GenRC: Generative 3D Room Completion from Sparse Image Collections

Ming-Feng Li[1], Yueh-Feng Ku[2], Hong-Xuan Yen[2], Chi Liu[4], Yu-Lun Liu[3], Albert Y. C.[4], Cheng-Hao Kuo[4], and Min Sun[2,4]

[1] Carnegie Mellon University
[2] National Tsing Hua University
[3] National Yang Ming Chiao Tung University
[4] Amazon

Abstract. Sparse RGBD scene completion is a challenging task especially when considering consistent textures and geometries throughout the entire scene. Different from existing solutions that rely on human-designed text prompts or predefined camera trajectories, we propose GenRC, an automated training-free pipeline to complete a room-scale 3D mesh with high-fidelity textures. To achieve this, we first project the sparse RGBD images to a highly incomplete 3D mesh. Instead of iteratively generating novel views to fill in the void, we utilized our proposed E-Diffusion to generate a view-consistent panoramic RGBD image which ensures global geometry and appearance consistency. Furthermore, we maintain the input-output scene stylistic consistency through textual inversion to replace human-designed text prompts. To bridge the domain gap among datasets, E-Diffusion leverages models trained on large-scale datasets to generate diverse appearances. GenRC outperforms state-of-the-art methods under most appearance and geometric metrics on ScanNet and ARKitScenes datasets, even though GenRC is not trained on these datasets nor using predefined camera trajectories. Project page: https://minfenli.github.io/GenRC/

Keywords: 3D synthesis · Panorama inpainting · Diffusion models

1 Introduction

3D Scenes are essential in a diverse range of applications, including virtual reality, augmented reality, computer graphics, and game development. Conventional approaches to acquire a 3D scene are formulated as reconstruction by fitting multiple observations such as point clouds or multi-view images. Starting from the seminal works [25,28], neural implicit representations have become the most popular type of methods as they demonstrated great accuracy and flexibility in reconstruction and rendering. However, these approaches typically require dense observation of the scene for high-quality interpolation but struggle to extrapolate (or generate) the missing part of the scene.

In this work, we explore a generative method for completing a 3D room with a sparse collection of RGBD images (refer to the sparse observation shown

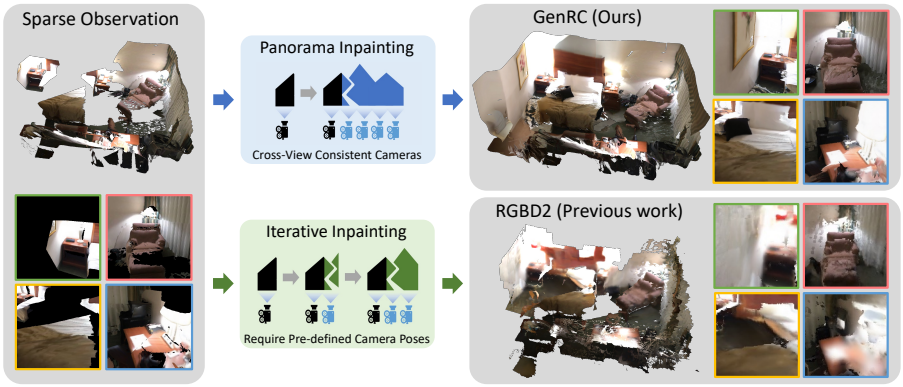


Fig. 1: Scene-level 3D mesh generation. GenRC (the blue path) directly generates a cross-view consistent panorama to complete the main portion of a scene, unlike the iterative methods (the green path) demonstrated in [15, 18] which require designed camera trajectories. GenRC can produce a comprehensive room-scale mesh with high-fidelity texture, even when provided with sparse RGBD observations. Compared with the previous method RGBD2 [18], GenRC excels in generating more complete meshes and high-fidelity images.

in Fig. 1). This task is challenging because the generation of absent parts of the scene is in 3D, which requires cross-view consistency in both appearance and geometry. To tackle this challenge, RGBD2 [18] constructed a diffusion model trained on RGBD data from the ScanNet dataset [9]. This model was then utilized to inpaint the missing parts within a scene iteratively along the poses of a given camera trajectory. Nevertheless, constrained by the limited quantity and diversity of training images, RGBD2 can only produce scene structures with low fidelity, and their visual styles closely resemble the training data from ScanNet. Additionally, since RGBD2 adopted an iterative approach to synthesize novel-view images with adjacent camera poses, it may produce cross-view inconsistent results very sensitive to the camera trajectory. Hence, RGBD2 requires manually selected camera trajectory to be provided as input.

Inspired by recent works to generate high-fidelity images using 2D text-to-image models [27, 31, 32, 34], we propose to leverage foundational diffusion models (e.g., Stable Diffusion [32]) for the completion task and design an automated and training-free pipeline to complete posed RGBD images to a room-scale 3D mesh without the need of human-designed text prompts and predefined camera trajectories. Our method comprises four key steps: (1) Firstly, we extract text embeddings as a token to represent the style of provided RGBD images via textual inversion (Sec. 3.3), and the token will be utilized in text prompts for the text-to-image model. (2) Next, we project the provided RGBD images to a 3D mesh. (3) Following that, we render a panoramic image from a selected room center, which contains missing parts of the scene. Due to the unique equirectangular geometry in panoramic images, standard 2D diffusion inpainting cannot be directly applied. We propose Equirectangular-Diffusion (referred to as E-Diffusion) which explicitly enforces equirectangular projection in the diffusion

process. Our E-Diffusion guided by textual inversion concurrently denoises these images (Sec. 3.4) and determine their depth via monocular depth estimation [2] (Sec. 3.5). This step results in a cross-view consistent panoramic RGBD image that completes the main portion of the mesh. (4) Lastly, we sample novel views from the mesh to fill in holes (Sec. 3.7). With the geometric and stylistic consistency guaranteed by E-Diffusion and textual inversion, our method can effectively generate cross-view consistent room structures without human-designed text prompts and camera trajectories, as shown in Fig. 1. Moreover, benefiting from Stable Diffusion trained on large-scale datasets [35], our method can generate high-fidelity and diverse room structures without any domain-specific training (e.g., training on the ScanNet dataset).

We evaluate our method on the ScanNet [9] and ARKitScenes [4] datasets. In contrast to RGBD2 trained on the ScanNet dataset, our approach demonstrates superior performance in both color and geometric metrics without the need for fine-tuning. Additionally, to showcase the cross-domain adaptability of our approach, we apply our method and RGBD2, which was trained on the ScanNet, to the ARKitScenes dataset. Our method shows better cross-domain adaptability and robustness in real-world scenarios. To summarize, our contributions are:

- Creating cross-view consistent 3D meshes for indoor scenes with sparse RGBD observations through our E-Diffusion (Sec. 3.4) and enhancing the stylistic and geometric coherence of scenes via textual inversion (Sec. 3.3) and a novel active sampling method (Sec. 3.6).
- Generating high-fidelity and diverse room structures on both ScanNet and ARKitScenes by leveraging Stable Diffusions [32] trained on large datasets [35].
- Showcasing a training-free automated pipeline for 3D indoor scene generation without requiring human-designed text prompts or predefined camera trajectories, setting it apart from previous methods.

2 Related Work

2.1 2D Diffusion Models

2D image generation has experienced remarkable progress in text-based control (e.g., GLIDE [27], DALL·E 2 [31], Stable Diffusion [32], and Imagen [34]), primarily driven by large-scale text-image datasets [35] and novel diffusion model architectures [14,32]. In addition to text descriptions, some studies also explored ways to control the output results through other kinds of input conditions such as reference images [23,33,43], sketches [40,44], or incomplete 3D shapes [8,26], etc. To generate precise text descriptions that describe user-desired outputs, [12] provides a framework for converting the concepts of images into text embeddings through optimization.

2.2 3D Shape Generation

Many methods [1,8,10,13,36] have demonstrated object-level 3D generation on a small number of existing 3D datasets such as ShapeNet [7], showing their capabilities to reconstruct 3D shapes of objects. However, due to the limited scale

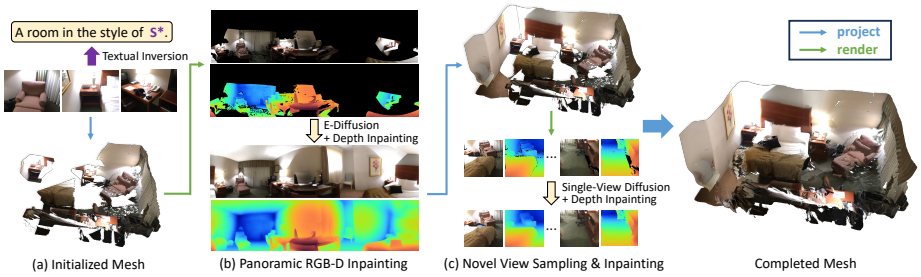


Fig. 2: Pipeline of GenRC: (a) Firstly, we extract text embeddings as a token to represent the style of provided RGBD images via textual inversion. Next, we project these images to a 3D mesh. (b) Following that, we render a panorama from a plausible room center and use equirectangular projection to render various viewpoints of the scene from the panoramic image. Then, we propose E-Diffusion that satisfies equirectangular geometry to concurrently denoise these images and determine their depth via monocular depth estimation, resulting in a cross-view consistent panoramic RGBD image. (c) Lastly, we sample novel views from the mesh to fill in the remaining holes.

and diversity of these datasets, these methods can only generate simple shapes and a limited number of classes. To overcome the scarcity of 3D datasets, recent approaches [17,20,22–24,29,38,41] expanded powerful 2D text-to-image models, such as Stable Diffusion [32], to 3D shape generation tasks. These approaches leverage pre-trained image diffusion models as priors, optimizing 3D models such as Neural Radiance Fields (NeRFs) via Score Distillation Sampling (SDS). Nevertheless, these methods encounter difficulties in processing large-scale 3D scenes with fine-grained textures due to the limited capacities of implicit representations. In contrast, [11,15,18,37] employ explicit representations, such as meshes, to generate or process room-scale 3D scenes and demonstrate high-fidelity visual details. Nevertheless, most of these methods require additional forms of guidance as input to yield ideal results, including detailed text prompts [11,15], carefully designed camera strategies [15,18], or providing initial meshes [37].

2.3 3D-consistent Scene Synthesis

Recent studies of neural implicit representations [25,28] have showcased their ability to produce high-quality reconstructions and synthesize novel views. While these methods often rely on a substantial number of overlapping images for high-quality interpolation, they struggle when it comes to extrapolating missing parts of a scene. In contrast, studies of perpetual view generation [5,11,19,21] aim to generate unseen parts of a scene, performing the synthesis of videos with a single RGB image as the start. However, these approaches only ensure continuity of appearance but not geometric consistency across different views.

2.4 Multi-view Diffusion Models

To generate images from various viewpoints of a scene while ensuring consistent cross-view appearances, multi-view diffusion model [3] harnessed diffusion mod-

els trained for 2D images and incorporated mechanisms to share local features of neighboring 2D images. In each denoising step, the model denoises the neighboring 2D images independently and interpolates the latents of their overlapping regions. These mechanisms guarantee the appearance consistency of generated 2D images across their overlapping areas. However, [3] cannot handle the special equirectangular geometry to produce geometrically correct panoramic images. Recently, [39,42] proposed to train diffusion models on panoramic datasets (Structured3D [46] and Matterport3D [6], respectively) for panorama inpainting. Due to the limited volume and diversity in panoramic datasets, these methods have not been evaluated on cross-datasets to highlight their generalizability. In this paper, we want to leverage powerful pre-trained diffusion models but also have the ability to handle equirectangular geometry. Our E-Diffusion (see Sec. 3) enforces the equirectangular geometry at the beginning steps of the multi-view diffusion process to get the scene geometry correct; then, we apply Texture Refinement diffusion process to enhance the local image quality.

3 Method

GenRC generates a complete 3D mesh with high-fidelity texture, conditioning on sparse RGBD observations. Specifically, given N RGB images $\{\mathbf{I}_i\}_{i=1}^N$, their depth maps $\{\mathbf{D}_i\}_{i=1}^N$ and associated camera poses $\{\mathbf{P}_i\}_{i=1}^N$, our method can generate a complete 3D mesh $M = (V, C, S)$ with the vertices V, vertex colors C, and the faces S. The core idea of our approach is to initially generate a cross-view consistent panoramic RGBD image that completes a main portion of the scene and then generate separate novel views to fill the remaining holes within the room. We describe our pipeline and components below.

3.1 Pipeline Overview

Our pipeline as show in Fig. 2 consists of the following steps: (1) In Fig. 2(a)-Top, we extract text embedding as a token to represent the style of provided RGBD images via textual inversion (see Sec. 3.3). The token will be utilized in text prompts as input to guide the inpainting. At the same time, we initialize a mesh with the given RGBD observations and camera poses by projecting RGB values to 3D and connecting neighboring pixels as triangles (see Fig. 2(a)-Bottom). (2) Next, we render a panoramic image from a selected room center, which contains missing parts of the scene (black parts in Fig. 2(b)-Top). (3) Then, in Fig. 2(b)-Bottom), we applied our proposed E-Diffusion (Sec. 3.4) to inpaint the missing parts in the panoramic image, and we determine their depth via monocular depth estimation [2] (Sec. 3.5). This step results in a cross-view consistent panoramic RGBD image that completes the main portion of the mesh. As the most critical step to ensure final completion quality, we leverage the sampling capability of E-Diffusion and introduce an active sampling method (Sec. 3.6) to pick an RGBD panorama that best matches the given geometry from multiple samples. (4) Lastly, we sample novel views from the mesh to fill in the remaining disconnected holes in an iterative manner (Sec. 3.7).

3.2 Preliminary: Inpainting 3D Meshes

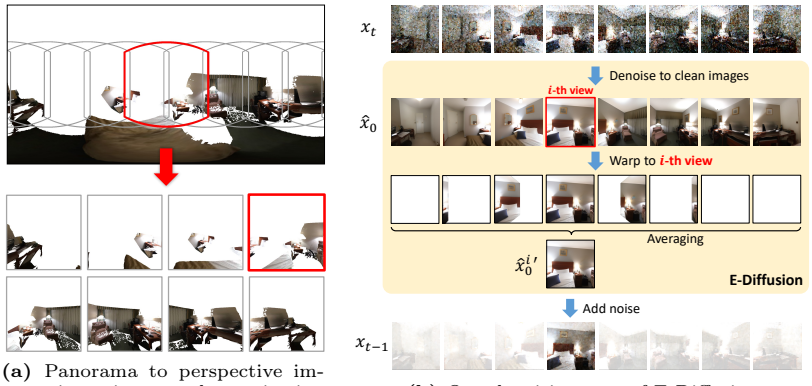
We first define the task of inpainting 2D images. Given an image $\mathbf{I}$, an inpainting mask $\mathbf{m}$ and a text prompt $\mathbf{T}$ as input, the text-to-image model G can generate a completed image $\mathbf{I}'$ by filling in the areas specified by $\mathbf{m}$ with the appropriate appearance. The process is represented as $\mathbf{I}' = G(\mathbf{I}, \mathbf{m}, \mathbf{T})$. Inpainting a 3D mesh can be decomposed into inpainting many 2D images assuming each 2D image captures a portion of the missing surfaces on the 3D mesh. However, the key challenge is that these 2D images are typically interconnected since a missing surface is often not completely observed by a single 2D image. Recent methods [15, 18] address this challenge by iteratively inpainting 2D images along a predefined camera trajectory comprising L poses $\{\hat{\mathbf{P}}_i\}_{i=1}^{L}$. At each step, they render one image $\hat{\mathbf{I}}_i$ from a novel view $\hat{\mathbf{P}}_i$ and generate a mask $\mathbf{m}_i$ indicating which pixels are covered by the mesh. Then, they utilize 2D diffusion models to get the inpainted image $\hat{\mathbf{I}}'_i$ and determine its depth map $\hat{\mathbf{D}}'_i$ with monocular depth estimation models. However, the completion quality of the iterative methods depends on the camera trajectory significantly. Hence, we propose a straightforward yet efficient approach to inpaint a panoramic image that completes a significant portion of the mesh with cross-view consistency.

3.3 Preliminary: Textual Inversion

Using text descriptions to precisely control the style or detailed content of a generated image can be challenging. This task can be significantly simplified once a reference image is given. [12] introduced *textual inversion* to convert a reference image into a token embedding as a textual token that represents the concepts of the image. Given a set of images from the same scene, we utilize *textual inversion* to extract the token S^* that best describes the style of these images. We extract the token S^* for each scene from N given RGB images $\{\mathbf{I}_i\}_{i=1}^{N}$ and then utilize the extracted S^* to describe the style of a room in text prompts. Thanks to textual inversion for stylistic coherency, our method could constantly use a fixed input prompt, "a simple and clean room in the style of S^*.", for our automated pipeline and generate images that have the closest semantics with input images without requiring human-designed text prompts.

3.4 Panorama Inpainting with Equirectangular-Diffusion

A panoramic image is not simply equal to a perspective image with higher resolution. For instance, in a panorama, horizontal lines located in the top and bottom regions should appear curved, not straight, as they should be stretched relative to those in the middle region (see Fig. 4a). Hence, pre-trained diffusion models for perspective images are not suitable for panorama inpainting. However, a panorama can be decomposed into multiple perspective images (see Fig. 3a) since a panorama uses equirectangular projection to represent a spherical surface, and the spherical surface can be represented by multiple 2D perspective images. Hence, the key is to inpaint multiple perspective images while



(a) Panorama to perspective images via equirectangular projection

(b) One denoising step of E-Diffusion

Fig. 3: Multi-view diffusion with equirectangular geometry. (a) Given an incomplete panoramic image, we first obtain several incomplete perspective images via equirectangular projection. (b) To denoise a perspective image at i-th view for one step, we first denoise all images to clean images and warp all the images to i-th view to get an averaged image. Then, we add random noise back to the averaged image to get a perspective image which is denoised for one step. Note that while we use images in RGB space here for illustration, the entire process is operated in latent space.

preserving cross-view consistency under equirectangular projection. We propose Equirectangular-Diffusion (referred to as E-Diffusion), a modified MultiDiffusion [3] for panorama inpainting that considers equirectangular geometry.

To inpaint a panorama with equirectangular geometry, E-Diffusion considers a set of overlapping views represented as M latent images $\{x^i\}_{i=1}^M$ with the resolution of 64×64 pixels, whose camera poses share the same position but have different orientations. For each i-th view, $\{x_T^i, ..., x_0^i\}_{i=1}^M$ is then defined as a series of noisy latent images produced by T steps of the reverse diffusion process (see Fig. 3b). $\{x_T^i\}_{i=1}^M$ are initialized as Gaussian noise and $\{x_0^i\}_{i=1}^M$ are the finally inpainted images. During each step t of the reverse diffusion process, we first predict noise ϵ_θ of x_t^i and obtain $\hat{x}_0^i$ for each i-th view:

$$\hat{x}_0^i = \frac{x_t^i - \sqrt{1-\alpha_t}\epsilon_\theta(x_t^i, t, \mathbf{I}, \mathbf{m}, \mathbf{T})}{\sqrt{\alpha_t}} \tag{1}$$

where $\mathbf{I}$ and $\mathbf{m}$ are the reference image and the inpainting mask in the pixel space, and $\mathbf{T}$ serves as the text prompt.

Then, to ensure multi-view consistency, for each i-th view of $\hat{x}_0^i$, all latent images are warped to this i-th view and perform averaging to get $\hat{x}_0^{i\prime}$:

$$\hat{x}_0^{i\prime} = \frac{\sum_j W_{j\to i}(x_0^j)}{\sum_j m_{j\to i}} \tag{2}$$

where $W_{j\to i}$ is the warp operation from j-th view to i-th view and $m_{j\to i}$ is a binary mask indicating which pixels are visible after warping. Finally, x_{t-1}^i is obtained by adding random noise ϵ to $\hat{x}_0^{i\prime}$:

$$x_{t-1}^i = \sqrt{\alpha_{t-1}}\hat{x}_0^{i\prime} + \sqrt{1-\alpha_{t-1}}\epsilon \tag{3}$$

(a) MultiDiffusion [3] **(b)** E-Diffusion w/o TR **(c)** E-Diffusion

Fig. 4: Comparision of methods for panorama generation. We crop two regions on each panorama and project them to perspective views (the red blocks above). (a) MultiDiffusion [3] can produce a high-resolution image. However, it doesn't satisfy the geometry of equirectangular projection (e.g., the straight lines on the ceiling in the panorama transforming into unrealistic curves in the perspective view). (b) Our proposed E-Diffusion (Sec. 3.4) can generate a panorama that preserves the equirectangular geometry. But without Texture Refinement (TR), the result looks blurry. (c) Applying the last 20 denoising steps for Texture Refinement (TR), our approach achieves the generation of a high-fidelity and high-resolution panorama that adheres to equirectangular geometry.

Texture Refinement (TR). Although our proposed method can generate a reasonable panorama with equirectangular geometry, it would be blurry since the high-frequency information is lost after many interpolations during the warping of latent pixels, as shown in Fig. 4b. To address this issue, in the last F denoising steps of the total T steps, we remove the equirectangular projection and use MultiDiffusion to refine more high-frequency texture, as shown in Fig. 4c. We use $F = 20$ for texture refinement and $T = 50$ as the total steps of the reverse diffusion process in this work.

3.5 Panorama Depth Inpainting

Our panorama depth inpainting method is based on a monocular depth estimation model [2], which can predict an initial depth map for a perspective image and refine the depth map recurrently based on the partial ground-truth depth. To ensure the predicted depth has the same scale as the depth rendered from the mesh, we (1) align the predicted depth to the rendered depth by finding an optimal scale parameter, as [15], after initial depth prediction and (2) use the rendered depth as ground-truth to refine the predicted depth.

Similar to how we ensure multi-view consistency in Sec. 3.4, for each view, we warp the distance maps from the other perspective views to this view and perform averaging to maintain geometric consistency. The *warp-and-average* operation will be performed after predicting initial depth maps for all perspective images and every time we refine the predicted depth with the rendered depth. We consider depth maps as the depth from the image plane to the object surface, while distance maps represent the distance from a camera origin to object surfaces. Distance maps can be converted from or back to perspective depth maps with camera intrinsic. Note that within our automated pipeline, the depth estimation model can be substituted with other depth estimation methods that incorporate the inpainting function.

(a) Initialized mesh **(b)** Bad panorama **(c)** Good panorama

Fig. 5: Active Sampling. Given an initialized mesh from two input views as shown in (a), we try to complete the mesh by inpainting the rendered panorama at the room center (yellow and green dots in (b) and (c)). However, input camera views are sometimes blocked by the mesh inpainted from an unreasonable panorama, as shown in (b). To address this issue, our active sampling strategy samples multiple panoramas as candidates and calculates their mean square errors of depth (Eq. (4)) with all input depth maps to pick the best panorama. This strategy prevents us from using bad panoramas that occlude the given camera views.

3.6 Active Sampling

While the proposed method can generate plausible panoramic RGBD images that match the initialized mesh, the input camera views are sometimes blocked by meshes inpainted from bad panoramic images, as shown in Fig. 5b. This is because the input camera poses and the depth maps suggest that the region between input camera positions and projected surfaces must be free space, but this hint is not considered during panorama generation. Following this idea, we introduce a novel active sampling strategy to pick a panoramic RGBD image that best matches the given priors, as shown in Fig. 5c, from multiple panorama samples. Specifically, given a mesh completed by an RGBD panorama, we can compute the mean square error between the N rendered depth maps $\{\mathbf{D}_i^{render}\}_{i=1}^N$ and the N input depth maps $\{\mathbf{D}_i\}_{i=1}^N$ along input camera poses $\{\mathbf{P}_i\}_{i=1}^N$:

$$MSE_{depth} = \frac{\sum_{i=1}^{N}(\mathbf{D}_i - \mathbf{D}_i^{render})^2}{N} \tag{4}$$

If the mean square error is not close to zero, it implies that the sampled panoramic image may have occluded part of the input views. We can sample a set of A panoramic image $\{\bar{\mathbf{I}}_i\}_{i=1}^A$ corresponding to their predicted depth maps $\{\bar{\mathbf{D}}_i\}_{i=1}^A$ and then pick the one with the minimum mean square error. We set $A = 3$ in this work.

3.7 Mesh Completion

After color and depth inpainting, we achieve a temporal mesh that is almost complete by projecting the RGBD panorama back to 3D. Depending on the use cases, the temporal mesh can be further completed in two different ways. To compare with RGBD2 on the benchmark with fixed camera trajectory, we iteratively inpaint the remaining holes along the fixed camera trajectory. Regardless of the benchmark comparison, we complete the scene by sampling additional

camera poses facing existing holes of the mesh. To select optimal camera poses that cover the big holes within the scene while ensuring an adequate portion of the scene for inpainting, we randomly sample multiple camera poses within the scene and select the poses that capture the viewpoint with the highest product of the number of unobserved pixels and the minimum depth, where the number of unobserved pixels infers the region where we want to inpaint and the minimum depth prevents us from selecting a close-up view on small holes. Then, we inpaint the scene with these selected poses iteratively. Please see supplementary materials for more mesh completion details and visual results.

4 Experiment

Our approach is evaluated on two indoor datasets: (1) ScanNet [9] to compare under RGBD2's setting, and (2) ARKitScenes [4] to showcase the difference in cross-domain adaptability among GenRC and other state-of-the-art methods.

4.1 Setup

Baselines. RGBD2 [18] is the state-of-the-art 3D scene generation method for sparse RGBD inputs on ScanNet. Its key component is a four-channel (i.e., RGBD) diffusion model trained on ScanNet. In addition, to build a baseline powered by Stable Diffusion [32], we modify Text2Room (T2R) [15], a text-to-mesh method, to enable it for completing the mesh from sparse RGBD inputs. We refer it to as T2R+RGBD. We use textual inversion as the textual input to T2R+RGBD and initialize the partial 3D mesh for T2R+RGBD to iteratively inpaint RGB and depth.

Datasets. ScanNet [9] is a well-known RGBD video dataset for indoor generation tasks. We follow the same experiment setting shown in [18], which provides 18 scenes from ScanNet with RGB images and depth maps both in the resolution of 240×180. At each scene, a certain percentage of images will be uniformly sampled as the sparse set of input images, and the rest of them will be used as testing images for computing metrics. Besides, the in-order camera poses of testing images will be used as fixed camera trajectories for two baseline methods and our mesh completion step in the following experiments. Please refer to the supplementary materials for the Sensitivity Analysis of Camera Trajectories which show the impact on each method when in-order camera trajectories are not given. ARKitScenes [4] is a large real-world RGBD video dataset captured using handheld 2020 Apple iPad Pros, making it particularly well-suited for representing real-world use cases. Furthermore, it's worth noting that the depth maps in ARKitScenes exhibit relatively lower density when compared to the ScanNet dataset. Using the same experimental setup as with the ScanNet dataset, we sampled 20 scenes from the ARKitScenes dataset and filtered out the redundant frames. By conducting experiments on the ARKitScenes dataset, we demonstrate that our method exhibits superior cross-domain adaptability and robustness in real-world scenarios.

Evaluation Metrics. We evaluate the performance by comparing RGB and depth renderings from the completed meshes with ground-truth RGB and depth in the testing sets. For visual quality, we compute the Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index Measure (SSIM), and Learned Perceptual Image Patch Similarity [45] (LPIPS). For geometry quality, we computed the Mean Squared Error (MSE) on depth maps. To evaluate if a model can generate images with a similar style to the input RGB images, we compute CLIP Score [30] (CS) for high-level semantic similarity, which is the cosine similarity between the average of the input image CLIP embeddings and the average of the generated image CLIP embeddings. Note that quantitative evaluation is conducted along fixed camera trajectories with ground truth RGBD images. This does not show our strength in completing meshes outside the fixed camera trajectories. Please see supplementary for more qualitative comparisons on panoramic views.

4.2 Implementation Details.

We represent indoor scenes as meshes and utilize Pytorch3D [16] to implement mesh rasterization and fusion. As our text-to-image model, we utilize Stable Diffusion [32] fine-tuned for image inpainting tasks. The resolution of each output image will be 512×512 pixels. As for monocular depth estimation, we employ the IronDepth model [2] to estimate the depth of images generated from Stable Diffusion. To identify a plausible room center for panorama inpainting, we calculate the average of camera positions from input images to serve as the room center. The process of creating a panorama from a sparse set of RGBD observations typically takes approximately 3 minutes, while completing one scene consumes roughly half an hour when running on a single RTX 3090 GPU. To compare our method with RGBD2 on the benchmark with fixed camera trajectory, we inpaint the remaining holes of our generated mesh after panorama inpainting along the fixed camera trajectory for our mesh completion step. Note that the ground truth RGB images and depth maps are at 240×180 pixels resolution, whereas Stable Diffusion for T2R+RGBD and GenRC generates high-resolution images at 512×512 pixels. Hence, we employ Gaussian blur on rendered images and depth maps of these two methods with the kernel size of 5 to reduce high-frequency noise and sharp corners when computing metrics.

4.3 Results on ScanNet

As presented in Tab. 1, GenRC stands out with higher PSNR and SSIM scores, especially when dealing with sparse RGBD observations. The result demonstrates that GenRC excels at reconstructing the visual structures of scenes, as shown in Fig. 6. Additionally, in terms of depth mean square error, GenRC outperforms two baseline methods when sparse observations are provided. This remarkable performance can be attributed to our proposed panorama inpainting technique, as described in Sec. 3.4, which generates cross-view consistent panoramas. T2R+RGBD shows the lowest performance in geometric metrics, despite achieving competitive scores in feature-level and semantic similarity, such as

Table 1: Quantitative results on ScanNet. GenRC stands out with significantly higher PSNR and SSIM scores, especially when dealing with sparse RGBD observations. Additionally, in terms of mean square error in depth estimation, GenRC outperforms the two baseline methods, particularly when sparse observations are provided.

Methods	Visual												Geometric				Semantic			
	$PSNR_{color}(\uparrow)$				$SSIM_{color}(\uparrow)$				$LPIPS_{color}(\downarrow)$				$MSE_{depth}(\downarrow)$				$CS_{input}(\uparrow)$			
	5%	10%	20%	50%	5%	10%	20%	50%	5%	10%	20%	50%	5%	10%	20%	50%	5%	10%	20%	50%
T2R+RGBD [15]	12.9	15.0	16.6	17.2	0.449	0.542	0.591	0.608	0.573	0.492	0.444	0.423	0.57	0.22	0.12	0.12	0.75	0.79	**0.81**	**0.80**
RGBD2 [18]	13.7	16.0	17.5	**18.6**	0.501	0.562	0.598	0.609	0.565	0.488	0.446	0.417	0.38	0.15	**0.08**	**0.06**	0.69	0.71	0.72	0.73
Ours	**14.4**	**16.7**	**17.6**	18.2	**0.524**	**0.599**	**0.628**	**0.633**	**0.531**	**0.441**	**0.410**	**0.400**	**0.27**	**0.13**	0.09	0.09	**0.77**	**0.80**	**0.81**	**0.80**

Table 2: Quantitative results on ARKitScenes. For cross-domain adaptability evaluation, GenRC demonstrates superior performance in both visual and geometric metrics. Compared with RGBD2 trained on ScanNet, GenRC consistently outperforms RGBD2 in each metric, which reflects that GenRC is more suitable for diverse and extensive input data in the real world.

Methods	Visual												Geometric				Semantic			
	$PSNR_{color}(\uparrow)$				$SSIM_{color}(\uparrow)$				$LPIPS_{color}(\downarrow)$				$MSE_{depth}(\downarrow)$				$CS_{input}(\uparrow)$			
	5%	10%	20%	50%	5%	10%	20%	50%	5%	10%	20%	50%	5%	10%	20%	50%	5%	10%	20%	50%
T2R+RGBD [15]	12.0	13.2	14.6	15.5	0.408	0.458	0.521	0.551	0.687	0.624	0.548	0.509	0.66	0.69	0.35	0.27	0.82	0.84	0.86	0.86
RGBD2 [18]	12.2	13.9	15.2	16.6	0.463	0.502	0.541	0.564	0.665	0.596	0.532	0.474	0.51	0.29	0.20	0.11	0.78	0.80	0.81	0.81
Ours	**13.2**	**14.7**	**15.8**	**16.8**	**0.504**	**0.545**	**0.574**	**0.592**	**0.630**	**0.555**	**0.499**	**0.466**	**0.41**	**0.27**	**0.14**	**0.09**	**0.84**	**0.86**	**0.87**	**0.87**

LPIPS and CS. This outcome suggests that T2R+RGBD prioritizes generating high-fidelity images but doesn't effectively address the rational geometry of scenes. Most importantly, as RGBD2 was trained on ScanNet, it performs well when provided with dense RGBD observations as input (i.e., 20% or 50%) in PSNR and MSE depth. However, the experimental results show that such performance of RGBD2 does not generalize well across datasets.

4.4 Cross-domain Results on ARKitScenes

We assess the cross-domain adaptability of GenRC by comparing it with two baseline methods on ARKitScenes [4], without fine-tuning. As presented in Tab. 2, GenRC demonstrates superior performance in both visual and geometric metrics. In comparison to RGBD2 trained on ScanNet, GenRC consistently outperforms RGBD2 in all metrics. In addition, while the input observations are sparse (5% and 10%), both RGBD2 and T2R+RGBD fail to generate reasonable room structures but GenRC can still produce visually pleasing meshes, as shown in Fig. 7. These demonstrate our strength in dealing with sparse observations and show that GenRC based on Stable Diffusion [32] trained on large-scale datasets [35] can be more effective for diverse and extensive input data in the real world.

4.5 Ablation Studies

We carry out ablation studies to confirm the importance of each component of our method by removing them one at a time in Tab. 3. For more visual comparisons, please refer to supplementary materials.

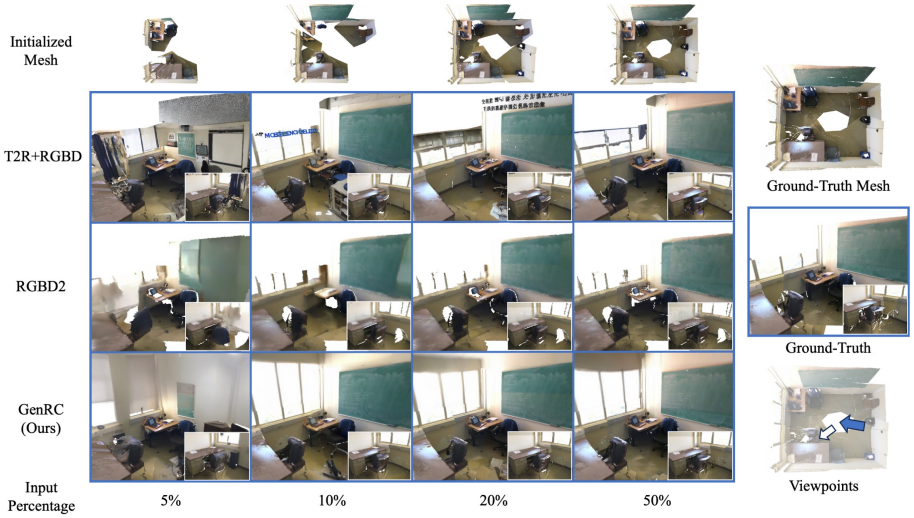


Fig. 6: Comparison with baselines on Scannet. We visualize the generated meshes of each method from two different viewpoints. Leveraging our proposed panorama inpainting technique (Sec. 3.4), GenRC can produce a comprehensive room-scale mesh with high-fidelity texture, even when provided with sparse RGBD observations. In comparison to the prior method RGBD2 [18], GenRC excels in generating more complete meshes and high-fidelity images. Besides, while T2R+RGBD [15] achieves high-fidelity texture, it may generate cross-view inconsistent geometry and artifacts.

Panorama Inpainting. Firstly, we estimate the effectiveness of our proposed panorama inpainting method (Sec. 3.4). Without panorama inpainting, the PSNR declines and the depth mean square error increases significantly when the observations are sparse. The result underlines the importance of utilizing panorama inpainting to complete the main portion of a given mesh at once, instead of iteratively generating novel views to fill in the void. Next, we compare the ablation studies of two panorama inpainting methods, as described in Sec. 3.4: (1) directly applying MultiDiffusion for panorama inpainting (referred to as w/o E-Diffusion); (2) performing our proposed E-Diffusion without texture refinement (referred to as w/o texture refinement). In the case of directly applying MultiDiffusion for panorama inpainting (see Fig. 4a), the generated panorama doesn't adhere to the equirectangular geometry. Hence, both visual and geometric metrics decrease significantly when the observations are sparse. In the case without texture refinement (see Fig. 4b), the blurriness of inpainted panoramas slightly decreases the visual metrics. Also, the mean squared error of depth grows higher because the depth estimation model cannot predict the depth accurately for blurry panoramas without clear corners and details.

Active Sampling. Taking away active sampling yields the highest mean square error of depth, underlining its significance for maintaining geometric consistency with input views. Moreover, active sampling can enhance visual quality by preventing input views from being obstructed or occluded by generated meshes.

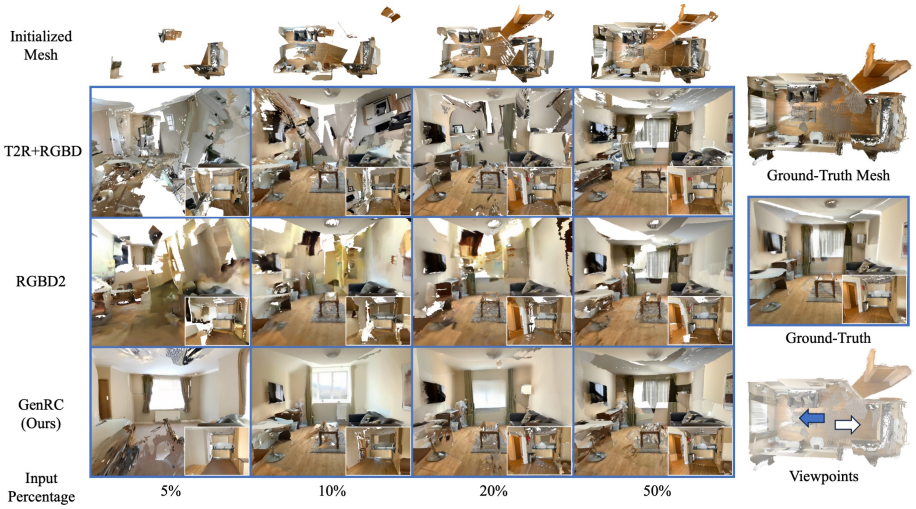


Fig. 7: Comparison with baselines on ArkitScenes. We visualize the generated meshes of each method from two different viewpoints. When the input observations are sparse (5% and 10%), both RGBD2 and T2R+RGBD fail to generate reasonable room structures but GenRC can still produce visually pleasing results.

Table 3: Ablation Studies. Ablation studies reflect the importance of each component in GenRC. The top three components with the highest scores in each metric are marked in red, orange, and yellow respectively. Please refer to Sec. 4.5 for discussions.

Methods	$\text{PSNR}_{\text{color}}(\uparrow)$			$\text{MSE}_{\text{depth}}(\downarrow)$			$\text{CS}_{\text{input}}(\uparrow)$		
	5%	10%	20%	5%	10%	20%	5%	10%	20%
w/o panorama inpainting	13.7	16.1	17.4	0.44	0.15	0.09	0.781	0.805	0.807
w/o E-Diffusion	13.8	16.4	17.5	0.32	0.16	0.09	0.765	0.799	0.808
w/o texture refinement	14.6	16.4	17.7	0.35	0.21	0.13	0.785	0.808	0.807
w/o active sampling	14.5	16.4	17.6	0.46	0.19	0.14	0.794	0.810	0.809
w/o textual inversion	14.7	16.5	17.6	0.30	0.16	0.12	0.792	0.811	0.809
Ours (full)	14.4	16.7	17.6	0.27	0.13	0.09	0.794	0.812	0.809

Textual Inversion. Textual inversion helps to generate images that show higher semantic similarity with provided images, resulting in higher CLIP scores.

5 Conclusion

In this work, we proposed GenRC, a training-free automated pipeline for 3D indoor scene generation. By leveraging the powerful pre-trained diffusion model [32] and our proposed E-Diffusion (Sec. 3.4) for cross-view consistent panoramas, GenRC can generate complete room-scale 3D meshes with high-fidelity texture given a sparse collection of RGBD images. Furthermore, we proposed an active sampling manner (Sec. 3.6) and utilized textual inversion to enhance the geometric and stylistic consistency of scenes. Notably, GenRC outperforms state-of-the-art methods under most appearance and geometric metrics on ScanNet and ARKitScenes datasets even though GenRC was not trained on these datasets.

Acknowledgements

This project is supported by the National Science and Technology Council (NSTC) and Taiwan Computing Cloud (TWCC) under the project NSTC 112-2634-F-002-006 and 113-2221-E-007-104.

References

1. Anciukevičius, T., Xu, Z., Fisher, M., Henderson, P., Bilen, H., Mitra, N.J., Guerrero, P.: Renderdiffusion: Image diffusion for 3d reconstruction, inpainting and generation. In: CVPR. pp. 12608–12618 (2023) 3
2. Bae, G., Budvytis, I., Cipolla, R.: Irondepth: Iterative refinement of single-view depth using surface normal and its uncertainty. BMVC (2022) 3, 5, 8, 11
3. Bar-Tal, O., Yariv, L., Lipman, Y., Dekel, T.: Multidiffusion: Fusing diffusion paths for controlled image generation. ICML (2023) 4, 5, 7, 8
4. Baruch, G., Chen, Z., Dehghan, A., Dimry, T., Feigin, Y., Fu, P., Gebauer, T., Joffe, B., Kurz, D., Schwartz, A., et al.: Arkitscenes: A diverse real-world dataset for 3d indoor scene understanding using mobile rgb-d data. arXiv preprint arXiv:2111.08897 (2021) 3, 10, 12
5. Cai, S., Chan, E.R., Peng, S., Shahbazi, M., Obukhov, A., Van Gool, L., Wetzstein, G.: Diffdreamer: Towards consistent unsupervised single-view scene extrapolation with conditional diffusion models. In: ICCV. pp. 2139–2150 (2023) 4
6. Chang, A., Dai, A., Funkhouser, T., Halber, M., Niessner, M., Savva, M., Song, S., Zeng, A., Zhang, Y.: Matterport3d: Learning from rgb-d data in indoor environments. International Conference on 3D Vision (3DV) (2017) 5
7. Chang, A.X., Funkhouser, T., Guibas, L., Hanrahan, P., Huang, Q., Li, Z., Savarese, S., Savva, M., Song, S., Su, H., et al.: Shapenet: An information-rich 3d model repository. arXiv preprint arXiv:1512.03012 (2015) 3
8. Cheng, Y.C., Lee, H.Y., Tulyakov, S., Schwing, A.G., Gui, L.Y.: Sdfusion: Multimodal 3d shape completion, reconstruction, and generation. In: CVPR. pp. 4456–4465 (2023) 3
9. Dai, A., Chang, A.X., Savva, M., Halber, M., Funkhouser, T., Nießner, M.: Scannet: Richly-annotated 3d reconstructions of indoor scenes. In: CVPR. pp. 5828–5839 (2017) 2, 3, 10
10. Erkoç, Z., Ma, F., Shan, Q., Nießner, M., Dai, A.: Hyperdiffusion: Generating implicit neural fields with weight-space diffusion. In: ICCV (2023) 3
11. Fridman, R., Abecasis, A., Kasten, Y., Dekel, T.: Scenescape: Text-driven consistent scene generation. NeurIPS (2023) 4
12. Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A.H., Chechik, G., Cohen-Or, D.: An image is worth one word: Personalizing text-to-image generation using textual inversion. arXiv preprint arXiv:2208.01618 (2022) 3, 6
13. Gao, J., Shen, T., Wang, Z., Chen, W., Yin, K., Li, D., Litany, O., Gojcic, Z., Fidler, S.: Get3d: A generative model of high quality 3d textured shapes learned from images **35**, 31841–31854 (2022) 3
14. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020) 3
15. Höllein, L., Cao, A., Owens, A., Johnson, J., Nießner, M.: Text2room: Extracting textured 3d meshes from 2d text-to-image models. ICCV (2023) 2, 4, 6, 8, 10, 12, 13

16. Johnson, J., Ravi, N., Reizenstein, J., Novotny, D., Tulsiani, S., Lassner, C., Branson, S.: Accelerating 3d deep learning with pytorch3d. In: SIGGRAPH Asia 2020 Courses, pp. 1–1 (2020) 11
17. Kasten, Y., Rahamim, O., Chechik, G.: Point-cloud completion with pretrained text-to-image diffusion models. NeurIPS (2023) 4
18. Lei, J., Tang, J., Jia, K.: Rgbd2: Generative scene synthesis via incremental view inpainting using rgbd diffusion models. In: CVPR. pp. 8422–8434 (2023) 2, 4, 6, 10, 12, 13
19. Li, Z., Wang, Q., Snavely, N., Kanazawa, A.: Infinitenature-zero: Learning perpetual view generation of natural scenes from single images. In: ECCV. pp. 515–534 (2022) 4
20. Lin, C.H., Gao, J., Tang, L., Takikawa, T., Zeng, X., Huang, X., Kreis, K., Fidler, S., Liu, M.Y., Lin, T.Y.: Magic3d: High-resolution text-to-3d content creation. In: CVPR. pp. 300–309 (2023) 4
21. Liu, A., Tucker, R., Jampani, V., Makadia, A., Snavely, N., Kanazawa, A.: Infinite nature: Perpetual view generation of natural scenes from a single image. In: ICCV. pp. 14458–14467 (2021) 4
22. Liu, M., Xu, C., Jin, H., Chen, L., Xu, Z., Su, H., et al.: One-2-3-45: Any single image to 3d mesh in 45 seconds without per-shape optimization. arXiv preprint arXiv:2306.16928 (2023) 4
23. Liu, R., Wu, R., Van Hoorick, B., Tokmakov, P., Zakharov, S., Vondrick, C.: Zero-1-to-3: Zero-shot one image to 3d object. In: ICCV. pp. 9298–9309 (2023) 3, 4
24. Metzer, G., Richardson, E., Patashnik, O., Giryes, R., Cohen-Or, D.: Latent-nerf for shape-guided generation of 3d shapes and textures. In: CVPR. pp. 12663–12673 (2023) 4
25. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. In: ECCV (2020) 1, 4
26. Müller, N., Siddiqui, Y., Porzi, L., Bulo, S.R., Kontschieder, P., Nießner, M.: Diffrf: Rendering-guided 3d radiance field diffusion. In: CVPR. pp. 4328–4338 (2023) 3
27. Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., Chen, M.: Glide: Towards photorealistic image generation and editing with text-guided diffusion models. arXiv preprint arXiv:2112.10741 (2021) 2, 3
28. Park, J.J., Florence, P., Straub, J., Newcombe, R., Lovegrove, S.: Deepsdf: Learning continuous signed distance functions for shape representation. In: CVPR (2019) 1, 4
29. Poole, B., Jain, A., Barron, J.T., Mildenhall, B.: Dreamfusion: Text-to-3d using 2d diffusion. arXiv preprint arXiv:2209.14988 (2022) 4
30. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: ICML. pp. 8748–8763. PMLR (2021) 11
31. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125 **1**(2), 3 (2022) 2, 3
32. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR. pp. 10684–10695 (2022) 2, 3, 4, 10, 11, 12, 14
33. Saharia, C., Chan, W., Chang, H., Lee, C., Ho, J., Salimans, T., Fleet, D., Norouzi, M.: Palette: Image-to-image diffusion models. In: ACM SIGGRAPH 2022 Conference Proceedings. pp. 1–10 (2022) 3

34. Saharia, C., Chan, W., Saxena, S., Li, L., Whang, J., Denton, E.L., Ghasemipour, K., Gontijo Lopes, R., Karagol Ayan, B., Salimans, T., et al.: Photorealistic text-to-image diffusion models with deep language understanding **35**, 36479–36494 (2022) 2, 3
35. Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., et al.: Laion-5b: An open large-scale dataset for training next generation image-text models **35**, 25278–25294 (2022) 3, 12
36. Shue, J.R., Chan, E.R., Po, R., Ankner, Z., Wu, J., Wetzstein, G.: 3d neural field generation using triplane diffusion. In: CVPR. pp. 20875–20886 (2023) 3
37. Song, L., Cao, L., Xu, H., Kang, K., Tang, F., Yuan, J., Zhao, Y.: Roomdreamer: Text-driven 3d indoor scene synthesis with coherent geometry and texture. ACM MM (2023) 4
38. Tang, J., Ren, J., Zhou, H., Liu, Z., Zeng, G.: Dreamgaussian: Generative gaussian splatting for efficient 3d content creation (2024) 4
39. Tang, S., Zhang, F., Chen, J., Wang, P., Furukawa, Y.: Mvdiffusion: Enabling holistic multi-view image generation with correspondence-aware diffusion. NeurIPS (2023) 5
40. Voynov, A., Aberman, K., Cohen-Or, D.: Sketch-guided text-to-image diffusion models. In: ACM SIGGRAPH 2023 Conference Proceedings. pp. 1–11 (2023) 3
41. Wang, H., Du, X., Li, J., Yeh, R.A., Shakhnarovich, G.: Score jacobian chaining: Lifting pretrained 2d diffusion models for 3d generation. In: CVPR. pp. 12619–12629 (2023) 4
42. Wu, T., Zheng, C., Cham, T.J.: Panodiffusion: Depth-aided 360-degree indoor rgb panorama outpainting via latent diffusion model. In: ICLR (2024) 5
43. Yang, B., Gu, S., Zhang, B., Zhang, T., Chen, X., Sun, X., Chen, D., Wen, F.: Paint by example: Exemplar-based image editing with diffusion models. In: CVPR. pp. 18381–18391 (2023) 3
44. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: ICCV. pp. 3836–3847 (2023) 3
45. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: CVPR. pp. 586–595 (2018) 11
46. Zheng, J., Zhang, J., Li, J., Tang, R., Gao, S., Zhou, Z.: Structured3d: A large photo-realistic dataset for structured 3d modeling. In: ECCV. pp. 519–535. Springer (2020) 5