

Pano2Room: Novel View Synthesis from a Single Indoor Panorama

Guo Pu

guopu@pku.edu.cn

Wangxuan Institute of Computer Technology, Peking University Beijing, China

Yiming Zhao

zhaoym@pku.edu.cn

Wangxuan Institute of Computer Technology, Peking University Beijing, China

Zhouhui Lian*

lianzhouhui@pku.edu.cn

Wangxuan Institute of Computer Technology, Peking University Beijing, China

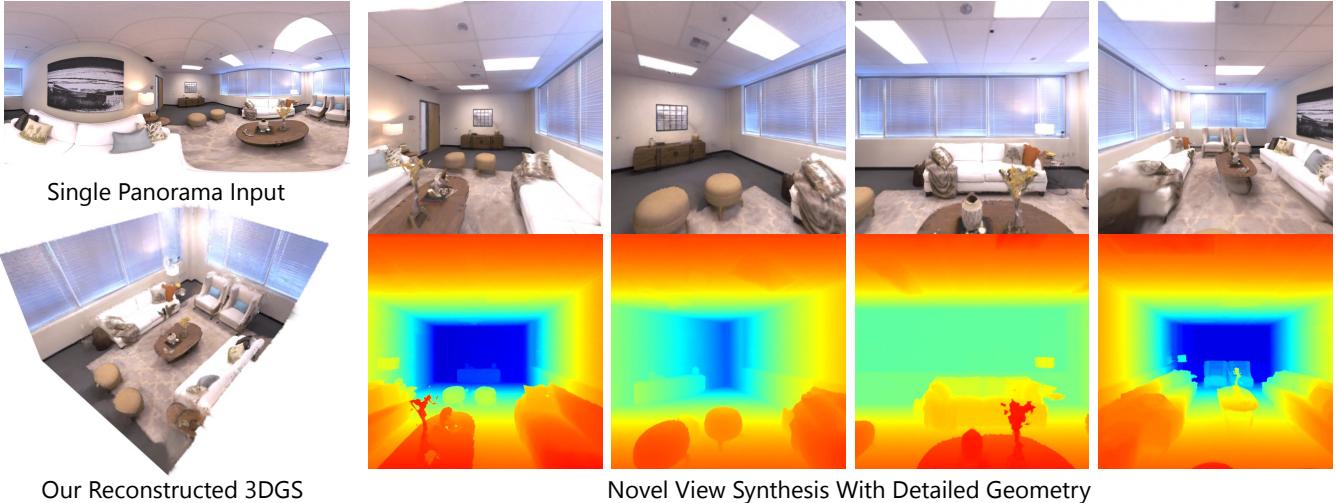


Figure 1: With a single panorama as input, the proposed Pano2Room automatically reconstructs the corresponding indoor scene with a 3D Gaussian Splatting field, capable of synthesizing photo-realistic novel views as well as high-quality depth maps. The panorama is generated using our panoramic RGBD inpainter based on any capture at a single location easily acquired by an average user. For better visualization of the 3D scene in all figures, Gaussian points or mesh blocking the room interior are deleted.

ABSTRACT

Recent single-view 3D generative methods have made significant advancements by leveraging knowledge distilled from extensive 3D object datasets. However, challenges persist in the synthesis of 3D scenes from a single view, primarily due to the complexity of real-world environments and the limited availability of high-quality prior resources. In this paper, we introduce a novel approach called Pano2Room, designed to automatically reconstruct high-quality 3D indoor scenes from a single panoramic image. These panoramic images can be easily generated using a panoramic RGBD inpainter from captures at a single location with any camera. The key idea is

to initially construct a preliminary mesh from the input panorama, and iteratively refine this mesh using a panoramic RGBD inpainter while collecting photo-realistic 3D-consistent pseudo novel views. Finally, the refined mesh is converted into a 3D Gaussian Splatting field and trained with the collected pseudo novel views. This pipeline enables the reconstruction of real-world 3D scenes, even in the presence of large occlusions, and facilitates the synthesis of photo-realistic novel views with detailed geometry. Extensive qualitative and quantitative experiments have been conducted to validate the superiority of our method in single-panorama indoor novel synthesis compared to the state-of-the-art. Our code and data are available at <https://github.com/TrickyGo/Pano2Room>.

*Corresponding author

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

SA Conference Papers '24, December 3–6, 2024, Tokyo, Japan

© 2024 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 979-8-4007-1131-2/24/12...\$15.00

<https://doi.org/10.1145/3680528.3687616>

CCS CONCEPTS

- Computing methodologies → Rendering; Image-based rendering.

KEYWORDS

Image-based Rendering, image-based modeling, texture synthesis and inpainting

ACM Reference Format:

Guo Pu, Yiming Zhao, and Zhouhui Lian. 2024. Pano2Room: Novel View Synthesis from a Single Indoor Panorama. In *SIGGRAPH Asia 2024 Conference Papers (SA Conference Papers '24), December 3–6, 2024, Tokyo, Japan*. ACM, New York, NY, USA, 11 pages. <https://doi.org/10.1145/3680528.3687616>

1 INTRODUCTION

Generating immersive 3D experiences is a crucial task in computer graphics, offering extensive practical applications in fields like Augmented Reality and Virtual Reality. While recent multi-view 3D reconstruction methods such as 3DGS (3D Gaussian Splatting) [Kerbl et al. 2023], NeuralRecon [Sun et al. 2021], and GFLF [Yin et al. 2023] have made significant progress in creating high-quality real-world scenes, they typically require a large number of images to generate accurate 3D representations for complex indoor scenes. The reliance on extensive image collection processes can be both costly and time-consuming, thereby limiting their wider practical applications. On the other hand, although single-view 3D generative methods such as RealFusion [Melas-Kyriazi et al. 2023], One-2-3-45 [Liu et al. 2024], and DreamGaussian [Tang et al. 2023] have made significant advancements in synthesizing 3D objects by distilling knowledge from extensive 3D object datasets, the task of single-view 3D indoor scene synthesis remains a tough challenge. This is primarily due to the complexity of real-world environments with open-vocabulary objects (objects with undefined/unrecognized semantics) and the limited availability of high-quality prior sources. In this paper, we present Pano2Room, a novel view synthesis method that leverages multiple priors to achieve the reconstruction of a complete 3DGS for indoor scenes with only a single panorama, as opposed to existing methods that rely on multi-view and multi-location captures.

Reconstructing the complete 3D scene from a single panorama presents significant challenges, as it requires interpreting the complex hidden structures within extremely limited 3D conditions offered by the single panoramic image. These challenges arise from the need to generate a 3D scene that not only faithfully preserves user-captured content and accurately reconstructs the geometry but also infers large-size occluded content with detailed 3D-consistent textures and geometry, integrating them consistently with existing scenes. Due to these challenges, existing single-view scene reconstruction methods, such as Text2Room [Höllein et al. 2023], PERF [Wang et al. 2024] and others, struggle not only to create 3D scenes that accurately preserve user-captured content but also to handle large-size occlusion scenarios.

To address these issues, we present Pano2Room, a novel view synthesis method that accurately preserves user-provided captures while generating 3D-consistent new textures and geometry within existing scenes. Specifically, we first convert the input panorama into a mesh and iteratively refine the mesh by leveraging a panoramic RGBD inpainter to generate occluded content and geometry while gradually incorporating the new content into the inpainted mesh. Finally, the inpainted mesh is converted to a 3DGS and trained with collected 3D-consistent pseudo novel views. The optimized 3DGS is capable of synthesizing photo-realistic novel views with detailed geometry.

In summary, the major contributions of this paper are threefold:

1. We introduce Pano2Room, a novel view synthesis method that generates a 3DGS from a single panorama. To the best of our

knowledge, Pano2Room is the first work capable of generating a complete 3DGS from a single panorama.

2. We propose a series of new modules specifically designed to improve performance and handle large-size occlusions, including a Pano2Mesh module with improved mesh filtering, a panoramic RGBD inpainter with improved inpainting quality and surface geometry, an iterative mesh refinement module with camera searching and geometry conflict avoidance strategy, and a Mesh2GS module to boost novel view synthesis quality.

3. We extensively evaluate the proposed Pano2Room on various challenging datasets, validating the state-of-the-art novel view synthesis quality in the single-panorama novel view synthesis task.

2 RELATED WORK

2.1 Single Image Novel View Synthesis

Single-image novel view synthesis approaches such as SynSin [Wiles et al. 2020], PNVS [Xu et al. 2021], InfiniteNature [Liu et al. 2021] and InfiniteNature-Zero [Li et al. 2022] address the challenge of single image novel view synthesis without interpreting each scene with explicit 3D representations, thus lacking 3D-consistency in the synthesized novel views. Incorporating explicit 3D representations, layer-based methods like SLIDE [Jampani et al. 2021], 3D-Photography [Shih et al. 2020], AdaMPI [Han et al. 2022] and SinMPI [Pu et al. 2023] represent a 3D scene using discrete layers, allowing them to generate high-quality synthesis results from a single input image. However, these methods are unsuitable for 360-degree scenes due to the structural limitations of the layered representations. Recent notable advancements in single-image 3D object generation, such as RealFusion [Melas-Kyriazi et al. 2023], SyncDreamer [Liu et al. 2023], DreamGaussian [Tang et al. 2023], and One-2-3-45 [Liu et al. 2024] distill knowledge from large-scale 3D object datasets, but they are not applicable to indoor scenes due to the complexity of real-world environments with numerous open-vocabulary objects.

Approaches based on NeRF [Mildenhall et al. 2021] and 3DGS [Kerbl et al. 2023] have demonstrated remarkable results in rendering novel views. Several single-view novel view synthesis methods have emerged, including NerfDiff [Gu et al. 2023], NerDi [Deng et al. 2023], PixelNerf [Cai et al. 2022], DietNeRF [Jain et al. 2021], Pix2NeRF [Cai et al. 2022], SinNeRF [Xu et al. 2022], OmniNeRF [Hsu et al. 2021], PERF [Wang et al. 2024], PixelSplat [Charatan et al. 2024], RealmDreamer [Shriram et al. 2024] and LucidDreamer [Chung et al. 2023]. Currently, PERF [Wang et al. 2024] stands as the state-of-the-art single-panorama novel view synthesis method. PERF trains a NeRF with novel views synthesized through a collaborative RGBD inpainting method, enabling high-quality NeRF rendering. However, due to the under-fitting of the NeRF with very few training views and dense sequential inpainting camera trajectory with possible undesired inpainting context and error accumulation, occluded areas in the novel views of PERF are prone to over-smoothed textures and meaningless geometry. Based on a single image or text prompts, LucidDreamer [Chung et al. 2023] trains a 3DGS by generating pseudo novel views through point cloud rendering. However, unlike meshes, point cloud rendering lacks surfaces, resulting in 3D-inconsistent pseudo novel views and heavy ghost artifacts in

the trained 3DGS, hence failing to apply to indoor scenarios with occlusions.

2.2 Indoor Novel View Synthesis

NeuralRoom [Yang et al. 2022] learns prior knowledge of objects via an offline stage and then synthesizes the room with unseen furniture arrangement. RoomDreamer [Song et al. 2023] utilizes diffusion models to edit a given mesh based on prompts. Control-Room3D [Schult et al. 2024] and CtrlRoom [Fang et al. 2023] create 3D scene meshes based on semantic layout proxies. While they are capable of creating virtual scenes, they do not reconstruct real scenes from user captures. These methods cannot be directly applied with only a panorama available.

From a single panorama, Auto3DIIndoor [Yang et al. 2018] utilizes geometric and semantic cues to recover indoor room layouts and typical indoor objects. DeepPanoContext [Zhang et al. 2021] estimates layouts and object poses, then reconstructs the scene with objects. Pano2CAD [Xu et al. 2017] estimates the geometry of a room and the 3D poses of objects from a single panorama. However, in real-world captures, open-vocabulary object detection and generation are challenging and error-prone, leading to these methods to fail in reconstructing complete real-world scenes with occlusions.

Text2Room [Höllein et al. 2023] and RGBD2 [Lei et al. 2023] generate 3D scene meshes based on a single input image. Text2Room generates new textures utilizing SD (Stable Diffusion) [Rombach et al. 2021] and predicts new geometries with IronDepth [Bae et al. 2022] in an iterative process to create 3D meshes of rooms. However, the linear depth alignment in Text2Room can result in cracks in surfaces, further leading to error accumulation and poor geometry. Additionally, the edge length filter proposed in Text2Room can unintentionally remove input textures and generate occlusion borders with noticeable irregularities. The reliance on Poisson surface reconstructions [Kazhdan et al. 2006] results in over-smoothing artifacts in synthesized novel views.

3 METHOD

With a single panorama as input, Pano2Room reconstructs the corresponding 3DGS Scene and enables high-quality novel view synthesis. The overview of Pano2Room is shown in Fig. 2. Given the input panoramic image I and its depth map D (predicted or converted from Cubemaps depth captured by standard RGBD cameras or Lidars), we first generate the initial mesh $Mesh_{init}$ using a Pano2Mesh module (Sec. 3.1). Subsequently, we iteratively refine the mesh into an inpainted mesh $Mesh_{inp}$ (Sec. 3.2). Finally the inpainted mesh is converted to a 3DGS GS (Sec. 3.3).

Concretely, D is captured or predicted by our panoramic RGBD inpainter (Sec. 3.2.1). During the iterative mesh refinement, we search iteratively for optimal camera viewpoints (3.2.2), generate occluded texture and geometry using the panoramic RGBD inpainter (Sec. 3.2.1), and we apply geometry conflict avoidance (Sec. 3.2.3) for generated content while collecting pseudo novel views. Following initialization, GS is optimized with the collected pseudo novel views.

3.1 Pano to Mesh

Generating 3D-consistent novel views from a single image necessitates dense and coherent surface rendering, challenging for point clouds, NeRF, or 3DGS, which demand equally dense captures and modeling. Inspired by Text2Room [Höllein et al. 2023], we generate the mesh of the input panorama by first performing triangulation among neighboring pixels in the image space and back-projecting the pixels to the 3D space. In this manner, we can faithfully preserve the content of the input image and enable novel view synthesis with 3D-consistent surfaces.

Specifically, we convert the input panorama I into an initial mesh $Mesh_{init} = \{V, C, F\}$ where V, C, F represent vertices, colors and faces of the mesh, respectively. As depicted in Fig. 3, we initialize V and C using the screen coordinates (u, v) and corresponding colors from the input panorama, respectively. Subsequently, we triangulate V based on (u, v) , connecting every four neighboring vertices in a grid to create two faces. This triangulation process yields a mesh connecting all adjacent points.

To disconnect vertices belonging to different objects and eliminate faces with excessive stretching, Text2Room [Höllein et al. 2023] utilizes edge length thresholds to remove F with edge lengths surpassing a predefined threshold. However, notice that the scale of the faces is proportional to the depth due to the back-projection from image to 3D space, this strategy tends to unintentionally eliminate content from the input panorama or the inpainted content situated at a large distance and also fails to disconnect objects that are close in proximity. This is because the edge length threshold is scale-dependent, whereas the triangulation process generates object surfaces with varying face scales according to their distances from the camera.

To mitigate this issue, we propose a depth edge filter that employs an edge detector to extract the depth edge mask M_D from the depth map D . Subsequently, the vertices within M_D are interpreted as the silhouettes of objects and are disjointed. This filter effectively filters out unwanted faces in a scale-invariant manner without compromising the input textures. Consequently, it enhances the smoothness of edges in occluded regions, leading to significant improvements in the results of the image inpainter. Illustrations demonstrating the effectiveness of the depth edge filter are provided in the supplemental materials.

Following mesh filtering in the image space, we convert the vertices from the image space (u, v) to the 3D space (x, y, z) based on the depth map D using spherical projection:

$$\begin{aligned}\phi &= v/H \cdot \pi, \theta = u/W \cdot 2\pi - 0.5\pi, \\ x &= \sin(\phi) \cdot \cos(\theta) \cdot D(\phi, \theta), \\ y &= \cos(\phi) \cdot D(\phi, \theta), \\ z &= \sin(\phi) \cdot \sin(\theta) \cdot D(\phi, \theta),\end{aligned}\tag{1}$$

where W and H represent the width and height of I respectively, and (θ, ϕ) denote the polar coordinates of (u, v) , respectively. Through the above projection, we obtain an initial mesh $Mesh_{init}$.

3.2 Iterative Mesh Completion

We propose an iterative refinement method for the initial mesh $Mesh_{init}$ by integrating the generated occluded content into the

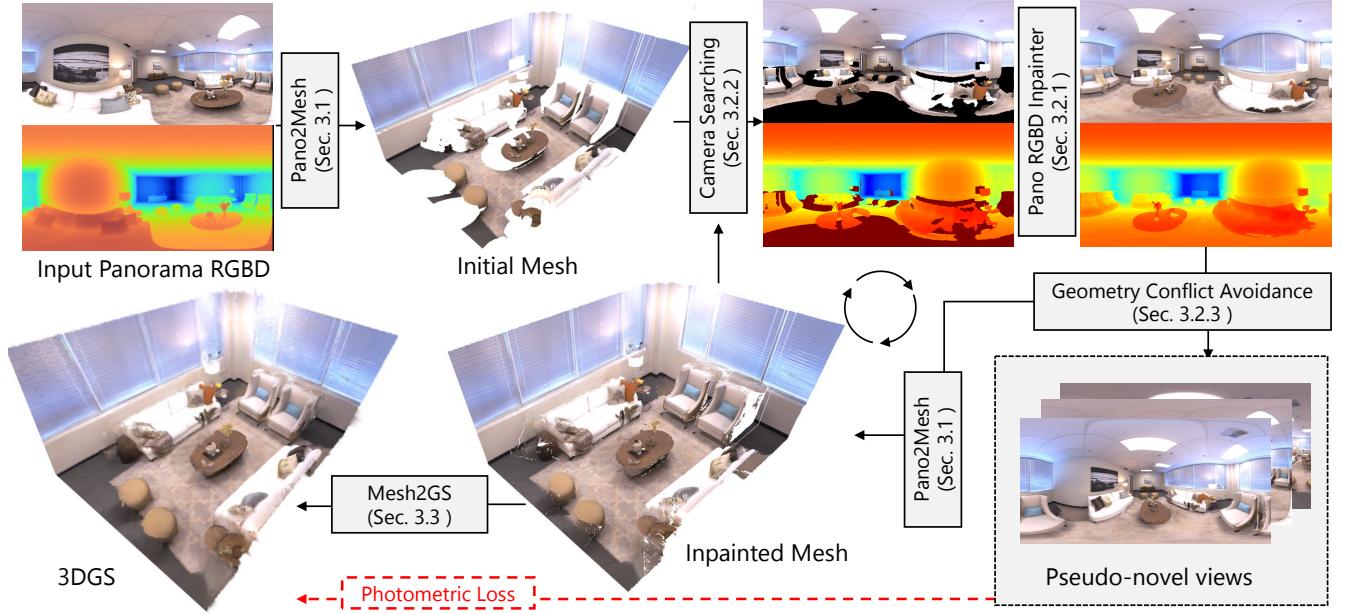


Figure 2: An overview of Pano2Room. With a panorama as input, we first predict the geometry of the panorama using the panoramic RGBD inpainter. Then we synthesize the initial mesh using a Pano2Mesh module. Next, we iteratively search for cameras with the least view completeness, and under the searched viewpoint, we render the existing mesh to obtain panoramic RGBDs with missing areas. To complete each rendered RGBD, we use the panoramic RGBD inpainter to generate new textures and predict new geometries. The new textures/geometries are iteratively fused into the existing mesh if no geometry conflict is introduced. Finally, the inpainted mesh is converted to a 3DGS and trained with collected pseudo novel views.

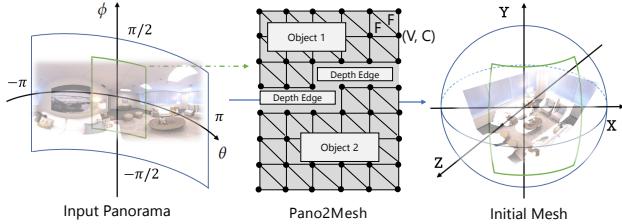


Figure 3: Demonstration of how to convert a panorama to a mesh. Initial mesh vertices and colors are derived from the input panorama’s pixels (depicted as black dots). Triangulation connects neighboring vertices to form faces (depicted by black lines). The edge map of the depth is utilized to disconnect faces representing different objects, ensuring accurate mesh generation.

existing mesh. The occluded textures are produced through image inpainting using pre-trained image generators. In the image inpainting task, having more context often aids the inpainter in generating new content that is consistent with the existing content. Since panoramas contain complete context in a single location, we sample panoramic views for inpainting instead of perspective views, which have a limited field of view and less context.

In each iteration, we start by identifying a camera viewpoint $Cams_i$ with the least view completeness (Sec. 3.2.2). We then render the existing mesh using a mesh renderer R_{Mesh} with a vertex color shader to obtain the rendered panorama image I_{render} , mask M_{render} , and depth map D_{render} . Next, we employ a panoramic RGBD inpainter to generate new textures I_{inp} and predict new geometry D_{inp} for I_{inp} . Finally, we generate a mesh for the newly generated content $I_{inp} * (1 - Mask_{render})$ using the Pano2Mesh module (Sec. 3.2) and merge this new mesh into the existing mesh. After N_{mesh} iterations, we obtain the inpainted mesh $Mesh_{inp}$. Further details are provided in the subsequent subsections.

3.2.1 Panoramic RGBD Inpainter. To synthesize the occluded content, we propose a panoramic RGBD inpainter including a panoramic image inpainter that generates high-quality textures consistent within the existing scene and a panoramic depth inpainter that predicts detailed new geometries aligned with existing surfaces.

As depicted in Fig. 4, with the rendered panoramic image I_{render} and its depth D_{render} as input, the panoramic RGBD inpainter generates the inpainted panoramic image I_{inp} and the panoramic depth inpainter predicts the inpainted depth D_{inp} .

Panoramic Image Inpainter. Existing panorama inpainting methods face limitations in generating high-quality content, particularly at high resolutions. For example, diffusion-based techniques such

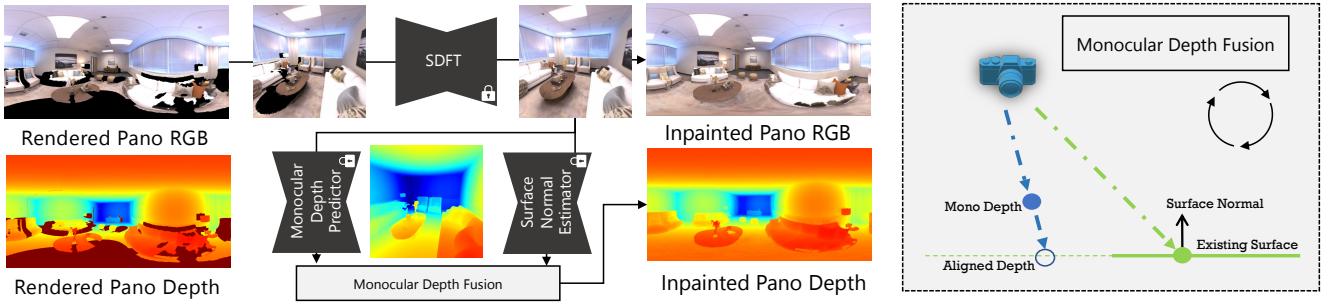


Figure 4: Panoramic RGBD Inpainter. We first inpainted the rendered panoramic image using SDFT. Then, the depth of inpainted content is estimated by a pre-trained monocular depth predictor and seamlessly fused into the rendered panoramic depth, creating inpainted panoramic depth with detailed new geometry aligned with existing geometries and enforced surface normals.

as MultiDiffusion [Bar-Tal et al. 2023] tend to produce pseudo-panoramic images that do not strictly conform to the spherical projection relationship.

To address these limitations, we inpaint I_{render} by leveraging the powerful image generation model SD (Stable Diffusion) [Rombach et al. 2021], which uses semantic masks as the conditional input. The conditional latent diffusion model is trained by minimizing:

$$L_{LDM} := \mathbb{E}_{\epsilon(x), y, \epsilon \sim N(0, 1), t} [\|\epsilon - \epsilon_\theta(z_t, t, m)\|_2^2], \quad (2)$$

where $t = 1, 2, \dots, T$ denotes the time step, z_t is the noisy version of the latent vector z of the input x , m represents the mask, ϵ is the noise schedule, and ϵ_θ denotes the time-conditioned U-Net, respectively.

As SD takes perspective images as input rather than panoramic images, we iteratively sample $k \in K$ perspective images $I_{render}(k)$ from I_{render} , employ SD to generate an inpainted perspective image $I_{inp}(k)$, then warp $I_{inp}(k)$ back to I_{render} . This iterative process ensures that each sampled $I_{render}(k)$ is conditioned on the existing panoramic image, which includes all previously inpainted content. Formally, the tangent projection images $I_{render}(k)$ is produced by the $K = 20$ faces of an icosahedron that uniformly covers the sphere's surface.

However, using perspective images as input, SD cannot capture the full style of the scene, leading to inconsistent generated new textures. To faithfully capture the style and features of the input panorama, we propose to fine-tune SD (SDFT) for each panorama. Inspired by SinMPI [Pu et al. 2023], we create pseudo inpainting pairs by employing monocular mesh construction and projection to fine-tune SD for each panorama, as depicted in Fig. 5.

In pursuit of rapid convergence and parameter efficiency, we employ Low-rank Adaptation (LoRA) [Hu et al. 2021] to fine-tune the inpainting model for each individual scene. Specifically, we maintain SD in a frozen state and only update the low-rank matrices of the self-attention layers within the U-Net during the training process. The optimization objective remains consistent with Eq. 2. We optimize the parameters of the self-attention layers due to their critical role in image inpainting tasks. Relevant ablation study results provided in supplemental materials show that the SDFT inpainter is able to capture the style and features of the input

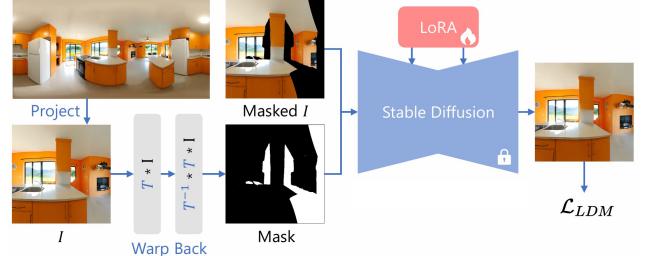


Figure 5: SDFT: Fine-tuning Stable Diffusion on the input panorama to learn the styles and features.

panorama, and generate new textures that are more consistent with the scene.

Taking advantage of the powerful image generation capabilities of SDFT, completing the inpainting process ensures a continuous and seamless inpainted panoramic image preserving the original resolution with high-quality inpainting.

Panoramic Depth Inpainter. The depth of a panorama includes visible surfaces in an indoor scene, incorporating complex textures and geometries with intricate details. Following 360MonoDepth [Rey-Area et al. 2022], to inpaint D_{inp} based on D_{render} , we first predict the depth maps $D_{inp}(k)$ from multiple perspective projects $I_{inp}(k)$ of I_{inp} using a pretrained monocular depth estimator DPT [Ranftl et al. 2021], and then fuse these $D_{inp}(k)$ to D_{render} to produce D_{inp} .

To iteratively fuse $D_{inp}(k)$ into D_{render} , we optimize the re-scale factors of $D_{inp}(k)$ to obtain a complete, smooth inpainted depth D_{inp} with detailed geometry. Specifically, D_{inp} is initialized as the pixel-wise average of the spherical projection of all $D_{inp}(k)$. For each $D_{inp}(k)$, we aim to re-scale $D_{inp}(k)$ by a scaling factor $s(k)$ and an offset $o(k)$ to obtain the optimized depth $\tilde{D}_{inp}(k) = s(k)D_{inp}(k) + o(k)$. The optimization objective is defined as:

$$\arg \min_{\{s(k), o(k)\}} E_{fix} + E_{align} + E_{normal} + E_{smooth}, \quad (3)$$

where the existing depth is fixed during depth fusion through:

$$E_{fix} = \sum_{k \in K} ((D_{render}(k) - \tilde{D}_{inp}(k)) * M_{render}(k))^2. \quad (4)$$

Depth is constrained to be consistent across all tangent views through the alignment loss:

$$E_{align} = \sum_{(a,b) \in K} \sum_{x \in \Omega(a,b)} (\tilde{D}_{inp}(a)(x) - \tilde{D}_{inp}(b)(x))^2, \quad (5)$$

where a, b denote any two \tilde{D}_{inp} with overlapping pixels x from the overlapping regions $\Omega(a, b)$. The spatial smoothness between neighboring grid-points m and n is ensured by applying the smoothness loss:

$$E_{smooth} = \sum_{k \in K} \sum_{(m,n)} \|s_k^m - s_k^n\|_2^2 + \|o_k^m - o_k^n\|_2^2. \quad (6)$$

In indoor scenes, dense and coherent surface rendering is one of the most critical factors for high-quality novel view synthesis. We propose to enforce surface normal constraints by using pre-trained surface normal estimators and regulating D_{inp} with a surface normal loss:

$$E_{normal} = \sum_{k \in K} (N_{I_{inp}}(k) - N_{\tilde{D}_{inp}}(k))^2, \quad (7)$$

where the surface normals $N_{I_{inp}}$ and $N_{\tilde{D}_{inp}}$ are estimated by pre-trained surface normal estimators [Bae et al. 2022].

After N_{depth} iterations of optimization, the monocular depth maps are seamlessly fused into the rendered panoramic depth, creating an inpainted panoramic depth D_{inp} with detailed new geometry aligned with existing geometry and enforced surface normals.

3.2.2 Searching Inpainting Viewpoints. Existing methods such as PERF [Wang et al. 2024] complete existing scenes through iterative inpainting of occluded content along pre-defined camera trajectories in a sequence. However, inpainting under any inappropriate camera viewpoint with small grazing angles between the view direction and inpainted surface normal, or conducting multiple inpainting steps for any occluded area leads to lower-quality new texture and geometry generation. This deterioration occurs because image inpainters rely on context for filling, but the accumulated errors of image inpainters and depth estimators lead to a decline in the quality of the input context, resulting in inferior filling quality, including cluttered and uneven surfaces.

Based on this observation, our objective is to minimize the number of cameras and inpainting steps required to complete a scene. In each scene, we search for viewpoints candidates as follows: First, we identify the room boundaries according to the panoramic depth map. Following the Atlanta world assumption [Schindler and Dellaert 2004], the middle line of the depth map signifies the horizontal boundary of the scene, while the distance between the ceiling and floor indicates the vertical boundaries. Subsequently, we uniformly sample in the room space to define all possible cameras. Next, we propose a camera search strategy aiming to reduce inpainting steps by identifying cameras that cover the largest areas needing completion in each iteration. Specifically, from a pre-defined set of potential camera viewpoints $Cams_k$ across the scene that covers the majority areas of the scene, we search for viewpoints with the least view completeness:

$$\arg \min_{\{k \in K\}} \sum_{(u,v) \in (H,W)} M_{render}(k)(u,v), \quad (8)$$

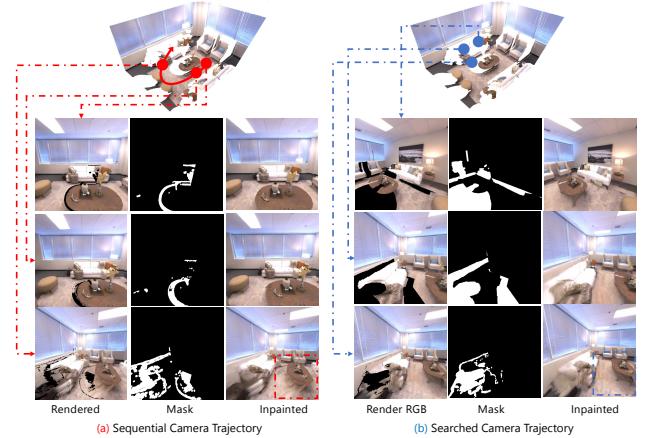


Figure 6: The effectiveness of the proposed camera search strategy. The strategy with sequential camera trajectory leads to multiple inpainting steps performed in an occluded space and produces blurry inpainting results, while our method with searched viewpoints facilitates the generation of plausible new textures and geometry.

where (u, v) denote pixel coordinates and H, W represent the height and width of M_{render} . The $Cams_k$ search space are elaborated in supplemental materials.

As illustrated in Fig. 6, the sequential camera trajectory strategy results in multiple inpainting steps in occluded spaces, yielding blurry inpainting results. In contrast, utilizing searched viewpoints, our method markedly enhances the generation performance of plausible new textures and geometry.

3.2.3 Geometry Conflict Avoidance. In each iteration of generating occluded content, newly added content may result in geometric conflicts with previously rendered viewpoints I_{render} or I_{inp} , thereby causing 3D-inconsistencies in pseudo novel views. Training 3DGs with 3D-inconsistent images can result in pronounced ghost artifacts and over-smoothing.

In order to avoid geometry conflicts, we propose a fast geometry conflict avoidance strategy based on mesh rendering. Upon fusing new mesh with the existing mesh, we conduct a render of the mesh from preceding viewpoints to detect any potential geometry conflicts. If any geometry conflicts are introduced, these new faces are abandoned. Specifically, to ensure that newly incorporated mesh does not disrupt previous views, we identify conflicting inpainted content using a mask:

$$M_{conflict}(k)(u, v) = \|I_{inp}(k)(u, v) - R_{mesh}(k)(u, v)\| > \varepsilon, \quad (9)$$

where (u, v) are the pixel coordinates and ε stands for the conflict threshold, respectively.

After the inpainting process, we run the Poisson surface reconstruction [Kazhdan et al. 2006] to close any remaining holes and obtain a watertight mesh.

3.3 Mesh to 3DGS

The Poisson surface reconstruction [Kazhdan et al. 2006] yields a complete mesh, but it tends to create over-smoothing texture and geometry, downgrading photo-realism and sharpness of rendered novel views. To address this problem, we propose to convert the mesh into a 3DGS GS and train GS with collected pseudo novel views I_{inp} to ensure that the rendering quality of GS matches the photo-realism of the images generated by our fine-tuned SD.

Concretely, we first construct the initial GS using the V and C of the mesh as the initial point cloud. Then, we train GS using perspective views $I_{inp}(k)$ of collected pseudo novel views I_{inp} . The 3DGS is rendered using point-based alpha-blending, which computes the color C of a pixel by blending N ordered points overlapping the pixel according to:

$$C = \sum_{i \in N} c_i \alpha_i \prod_{j=1}^{i-1} (1 - \alpha_j), \quad (10)$$

where c_i represents the color of each point, and α_i is given by evaluating a 2D Gaussian with covariance multiplied with a learned per-point opacity. To be specific, GS is optimized with collected pseudo novel view $I_{inp}(k)$ using photometric losses, comprising an L1 loss combined with a D-SSIM term:

$$L = L_1(R_{GS}, I_{inp}) + \lambda L_{D-SSIM}(R_{GS}, I_{inp}), \quad (11)$$

where R_{GS} denotes the 3DGS renderer.

Benefiting from our proposed iteratively mesh completion with the geometric conflict avoidance strategy, the collected pseudo novel views used for training GS are 3D consistent and photo-realistic. This results in an excellent quality of the trained GS, significantly enhancing the quality of novel view synthesis. As illustrated in Fig. 7, the optimized GS renders novel views with high quality and detailed geometry, effectively handling the over-smoothing artifacts introduced by the Poisson surface reconstruction.

4 EXPERIMENTS

4.1 Comparison with Previous Methods

4.1.1 Baselines. We compare our method with three state-of-the-art techniques: PERF [Wang et al. 2024], Text2Room [Höllein et al. 2023], and LucidDreamer [Chung et al. 2023]. PERF is a single-panorama novel-view synthesis method, Text2Room is a text-to-indoor-mesh generation method, and LucidDreamer is a 3DGS-based single-view novel view synthesis method. For fair comparison in each scene, all methods are provided with the identical input panoramic RGBD and camera trajectory to align generated novel views with corresponding ground-truth views. Since Text2Room and LucidDreamer take perspective images as input, we project the panoramic RGBD into cubemaps RGBD as their input.

4.1.2 Datasets. We evaluate all methods using the eight single-room scenes from the Replica dataset. We establish a consistent rendering camera trajectory comprising 150 poses for each method, with the cameras traversing the room and facing inward. Additionally, we also conduct comparisons on real-world captured panoramas in various indoor scenes from the Pano3D dataset [Albanis et al. 2021], the S2D3D dataset [Armeni et al. 2017], and the ZIND dataset [Cruz et al. 2021], as detailed in the supplemental materials.

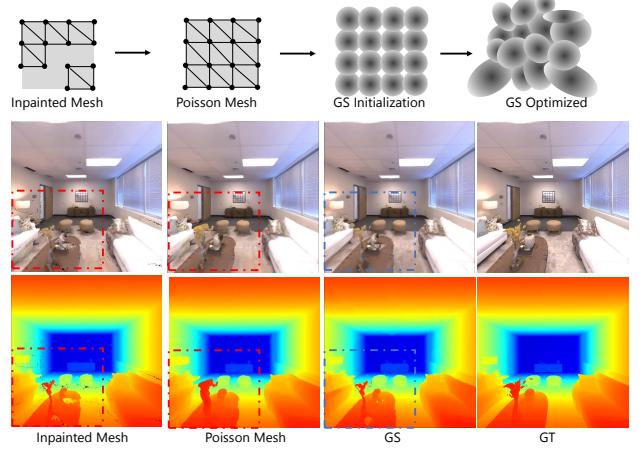


Figure 7: Demonstration of how to convert a mesh to a 3DGS. The optimized 3DGS renders novel views with high-quality and detailed geometry, effectively addressing the over-smoothing artifacts introduced by the Poisson surface reconstruction. Please zoom in for a better inspection.

4.1.3 Evaluation Metrics. We assess the reconstruction quality of the rendered views produced by each method, evaluating both the fidelity of how faithfully the method preserves the information of the input panorama and the image quality of synthesized novel views. The rendered views of each method are compared with corresponding ground-truth images using metrics including Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index (SSIM), and Learned Perceptual Image Patch Similarity (LPIPS).

4.1.4 Qualitative Comparison. Fig. 8 shows the qualitative comparison of rendered novel views. PERF gradually collects pseudo novel-views around the original location, making it challenging to inpaint large areas with appropriate camera positions and contextual information. The synthesized novel views of PERF are prone to over-smoothed textures and meaningless interpolated geometry, especially in occluded areas. This is caused by the under-fitting of NeRF with very few training views and dense sequential inpainting camera trajectory with possible undesired inpainting context and error accumulation. LucidDreamer is prone to strong artifacts due to the fact that point cloud rendering lacks surfaces which means when the camera is close to surfaces, the gaps in the point cloud become apparent. This results in 3D-inconsistent collected novel views for the training of 3DGS, leading to heavy ghost artifacts in the trained 3DGS. Text2Room suffers from the artifacts brought by Poisson Mesh Reconstruction and undesired inpainting results. Our novel views demonstrate superior photorealism, exhibiting high-quality texture, geometry, and structural fidelity. Please refer to the supplemental materials and accompanying video for more comprehensive analysis and visualizations.

4.1.5 Quantitative Comparison. Table 1 presents the quantitative evaluation metrics on the Replica dataset. Consistent with the visual comparisons, Pano2Room achieves the best performance across most scenes in terms of both reconstruction and novel view quality.

Table 1: Quantitative comparison on the Replica dataset.

Scene Method	Pano2Room			PERF			Text2Room			LucidDreamer		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
Room 0	23.29	0.788	0.168	22.58	0.767	0.167	21.55	0.779	0.151	17.50	0.611	0.470
Room 1	25.44	0.886	0.090	26.10	0.854	0.099	24.92	0.884	0.079	18.03	0.681	0.484
Room 2	25.06	0.874	0.127	23.75	0.858	0.121	22.21	0.862	0.138	17.40	0.719	0.456
Office 0	25.15	0.891	0.105	21.15	0.871	0.124	21.68	0.863	0.126	19.25	0.746	0.418
Office 1	31.77	0.953	0.043	30.98	0.938	0.070	23.92	0.935	0.078	20.91	0.788	0.332
Office 2	23.22	0.899	0.087	20.83	0.874	0.098	20.37	0.885	0.090	14.39	0.678	0.450
Office 3	20.93	0.858	0.143	19.69	0.825	0.159	15.77	0.789	0.232	13.98	0.679	0.446
Office 4	26.65	0.935	0.064	22.64	0.904	0.086	23.53	0.922	0.065	15.79	0.743	0.415

metrics, validating its state-of-the-art capability in single-panorama novel view synthesis for indoor scenes. Our method takes approximately 40 minutes to generate a complete 3DGS scene based on a single panorama that can be rendered at 156 FPS, outperforming the state-of-the-art single-panorama novel view synthesis method, PERF. Results of efficiency comparison can be found in supplemental materials.

4.2 Ablation Study

Fig. 9 illustrates the ablation study results of our method conducted on the representative Room-0 scene from the Replica dataset featuring 150 rendered views and the average PSNR score across all views. We evaluate our method without each key component namely Stable Diffusion finetuning (SDFT), Surface Normal Constraint (SN), Camera Searching strategy (CS), Geometry Conflict Avoidance strategy (GCA), and Mesh2GS module (GS). More specifically, (1) **w/o SDFT**: Employing Stable Diffusion without finetuning, resulting in increased variance in the generated content. This variance leads to a decrease in reconstruction scores due to lower consistency in the generated styles. (2) **w/o SN**: Without surface normal constraints, the quality of surface geometry completion decreases, resulting in cracks in surfaces. (3) **w/o CS**: Using sequential poses in iterative inpainting causes large occlusion areas to be inpainted from inappropriate camera poses multiple times, leading to low-quality inpainted texture due to error accumulation. (4) **w/o GCA**: Without geometry conflict avoidance, geometry conflicts occur where the inpainted new content blocks the user-captured content and previous inpainted content, causing floater artifacts. (5) **w/o GS**: The rendering results of the initial 3D mesh are over-smoothed, resulting in the loss of many details from user captures and low-quality novel view synthesis.

5 IMPLEMENTATION DETAILS

We performed all experiments utilizing an Nvidia A40 GPU. The mesh completion iteration N_{mesh} is adaptive and depends on the occlusion size of the particular scene, typically falling within the range $2 < N_{mesh} < 16$. The monocular depth fusion iterations are set to $N_{depth} = 3000$. The geometry conflict threshold is set to $\epsilon = 5$. The process of generating the final 3DGS field from a panorama typically takes approximately 40 minutes. This includes a 20-minute per-scene fine-tuning phase of Stable-Diffusion, 10

minutes for panorama mesh generation, and 10 minutes for 10000 iterations of 3DGS optimization.

6 LIMITATIONS

Our method comprises several consecutive steps utilizing multiple pre-trained models, each producing intermediate results that may contain errors. However, there are no mechanisms in place within subsequent steps to correct these errors, leading to error accumulation. For instance, during the depth estimation step, monocular depth estimators encounter challenges with highly reflective and transmissive materials such as glass, and these errors propagate through subsequent steps. We rely on pre-trained monocular depth estimators, but increased distance can degrade depth quality. In large indoor settings like long hallways, distant objects will lack detailed geometry. The mesh is iteratively constructed by back-projecting image-space grid mesh to world space, which means the mesh projected in distance will be sparse with blurry textures when closely inspected. In addition, this paper assumes that the input panorama captures fundamental room structures to identify camera search space.

7 CONCLUSION

In this paper, we proposed Pano2Room that generates a high-quality 3DGS scene from a single panorama. To achieve this goal, we designed several new modules including a Pano2Mesh module for constructing a panorama's mesh, a panoramic RGBD inpainter designed to generate the occluded content of a scene, an iterative mesh refinement module with camera searching and geometry conflict avoidance strategy to enhance the quality of inpainting, and a Mesh2GS module to boost the quality of novel view synthesis. Through extensive evaluations on various panorama datasets, we demonstrated that Pano2Room achieves state-of-the-art reconstruction quality in single-panorama novel view synthesis.

ACKNOWLEDGMENTS

This work was supported by National Natural Science Foundation of China (Grant No.: 62372015), Center For Chinese Font Design and Research, Key Laboratory of Intelligent Press Media Technology, and State Key Laboratory of General Artificial Intelligence.

REFERENCES

- Georgios Albanis, Nikolaos Zioulis, Petros Drakoulis, Vasileios Gkitsas, Vladimiros Sterzentsenko, Federico Alvarez, Dimitrios Zarpalas, and Petros Daras. 2021. Pano3D: A holistic benchmark and a solid baseline for 360deg depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3727–3737.
- I. Armeni, A. Sax, A. R. Zamir, and S. Savarese. 2017. Joint 2D-3D-Semantic Data for Indoor Scene Understanding. *ArXiv e-prints* (Feb. 2017). arXiv:1702.01105 [cs.CV]
- Gwangbin Bae, Ignas Budvytis, and Roberto Cipolla. 2022. IronDepth: Iterative Refinement of Single-View Depth using Surface Normal and its Uncertainty. In *British Machine Vision Conference (BMVC)*.
- Omer Bar-Tal, Lior Yaniv, Yaron Lipman, and Tali Dekel. 2023. MultiDiffusion: Fusing diffusion paths for controlled image generation. (2023).
- Shengqu Cai, Anton Obukhov, Dengxin Dai, and Luc Van Gool. 2022. Pix2NeRF: Unsupervised Conditional p-GAN for Single Image to Neural Radiance Fields Translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3981–3990.
- David Charanat, Sizhe Lester Li, Andrea Tagliasacchi, and Vincent Sitzmann. 2024. PixelSplat: 3d gaussian splats from image pairs for scalable generalizable 3d reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 19457–19467.
- Jae young Chung, Suyoung Lee, Hyeyongjin Nam, Jaerin Lee, and Kyoung Mu Lee. 2023. LucidDreamer: Domain-free generation of 3d gaussian splatting scenes. *arXiv preprint arXiv:2311.13384* (2023).
- Steve Cruz, Will Hutchcroft, Yuguang Li, Naji Khosravan, Ivaylo Boyadzhiev, and Sing Bing Kang. 2021. Zillow Indoor Dataset: Annotated Floor Plans With 360° Panoramas and 3D Room Layouts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2133–2143.
- Congyu Deng, Chiyu Jiang, Charles R Qi, Xincheng Yan, Yin Zhou, Leonidas Guibas, Dragomir Anguelov, et al. 2023. NerDi: Single-view nerf synthesis with language-guided diffusion as general image priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 20637–20647.
- Chuan Fang, Xiaotao Hu, Kunming Luo, and Ping Tan. 2023. Ctrl-Room: Controllable Text-to-3D Room Meshes Generation with Layout Constraints. *arXiv preprint arXiv:2310.03602* (2023).
- Jiatao Gu, Alex Trevithick, Kai-En Lin, Joshua M Susskind, Christian Theobalt, Lingjie Liu, and Ravi Ramamoorthi. 2023. NerfDiff: Single-image view synthesis with nerf-guided distillation from 3d-aware diffusion. In *International Conference on Machine Learning*. PMLR, 11808–11826.
- Yuxuan Han, Ruicheng Wang, and Jiaolong Yang. 2022. Single-view view synthesis in the wild with learned adaptive multiplane images. In *ACM SIGGRAPH 2022 Conference Proceedings*. 1–8.
- Lukas Höllerin, Ang Cao, Andrew Owens, Justin Johnson, and Matthias Nießner. 2023. Text2Room: Extracting textured 3d meshes from 2d text-to-image models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 7909–7920.
- Ching-Yu Hsu, Cheng Sun, and Hwann-Tzong Chen. 2021. Moving in a 360 world: Synthesizing panoramic parallaxes from a single panorama. *arXiv preprint arXiv:2106.10859* (2021).
- Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685* (2021).
- Ajay Jain, Matthew Tancik, and Pieter Abbeel. 2021. Putting nerf on a diet: Semantically consistent few-shot view synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 5885–5894.
- Varun Jampani, Huiwen Chang, Kyle Sargent, Abhishek Kar, Richard Tucker, Michael Krainin, Dominik Kaeser, William T Freeman, David Salesin, Brian Curless, et al. 2021. SLIDE: Single image 3d photography with soft layering and depth-aware inpainting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 12518–12527.
- Michael Kazhdan, Matthew Bolitho, and Hugues Hoppe. 2006. Poisson surface reconstruction. In *Proceedings of the fourth Eurographics symposium on Geometry processing*, Vol. 7. 0.
- Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 2023. 3D Gaussian Splatting for Real-Time Radiance Field Rendering. *ACM Transactions on Graphics* 42, 4 (2023).
- Jiabao Lei, Jiapeng Tang, and Kui Jia. 2023. RGBD2: Generative Scene Synthesis via Incremental View Inpainting Using RGBD Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8422–8434.
- Zhengqi Li, Qianqian Wang, Noah Snavely, and Angjoo Kanazawa. 2022. InfiniteNature-Zero: Learning Perpetual View Generation of Natural Scenes from Single Images. In *European Conference on Computer Vision*. Springer, 515–534.
- Andrew Liu, Richard Tucker, Varun Jampani, Ameesh Makadia, Noah Snavely, and Angjoo Kanazawa. 2021. Infinite nature: Perpetual view generation of natural scenes from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 14458–14467.
- Minghua Liu, Chao Xu, Haian Jin, Linghao Chen, Mukund Varma T, Zexiang Xu, and Hao Su. 2024. One-2-3-45: Any single image to 3d mesh in 45 seconds without per-shape optimization. *Advances in Neural Information Processing Systems* 36 (2024).
- Yuan Liu, Cheng Lin, Zijiao Zeng, Xiaoxiao Long, Lingjie Liu, Taku Komura, and Wenping Wang. 2023. SyncDreamer: Generating Multiview-consistent Images from a Single-view Image. *arXiv preprint arXiv:2309.03453* (2023).
- Luke Melas-Kyriazi, Iro Laina, Christian Rupprecht, and Andrea Vedaldi. 2023. RealFusion: 360deg reconstruction of any object from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8446–8455.
- Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. 2021. Nerf: Representing scenes as neural radiance fields for view synthesis. *Commun. ACM* 65, 1 (2021), 99–106.
- Guo Pu, Peng-Shuai Wang, and Zhouhui Lian. 2023. SinMPI: Novel View Synthesis from a Single Image with Expanded Multiplane Images. In *SIGGRAPH Asia 2023 Conference Papers*. 1–10.
- René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. 2021. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 12179–12188.
- Manuel Rey-Area, Mingze Yuan, and Christian Richardt. 2022. 360MonoDepth: High-resolution 360deg monocular depth estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3762–3772.
- Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2021. High-Resolution Image Synthesis with Latent Diffusion Models. *arXiv:2112.10752* [cs.CV]
- Grant Schindler and Frank Dellaert. 2004. Atlanta world: An expectation maximization framework for simultaneous low-level edge grouping and camera calibration in complex man-made environments. In *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004*, Vol. 1. IEEE, I–I.
- Jonas Schult, Sam Tsai, Lukas Höllerin, Bichen Wu, Jialiang Wang, Chih-Yao Ma, Kunpeng Li, Xiaofang Wang, Felix Wimbauer, Zijian He, et al. 2024. ControlRoom3D: Room generation using semantic proxy rooms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6201–6210.
- Meng-Li Shih, Shih-Yang Su, Johannes Kopf, and Jia-Bin Huang. 2020. 3D photography using context-aware layered depth inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8028–8038.
- Jaidev Shriram, Alex Trevithick, Lingjie Liu, and Ravi Ramamoorthi. 2024. Realm-Dreamer: Text-driven 3d scene generation with inpainting and depth diffusion. *arXiv preprint arXiv:2404.07199* (2024).
- Liangchen Song, Liangliang Cao, Hongyu Xu, Kai Kang, Feng Tang, Junsong Yuan, and Zhao Yang. 2023. RoomDreamer: Text-Driven 3D Indoor Scene Synthesis with Coherent Geometry and Texture. In *Proceedings of the 31st ACM International Conference on Multimedia*. 6898–6906.
- Jiaming Sun, Yiming Xie, Linghao Chen, Xiaowei Zhou, and Hujun Bao. 2021. NeuralRecon: Real-time coherent 3D reconstruction from monocular video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 15598–15607.
- Jiaxiang Tang, Jiawei Ren, Hang Zhou, Ziwei Liu, and Gang Zeng. 2023. DreamGaussian: Generative gaussian splatting for efficient 3d content creation. *arXiv preprint arXiv:2309.16653* (2023).
- Guangcong Wang, Peng Wang, Zhaoxi Chen, Wenping Wang, Chen Change Loy, and Ziwei Liu. 2024. PERF: Panoramic Neural Radiance Field from a Single Panorama. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* (2024).
- Olivia Wiles, Georgia Gkioxari, Richard Szeliski, and Justin Johnson. 2020. SynSin: End-to-end view synthesis from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7467–7477.
- Dejia Xu, Yifan Jiang, Peihao Wang, Zhiwen Fan, Humphrey Shi, and Zhangyang Wang. 2022. SinNerf: Training neural radiance fields on complex scenes from a single image. In *European Conference on Computer Vision*. Springer, 736–753.
- Jiu Xu, Björn Stenger, Tommi Kerola, and Tony Tung. 2017. Pano2CAD: Room layout from a single panorama image. In *2017 IEEE winter conference on applications of computer vision (WACV)*. IEEE, 354–362.
- Jiale Xu, Jia Zheng, Yanyu Xu, Rui Tang, and Shenghua Gao. 2021. Layout-guided novel view synthesis from a single indoor panorama. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 16438–16447.
- Bangbang Yang, Yinda Zhang, Yijin Li, Zhaopeng Cui, Sean Fanello, Hujun Bao, and Guofeng Zhang. 2022. Neural rendering in a room: amodal 3d understanding and free-viewpoint rendering for the closed scene composed of pre-captured objects. *ACM Transactions on Graphics (TOG)* 41, 4 (2022), 1–10.
- Yang Yang, Shi Jin, Ruiyang Liu, Sing Bing Kang, and Jingyi Yu. 2018. Automatic 3d indoor scene modeling from single panorama. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3926–3934.
- Ruihong Yin, Sezer Karaoglu, and Theo Gevers. 2023. Geometry-guided Feature Learning and Fusion for Indoor Scene Reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 3652–3661.
- Cheng Zhang, Zhaopeng Cui, Cai Chen, Shuaicheng Liu, Bing Zeng, Hujun Bao, and Yinda Zhang. 2021. DeepPanoContext: Panoramic 3d scene understanding with holistic scene context graph and relation-based optimization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 12632–12641.

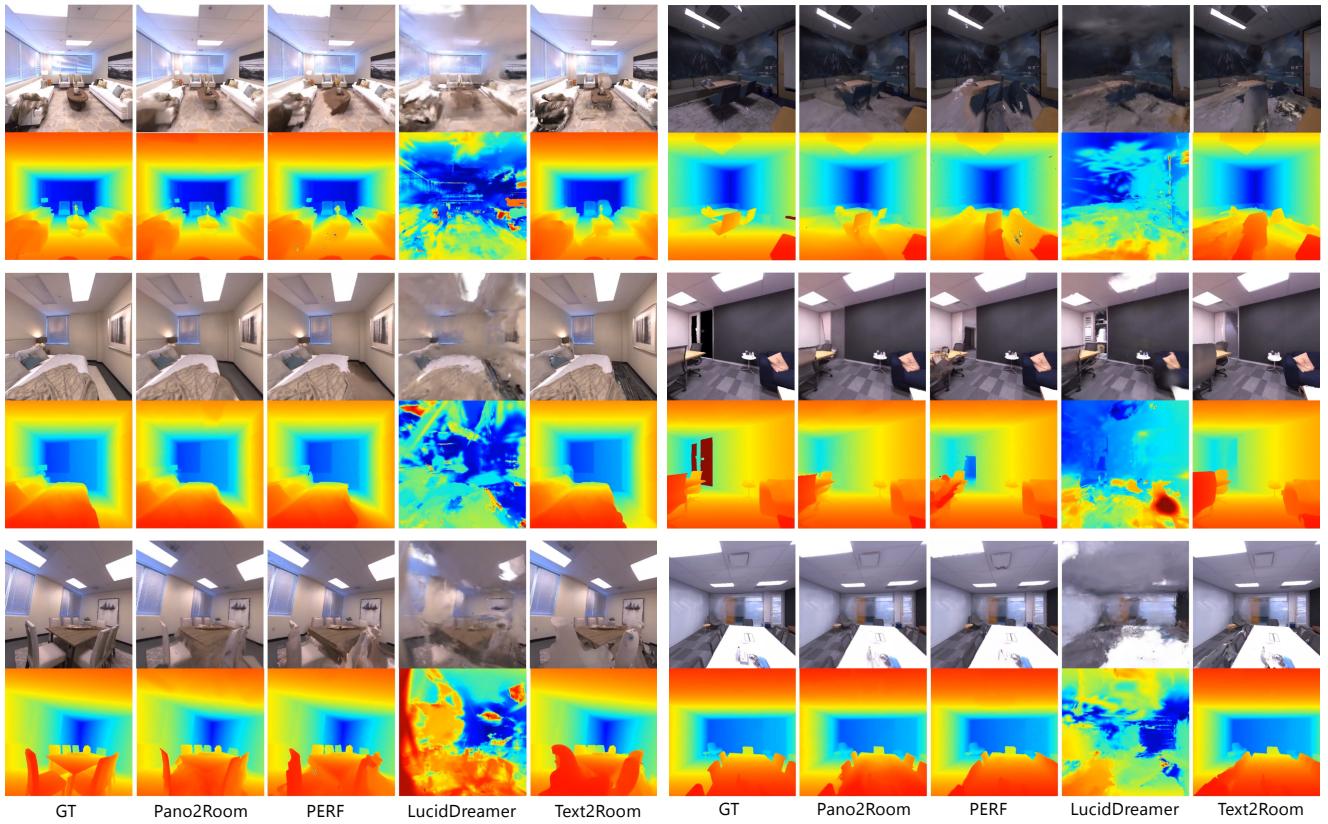


Figure 8: Comparison of novel view synthesis and depth map of each method with the corresponding ground truth. Novel views synthesized by Text2Room and LucidDreamer are prone to strong artifacts. PERF tends to generate over-smoothed occluded areas with interpolated geometries. Our novel views are more photorealistic, considering texture, geometry, and structure. Please zoom in for a better inspection.

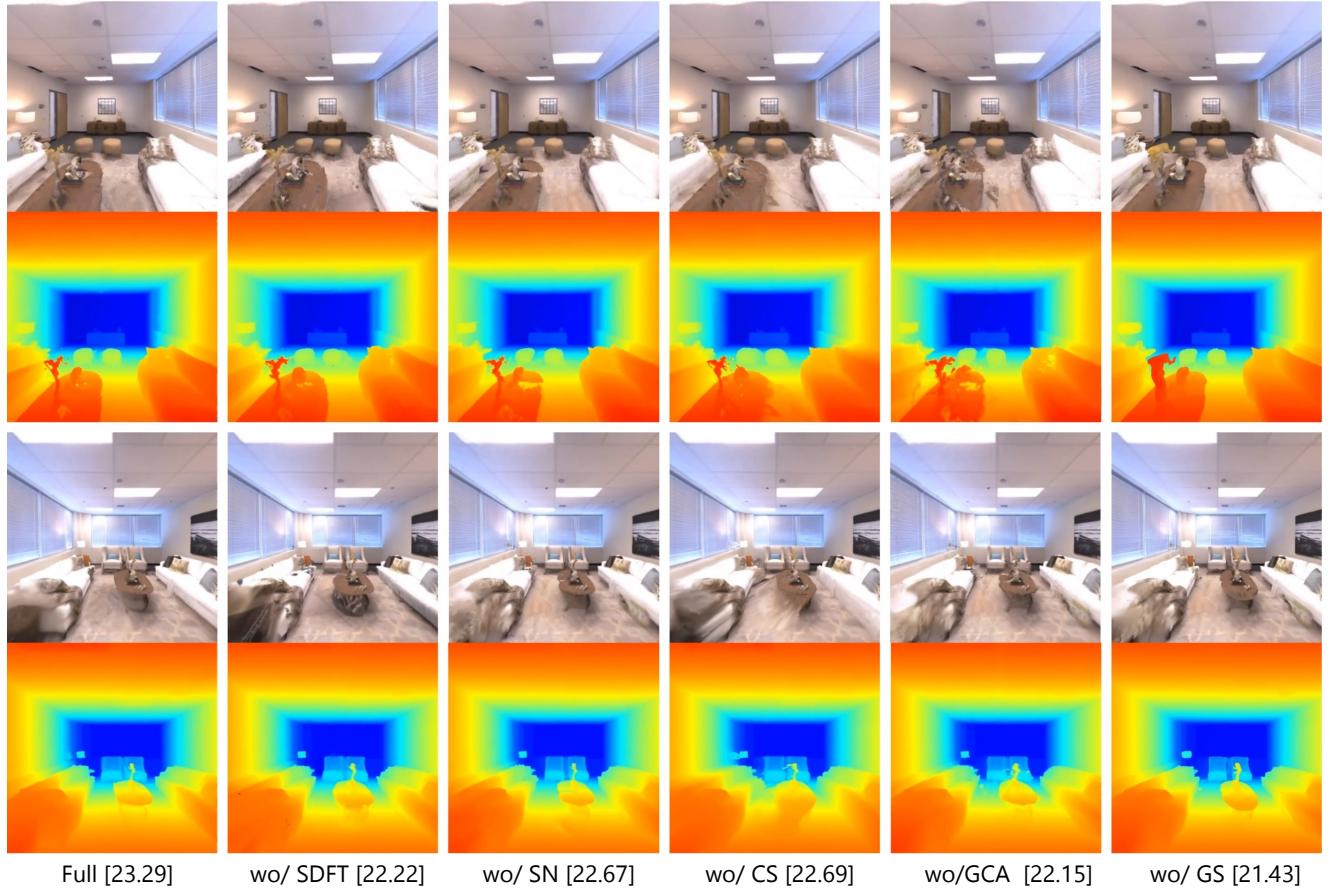


Figure 9: Ablation study results conducted on the representative Room-0 scene from the Replica dataset featuring 150 rendered views and the average PSNR score across all views. We evaluate our method without each key component including Stable Diffusion finetuning (SDFT), Surface Normal Constraints (SN), Camera Searching strategy (CS), Geometry Conflict Avoidance strategy (GCA), and Mesh2GS module (GS). The corresponding PSNR Scores are provided within the square blankets for quantitative comparison. Please zoom in for a better inspection.