive is defined as:

$$\mathcal{G}_i(\mathbf{x}) = \exp \left(-\frac{1}{2} (\mathbf{x} - \mathbf{x}_i)^\top \boldsymbol{\Sigma}_i^{-1} (\mathbf{x} - \mathbf{x}_i) \right).$$

The covariance matrix Σ_i is decomposed into rotation and scaling components:

$$\Sigma_i = RSS^\top R^\top,$$

where $R \in R^4$ is a quaternion-based rotation matrix, and $S \in R^3$ is a scaling vector. For visualization from a specific viewpoint, each primitive is transformed into camera coordinates via a viewing transformation matrix $\mathbf{W}$ and projected onto the image plane using a local affine transformation matrix $\mathbf{J}$. The resulting covariance matrix Σ'_i is computed as:

$$\Sigma'_i = \mathbf{J}\mathbf{W}\Sigma_i\mathbf{W}^\top\mathbf{J}^\top.$$

By removing the third row and column from Σ'_i , we obtain a 2D covariance matrix Σ_i^{2D} and its corresponding 2D Gaussian $\mathcal{G}_i^{2D}$. These 2D Gaussians are composited using alpha blending to render the final image:

$$\mathbf{c}(\mathbf{x}) = \sum_{i=1}^N \mathbf{c}_i o_i \mathcal{G}_i^{2D}(\mathbf{x}) \prod_{j=1}^{i-1} (1 - o_j \mathcal{G}_j^{2D}(\mathbf{x})).$$

Here, $\mathbf{c}_i \in [0, 1]^3$ denotes view-dependent colors, and $o_i \in [0, 1]$ denotes opacity.

CHAPTER 4

Method

4.1 Overview

Our goal is to reconstruct a large-scale, photorealistic 3D scene from a single RGB image. This is a fundamentally ill-posed problem due to the absence of precise camera and geometry information, requiring plausible view extrapolation beyond the visible content. Prior approaches either hallucinate novel content with little geometric grounding or reconstruct only partial scenes visible from the input image.

To tackle these challenges, we introduce a three-stage hybrid pipeline (Fig. 4.1) that combines text-guided image inpainting, fast surfel-based 3D scaffolding, video diffusion-based multiview synthesis, and fast 3D reconstruction. In Stage 1, we generate a “coarse scene” using depth-aware layered inpainting and initialize a structured FLAGS representation. In Stage 2, we render this scene from a designed camera trajectory and feed the resulting occlusion-prone videos into a pre-trained video diffusion model to synthesize novel views with temporal and spatial consistency. In Stage 3, we aggregate views from both stages to reconstruct the final scene using 3D Gaussian Splatting, achieving geometry-aware consistency and high visual fidelity.

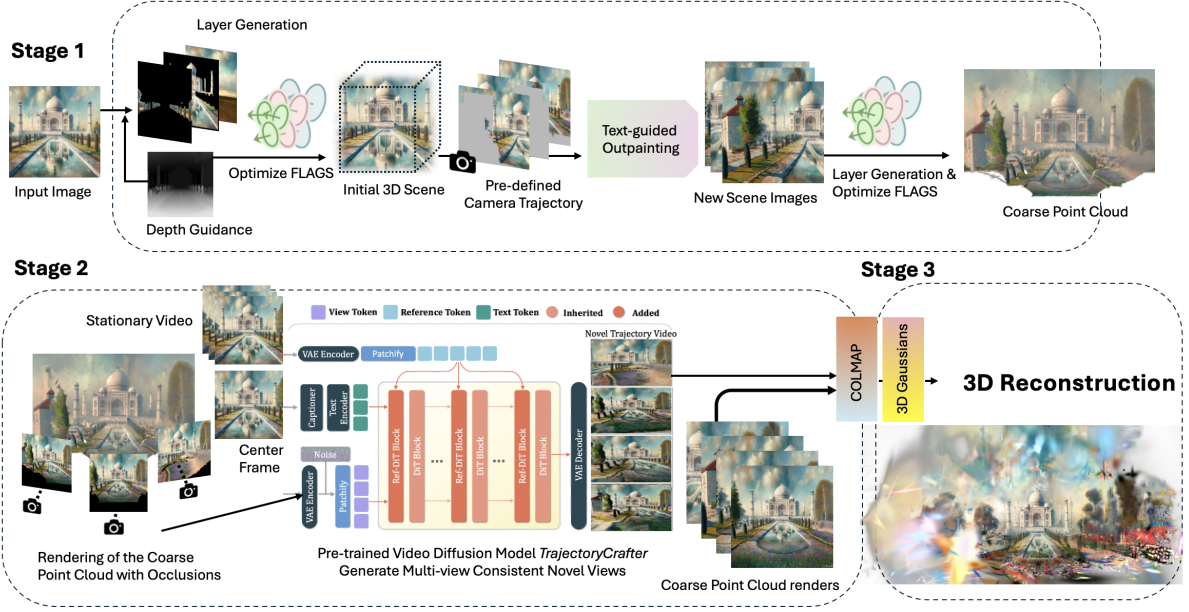


Figure 4.1: Pipeline overview. Starting from a single image, we reconstruct a 3D scene using a hybrid approach of FLAGS-based surfel initialization, video diffusion-based view synthesis, and final scene optimization using 3D Gaussian Splatting.

4.2 Coarse Scene Generation

4.2.1 Layered Inpainting and FLAGS-based Initialization

Given a single input image I_{scene} , we first initialize a canonical view using the WonderWorld pipeline, which constructs a foreground-background-sky layered representation by estimating visible depth from a known camera pose. Specifically, we apply guided depth rendering through the Marigold diffusion model [Ke et al., 2024] to generate a visible depth map $D_{\text{guide}} \in R^{H \times W}$, rendered from a reference viewpoint. We then compute binary visibility masks $M_{\text{fg}}, M_{\text{bg}}, M_{\text{sky}} \in \{0, 1\}^{H \times W}$ using a combination of semantic segmentation and depth gradient heuristics.

To hallucinate occluded or out-of-view regions, we apply Stable Diffusion Inpainting on the masked image. However, directly inpainting a novel view may lead to inconsistencies between the predicted and rendered depth. To address this, WonderWorld introduces a training-free *guided depth diffusion* method, which refines the depth estimation for the

inpainted view by aligning it to the visible depth in D_{guide} . Formally, we sample a latent depth representation $\mathbf{d}_t$ and iteratively denoise it using:

$$\mathbf{d}_{t-1} = \text{Denoise}(\mathbf{d}_t, t, \hat{\epsilon}_t), \quad \hat{\epsilon}_t = \text{UNet}(\mathbf{d}_t, I_{\text{scene}}, t) - s_t \mathbf{g}_t$$

$$\mathbf{g}_t = \nabla_{\mathbf{d}_t} \|D_{t-1} \odot M_{\text{guide}} - D_{\text{guide}} \odot M_{\text{guide}}\|_2^2$$

where $\mathbf{g}_t$ is a guidance term that penalizes deviation from existing depths, and s_t is a time-dependent scalar.

Once the guided depth D_{scene} is decoded, we lift visible pixels from each inpainted frame into 3D using FLAGS. Each surfel is initialized directly from per-pixel geometry and camera intrinsics. By separately optimizing each layer and compositing them via alpha blending, we produce an efficient, layered representation of the scene.

4.2.2 Prompt Design for Inpainting and Scene Coherence

Text prompts used in the Stable Diffusion inpainting process play a crucial role in controlling the image quality, spatial coherence, and stylistic fidelity of the hallucinated content. In the original WonderWorld pipeline [Wu, 2025], prompts are automatically refined using large language models (LLMs) such as ChatGPT to enhance specificity and diversity. In our approach, we instead curate prompts manually to better match the input image and avoid undesired hallucinations. Table 4.1 demonstrates some examples of refined prompts, and Fig.4.2 shows an ablation study that compares the inpaint model’s performance on baseline prompts and refined prompts.

Our self-designed prompts offer several advantages:

- **Reduced Occlusion Risk:** Prompts are crafted to guide inpainting toward object classes that are less likely to occlude the original scene structure (e.g., avoiding trees in architectural scenes).

- **Scene Coherence:** Prompts are consistent across viewpoints, reducing semantic drift and ensuring that generated content smoothly aligns.
- **Style Consistency:** Manual prompts allow us to adhere to the aesthetic or artistic style present in the input image (e.g. realistic landscape V.S. Monet-painting-style landscape)

Table 4.1: Comparison of Baseline and Refined Input Prompt Examples

Scene Name	Baseline Prompts	Refined Input Prompts
“Taj Mahal Monet”	“Monet style meadow” “Monet style garden” “Monet style village” “Monet style lakeside” “Monet style lake” “Monet style town” “Monet style flowerbeds and fountain” “Monet style flowerbeds and fountain”	“Monet-style meadow of poppies and cornflowers beneath soft morning haze” “Monet-style garden with Japanese bridge” “Monet-style village with church spire” “Monet-style lakeside terrace with a wooden jetty and banks of water irises on calm water” “Monet-style lake with water lilies” “Monet-style town waterfront at the far bend of the lake, pastel stone façades lining the quay, arcaded market hall above a cobblestone promenade that continues the shoreline path”
“Forbidden City”	“ancient garden” “ancient city” “ancient temple” “ancient palace” “ancient square” “ancient pond”	“Chinese-style imperial garden with ancient temples and carved stone dragon paths” “Ancient Chinese city with pagodas and arched bridges” “Ancient temple with ornate red columns and golden roof tiles” “Ancient palace courtyard with blooming flowers, crimson palace walls and walkways” “Ancient square with stone paving and ceremonial gates” “Ancient pond with lotus flowers and pavilion reflections”

4.2.3 Camera Trajectory and Mask Design

To ensure effective scene expansion, we define a composite camera trajectory $T = \{T_1, \dots, T_n\}$ that combines leftward rotation, rightward rotation, and zoom-out motion. This combination encourages wide spatial coverage and prevents the scene from becoming fragmented or disconnected. Key criteria for trajectory design include:

- **Connectivity:** Ensures overlapping regions between successive views.
- **Low Occlusion:** Avoids trajectories that cause excessive foreground disocclusion



Figure 4.2: Ablation Study on how prompt design affects stable diffusion inpainting. All results are generated from the same view with the same key scene element (e.g. palace).

(especially sky).

- **Coverage:** Spans a diverse angular range to extend the visible scene volume.

Visualization of an example trajectory can be find in Fig.4.3.

Each rendered view from the FLAGS scene is analyzed for newly emerged regions using Gaussian opacity. Pixels with opacity $o < 0.01$ are flagged as previously unseen region. However, this threshold alone is insufficient for views that newly expose the sky layer or thin background structures. To address this, we apply morphological dilation to the binary mask using an adaptive kernel. The kernel size is determined based on the relative rotation angle of the current view with respect to the previous view:

$$k_{t+1} = k_t + \lambda \cdot \theta_{\text{rot}}$$

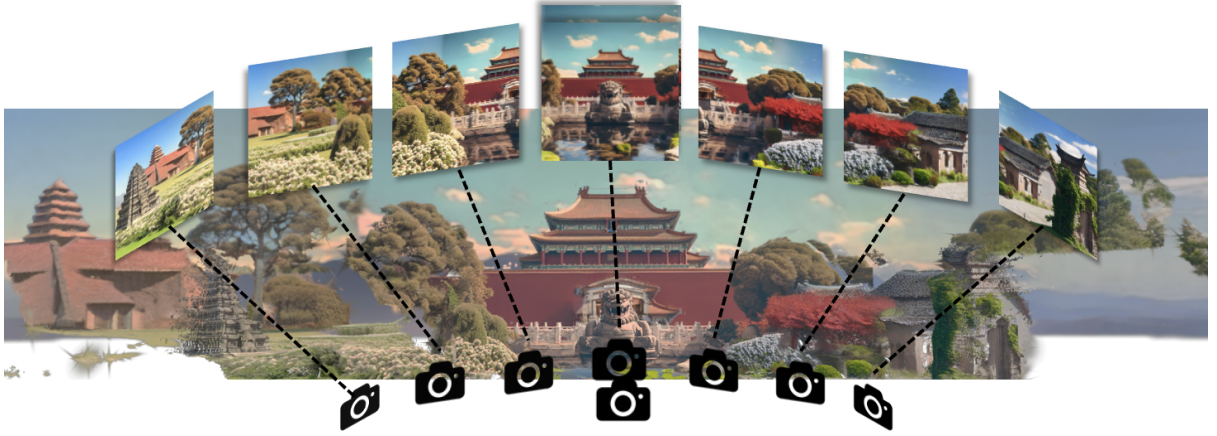


Figure 4.3: Visualization of an example camera trajectory and corresponding views generated from the scene.

where θ_{rot} is the angular deviation from the center frame, and λ is a scaling factor. This dilation step prevents views containing holes or occluded regions to be used as conditional images into the inpaint pipeline, which would result in the generation of low-quality, distorted or disconnected inpainting results, as shown in Fig. 4.4.

In summary, our careful design of both render camera trajectory and mask ensures that the extended views provide maximal useful content for downstream reconstruction, while maintaining topological and semantic consistency with the original input.



Figure 4.4: Ablation Study on dilation kernel size. The left column shows results using a fixed kernel size; the right column shows results using adaptive kernel size. Two sets of results are generated from the same view with the same prompt.

4.3 Multi-view Synthesis with Video Diffusion

4.3.1 Model Architecture

The video diffusion backbone utilized in our pipeline builds upon the recent *TrajectoryCrafter* model [Xie, 2024], which leverages a conditional diffusion architecture for synthesizing novel view videos. TrajectoryCrafter operates by first transforming a given source video sequence into a dynamic point cloud representation and subsequently rendering novel views according to a predefined camera trajectory.

Specifically, given a source video $I^s = \{I_i^s\}_{i=1}^n \in R^{n \times 3 \times h \times w}$, TrajectoryCrafter performs monocular depth estimation to obtain depth maps $D^s = \{D_i^s\}_{i=1}^n$. These depth maps enable a lifting of 2D pixels into a dynamic 3D point cloud representation, computed via inverse perspective projection:

$$P_i = \Phi^{-1}([I_i^s, D_i^s], K),$$

where Φ^{-1} denotes inverse perspective projection, and K is the camera intrinsic matrix.

Subsequently, novel viewpoints $T^r = \{T_i^r\}_{i=1}^n$ are rendered from the dynamic point cloud:

$$I_i^r = \Phi(T_i^r \cdot P_i, K),$$

alongside binary visibility masks M_i^r , indicating occlusions and missing regions. The core innovation in TrajectoryCrafter lies in its dual-stream conditioning diffusion model architecture: (1) **view tokens**, derived from point cloud renderings and masks, and (2) **reference tokens**, extracted from the original source frames. Both streams are jointly processed through stacked Diffusion Transformer (DiT) blocks with dedicated cross-attention mechanisms. These blocks incorporate detailed information from reference tokens into view tokens, thus significantly improving both geometric consistency and photorealism in the generated views.

Connection to our method. The dual-stream architecture benefits our pipeline as it inherently disentangles geometry (view tokens) and appearance (reference tokens). This allows us to explicitly leverage high-quality geometry estimated during Stage 1, thus enhancing consistency between our synthesized novel views and our FLAGS-generated coarse point cloud.

4.3.2 Stationary-Video Conditioning with FLAGS Renders

Instead of conditioning on dynamic video sequences as in the original TrajectoryCrafter framework, our method uses a single stationary video, created by duplicating the initial input image across multiple frames. The reason for doing this is that one frame is sufficient for reference purposes, as we also directly inject information from the original views with renderings of the coarse point cloud. Formally, given an initial FLAGS-based coarse point cloud $\mathcal{P}$ from Stage 1, we render multiple views along a predefined trajectory T^r :

$$I_i^r = \Phi(T_i^r \cdot \mathcal{P}, K), \quad M_i^r = 1[o_i < 0.01],$$

where M_i^r is computed based on FLAGS opacity thresholding.

By substituting dynamic point cloud renderings from TrajectoryCrafter with FLAGS renderings, our adapted pipeline achieves several key advantages:

- **Enhanced Geometric Consistency:** Using FLAGS renderings preserves the precise geometry from the coarse reconstruction stage, eliminating the need for dynamic depth-warping operations, which may introduce projection errors.
- **Reduced Hallucination Artifacts:** As view generation is directly guided by the explicit geometry from FLAGS, the diffusion model is less prone to generating content inconsistent with existing geometry.
- **Style and Appearance Consistency:** The stationary video ensures stable global style characteristics, derived from the original input image, are propagated to novel

synthesized views.

4.3.3 Inference Procedure and Implementation Details

During inference, our pipeline integrates FLAGS renders and stationary video conditioning into the dual-stream TrajectoryCrafter model. The inference protocol proceeds as follows:

Token Encoding. Both FLAGS-rendered views and corresponding visibility masks (I_i^r, M_i^r) are encoded into latent representations using the pretrained TrajectoryCrafter VAE encoder. Similarly, the stationary source frame I_0^s is encoded as reference tokens.

Dual-stream Conditioning. We input the latent view tokens and reference tokens separately into the Diffusion Transformer. At intermediate Ref-DiT blocks, reference tokens guide the cross-attention layer to inject detailed appearance cues into view tokens, balancing geometric accuracy with photorealism.

Diffusion Process. The model iteratively denoises initial random latent noise conditioned on the two token streams and the generated text prompt from the central frame of the reference video (which is simply the static frame). The denoising process follows the standard DDPM objective:

$$\min_{\theta} E_{t,\epsilon} [\|\hat{\epsilon}_{\theta}(x_t, t, v, r) - \epsilon\|^2],$$

where ϵ is noise and t is the diffusion timestep.

Output and Use Case. The resulting synthesized novel views exhibit improved visual continuity, much fewer occluded areas, and are consistent with Stage 1 geometry. These refined views are crucial in filling holes and occlusions previously observed in the FLAGS-based coarse reconstruction, forming a complete input set for Stage 3 optimization with

3D Gaussian Splatting. Fig. 4.5 demonstrates visualizations of our pipeline’s performance on input views with large incomplete areas.

Limitations Noticable in the visualization, there are still some occlusions existing in parts with sky coverage. To address this, manual mask extraction is desired. In addition, we can see that the diffusion model causes blurriness and loss of details. This is one of the fundamental issues with video diffusion models, and as of now, there is not a particularly good solution to it other than additional training or finetuning of the model itself.

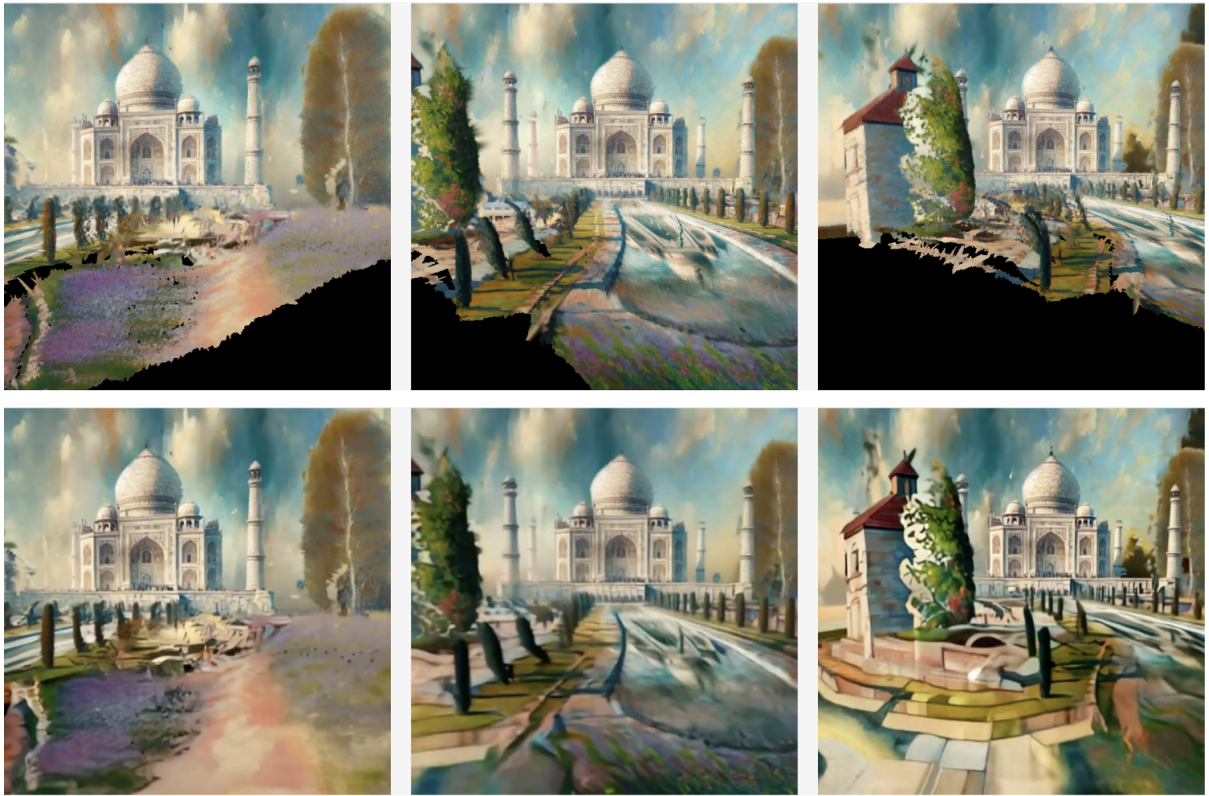


Figure 4.5: Performance of our Stationary Video Conditioning with FLAGS renders. Row 1 are the input views of the coarse point cloud in stage 1, row 2 are the diffusion model outputs.

4.4 3D Reconstruction

4.4.1 COLMAP-based Camera Estimation and Refinement

To construct a refined 3D scene from synthesized novel views, we first estimate camera poses using COLMAP, a widely-used Structure-from-Motion (SfM) and Multi-View Stereo (MVS) pipeline. COLMAP identifies and matches feature keypoints between input images to estimate camera parameters and produce sparse point clouds via incremental bundle adjustment.

The reason we use COLMAP to estimate poses instead of directly reusing our predefined camera trajectories from Stage 1 is that the video diffusion model inherently introduces subtle artifacts, spatial misalignments, and small-scale movements in the generated views. As diffusion models are trained on dynamic video datasets, outputs do not strictly adhere to precise scene geometry, resulting in slight deviations from the original poses. Using the original camera trajectories thus leads to blurred or ghosted reconstruction outcomes. By contrast, COLMAP re-estimates a coherent set of camera poses directly from the generated imagery, significantly improving the sharpness and geometric consistency of the final reconstruction (Fig. 4.6).

However, due to COLMAP’s reliance on robust keypoint detection and matching, the inherent variability and potential inconsistencies introduced by the video diffusion model (TrajectoryCrafter) pose significant challenges. Specifically, COLMAP struggles with frames that exhibit low detail, blurriness, noise artifacts, or insufficient overlap with adjacent views, leading to unstable or failed camera pose estimation.

To mitigate these issues, our pipeline incorporates the following solutions:

- **Enhancing Image Details:** We preprocess generated views to enhance detail clarity, to ensure robust keypoint extraction.
- **Frame Filtering by Quality Scores:** We automatically filter out frames likely to

cause COLMAP failures by computing an aggregate *quality score*, which evaluates:

- *Sharpness*: Calculated via variance of the Laplacian.
 - *Contrast*: Evaluated through histogram analysis.
 - *Detail Density*: Estimated using edge detection methods.
 - *Noise Level*: Quantified through noise estimation metrics.
 - *Content Uniformity*: Assessed to remove overly homogeneous frames.
 - *Compression Artifacts*: Detected to exclude heavily compressed or corrupted frames.
- **Optimized Rendering Trajectory**: We carefully design the rendering trajectory of the coarse point cloud to guarantee sufficient spatial overlap among views, thereby improving COLMAP’s stability and performance.

Through these enhancements, we significantly increase COLMAP’s robustness, enabling more reliable camera pose recovery despite diffusion-generated inconsistencies.



Figure 4.6: Comparison of reconstruction quality using predefined camera trajectories (left) versus COLMAP-estimated trajectories (right). COLMAP effectively mitigates blur and ghosting artifacts caused by slight deviations introduced by the video diffusion model.

4.4.2 Final Scene Reconstruction with 3D Gaussian Splatting

Once robust camera poses are determined, we reconstruct the final high-fidelity 3D scene using 3D Gaussian Splatting (3DGS). This method models the scene geometry and appearance using numerous Gaussian primitives optimized for photorealistic rendering and efficient representation.

Specifically, we initialize the reconstruction using both:

- Original high-quality renders from the coarse FLAGS-based point cloud (Stage 1), leveraging their strong initial geometry and strongest level of details.
- Additional synthesized views generated by TrajectoryCrafter (Stage 2), which fill gaps and occluded regions to yield a more comprehensive reconstruction.

By jointly optimizing Gaussian primitives across these combined view sources, our pipeline effectively integrates robust geometric priors from Stage 1 and the expanded viewpoint coverage from Stage 2. This hybrid reconstruction approach ensures the final scene achieves high geometric accuracy, visual coherence, and reduced artifact presence, overcoming common challenges posed by single-image reconstruction tasks.

CHAPTER 5

Experiment

5.1 Implementation Details

Our pipeline takes around 20 minutes to 40 minutes to run on a NVIDIA L40 GPU with 46GB VRAM, depending on the number of views used in Stage 3 and the speed of COLMAP. Stage 1 takes around 5 minutes to outpaint 8 times, and Stage 2 takes around 5 minutes to run the video diffusion model once and get 50 generated views. The experiments are conducted primarily using scenes from the WonderWorld dataset, which offers diverse initial single-image inputs with various complexity levels.

5.1.1 Diverse Style Generation

One significant advantage of our approach is the capability to support diverse visual styles and scene aesthetics. By careful manual prompt design during Stable Diffusion inpainting, our pipeline successfully generates consistent and coherent 3D scenes across various stylistic categories (Fig. 5.1)

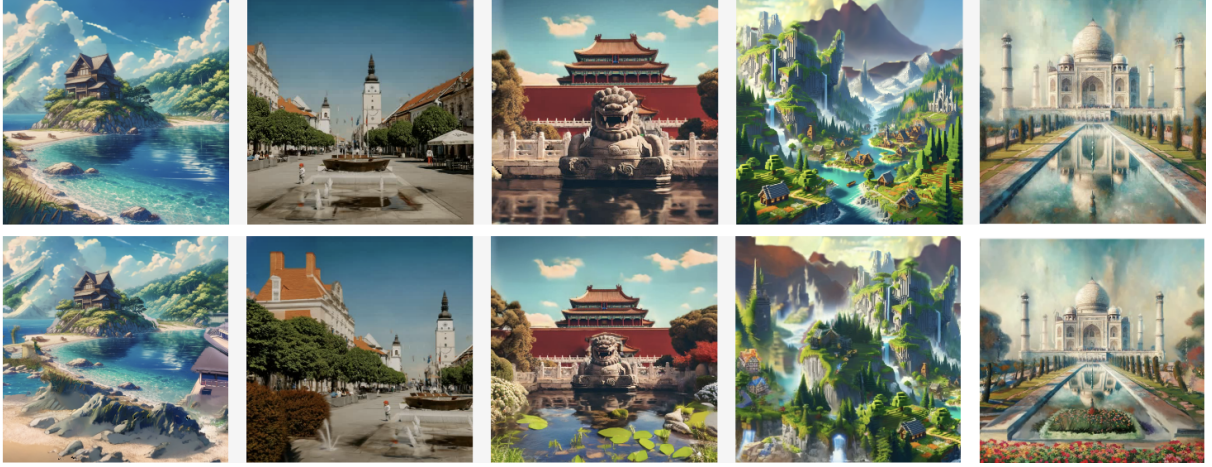


Figure 5.1: Our pipeline handles scene generation of diverse artistic styles, including realistic, animation, Monet paintings, fantasy, and urban.

5.1.2 Baseline Comparison

We evaluate our method against the WonderWorld baseline, which provides state-of-the-art interactive scene generation from single images. Our method effectively reduces occlusions and holes in generated views due to explicit diffusion-based refinement (Fig. 5.2). While we sacrifice the sharpness of our rendered results with our use of video diffusion model, we get a more connected scene with much more 3D precision and less artifacts.

5.1.3 Qualitative Result Analysis

We present qualitative results for two distinct input images representing diverse scenarios:

1. **Forbidden City (realistic)**: Complex architectural scene with detailed landscaping.
2. **Taj Mahal Monet (artistic)**: Artistically stylized scene inspired by Monet.

Given the inherent absence of ground-truth views in single-image 3D reconstruction scenarios, a direct quantitative assessment of correctness or accuracy is not feasible.



Figure 5.2: Left: WonderWorld Baseline. Right: Ours. Both rendered from same viewpoint.

Instead, we employ standard reference-based perceptual metrics commonly used in image synthesis research—Structural Similarity Index Measure (SSIM), Peak Signal-to-Noise Ratio (PSNR), and Learned Perceptual Image Patch Similarity (LPIPS)—to evaluate the similarity between our reconstructed scene renderings and the synthesized novel views generated by the diffusion model. These metrics quantify visual fidelity relative to the synthesized reference images obtained from Stage 2, rather than absolute correctness or spatial and temporal consistency.

Metric results are summarized in Table 5.1, demonstrating high perceptual quality across diverse scenarios.

Table 5.1: Perceptual quality metrics for our pipeline on two representative input scenes.

Scene	SSIM $\uparrow$	PSNR $\uparrow$	LPIPS $\downarrow$
Forbidden City (realistic)	0.9052	28.43	0.1246
Taj Mahal Monet (artistic)	0.9682	36.85	0.0625

Figure 5.3 provides qualitative visualizations of our reconstructed scenes. These results underscore our method’s capability to produce detailed and visually coherent 3D reconstructions, despite being derived entirely from diffusion-generated novel views.



Figure 5.3: Rendering results from reconstructed 3DGS scenes based on two input styles. From left to right: Taj Mahal Monet (Input view), Taj Mahal Monet (Render), Forbidden City (Input view), Forbidden City (Render).

5.2 Limitations

While our method demonstrates significant improvements in single-image 3D reconstruction, several limitations remain, which we outline below:

Reliance on COLMAP for Camera Estimation. Our reconstruction pipeline heavily relies on COLMAP’s Structure-from-Motion (SfM) and Multi-View Stereo (MVS) frameworks. As previously mentioned, COLMAP depends fundamentally on robust keypoint matching and sufficient overlap between views. Due to inherent inconsistencies, artifacts, and occasional blur introduced by the video diffusion model (TrajectoryCrafter), COLMAP frequently struggles to achieve accurate camera pose estimations. Despite our mitigations such as quality-based frame filtering and image detail enhancement, this issue still limits the robustness and scalability of our method.

Instability of Video Diffusion Outputs. TrajectoryCrafter, like most video diffusion models, introduces inherent variability and inconsistency despite its conditioning on depth and explicit view transforms. Generated views often exhibit subtle but detrimental artifacts including slight spatial shifts, temporal inconsistencies, and blurriness. These issues degrade overall reconstruction quality, particularly in finely detailed or geometrically complex scenes, and currently have no robust solutions beyond further model fine-tuning or domain-specific training.

Dependence on Stable Diffusion Inpainting Quality. Our Stage 1 coarse scene initialization relies strongly on Stable Diffusion-based inpainting. The subsequent stages of our pipeline assume that the inpainted content is accurate, coherent, and geometrically plausible. However, failures or low-quality outputs from the inpainting step propagate through later stages, causing significant inaccuracies or artifacts in the final reconstruction. Ensuring high-quality and consistent inpainting results remains crucial for the entire reconstruction process.

Manual Prompt Tuning Requirement. At present, our method requires substantial manual tuning of inpainting prompts and design of camera trajectories to achieve desired visual styles and coherent scene geometry. Automated prompt generation strategies (e.g., via large language models) remain largely experimental and do not consistently meet our desired quality and stylistic consistency. Consequently, manual intervention in prompt engineering remain necessary, presenting an obstacle to full automation and broader applicability of our pipeline.

CHAPTER 6

Conclusion

In this thesis, we presented a novel hybrid pipeline for reconstructing large-scale, detailed 3D scenes from a single input image. By combining the robust generative capabilities of Stable Diffusion-based inpainting and multi-view consistent video diffusion models (TrajectoryCrafter) with efficient 3D representations (FLAGS and 3D Gaussian Splatting), we effectively addressed several key challenges inherent to single-image reconstruction, including occlusions, limited viewpoint coverage, and geometric inconsistencies.

Our method uniquely integrates a coarse scene generation stage using FLAGS to establish strong initial geometry, followed by TrajectoryCrafter-based view synthesis to enrich viewpoint coverage and detail consistency, ending in refined reconstruction via 3D Gaussian Splatting. Quantitative and qualitative evaluations demonstrate that our pipeline achieves significant improvements in reconstruction completeness, visual quality, and style diversity, notably surpassing baseline methods such as only using WonderWorld.

Nevertheless, several limitations remain, notably dependence on COLMAP for camera pose estimation, inherent instability and blurriness introduced by video diffusion models, sensitivity to the quality of Stable Diffusion inpainting, and the manual tuning requirement for effective prompts and camera trajectories. Addressing these limitations offers promising

goals for future research, including more robust automatic prompt engineering, improved diffusion models tailored specifically for novel-view synthesis, and more reliable camera estimation methods independent of traditional keypoint-based SfM pipelines.

Overall, our work advances the state of single-image 3D reconstruction by demonstrating the potential of combining generative diffusion models and explicit 3D representations, opening new possibilities for accessible, high-quality 3D content creation across a wide range of applications.

Bibliography

- [Bahmani and et al., 2025] Bahmani, S. and et al. (2025). VD3D: Taming large video diffusion transformers for 3d camera control. In *International Conference on Learning Representations (ICLR)*.
- [Chen et al., 2021] Chen, A., Xu, Z., Zhao, F., Zhang, X., Xiang, F., Yu, J., and Su, H. (2021). MVSNeRF: Fast generalizable radiance field reconstruction from multi-view stereo. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.
- [Chung and et al., 2023] Chung, J. and et al. (2023). Luciddreamer: Domain-free generation of 3d gaussian splatting scenes. arXiv preprint arXiv:2311.13384.
- [Fan and et al., 2024] Fan, Z. and et al. (2024). Instantsplat: Unbounded sparse-view pose-free gaussian splatting in seconds. arXiv preprint arXiv:2403.20309.
- [Gao et al., 2021] Gao, C., Saraf, A., Kopf, J., and Huang, J.-B. (2021). Dynamic view synthesis from dynamic monocular video. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.
- [Gao and et al., 2024] Gao, R. and et al. (2024). CAT3D: Create anything in 3d with multi-view diffusion models. In *Advances in Neural Information Processing Systems (NeurIPS)*.
- [Guangcong, 2023] Guangcong, e. a. (2023). Sparsenerf: Distilling depth ranking for

- few-shot novel view synthesis. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.
- [Ke et al., 2024] Ke, J., Ye, C., Danelljan, M., Tai, Y.-W., Tang, C.-K., and Yu, F. (2024). Marigold: Marrying diffusion and monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- [Kerbl et al., 2023] Kerbl, B., Kopanas, G., Leimkühler, T., and Drettakis, G. (2023). 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics (TOG)*, 42(4).
- [Li and et al., 2024] Li, L. and et al. (2024). Depthanything: Towards practical monocular depth estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- [Liu et al., 2021] Liu, A., Tucker, R., Jampani, V., Makadia, A., Snavely, N., and Kanazawa, A. (2021). Infinitenature: Perpetual view generation of natural scenes from a single image. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.
- [Liu and et al., 2023] Liu, R. and et al. (2023). Zero-1-to-3: Zero-shot one image to 3d object. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.
- [Luo, 2023] Luo, e. a. (2023). Videofusion: Diffusion models for video-to-video translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- [Mildenhall et al., 2020] Mildenhall, B., Srinivasan, P. P., Tancik, M., Barron, J. T., Ramamoorthi, R., and Ng, R. (2020). NeRF: Representing scenes as neural radiance fields for view synthesis. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

- [Müller et al., 2022] Müller, T., Evans, A., Schied, C., and Keller, A. (2022). Instant neural graphics primitives with a multiresolution hash encoding. *ACM Transactions on Graphics (TOG)*, 41(4).
- [Ranftl et al., 2020] Ranftl, R., Lasinger, K., Hafner, D., Schindler, K., and Koltun, V. (2020). Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*.
- [Rombach et al., 2022] Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- [Sargent and et al., 2024] Sargent, K. and et al. (2024). ZeroNVS: Zero-shot 360-degree view synthesis from a single real image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- [Schonberger and Frahm, 2016] Schonberger, J. L. and Frahm, J.-M. (2016). Structure-from-motion revisited. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- [Sun et al., 2024] Sun, W., Chen, S., Liu, F., Chen, Z., Duan, Y., Zhang, J., and Wang, Y. (2024). Dimensionx: Create any 3d and 4d scenes from a single image with controllable video diffusion. arXiv preprint arXiv:2411.04928.
- [Wang and et al., 2024] Wang, Z. and et al. (2024). MotionCtrl: A unified and flexible motion controller for video generation. In *ACM SIGGRAPH*.
- [Wu, 2025] Wu, e. a. (2025). Wonderworld: Fast and interactive scene generation using flags. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- [Wu and et al., 2024] Wu, R. and et al. (2024). Reconfusion: 3d reconstruction with

diffusion priors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[Xie, 2024] Xie, e. a. (2024). Trajectorycrafter: Camera-controlled video generation via diffusion models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

[Yu and et al., 2024] Yu, W. and et al. (2024). Viewcrafter: Taming video diffusion models for high-fidelity novel view synthesis. arXiv preprint arXiv:2409.02048.

[Zhang et al., 2024] Zhang, D. J., Paiss, R., Zada, S., Karnad, N., Jacobs, D. E., Pritch, Y., Mosseri, I., Shou, M. Z., Wadhwa, N., and Ruiz, N. (2024). Recapture: Generative video camera controls for user-provided videos using masked video fine-tuning. arXiv preprint arXiv:2411.05003.