GenRC: Generative 3D Room Completion from Sparse Image Collections

Supplementary Materials

1 Datasets and Metrics

1.1 Data Preprocessing

ScanNet. To prepare data from the ScanNet [4] dataset, we follow the same experiment setting shown in [9], which provides 18 scenes from ScanNet with RGB images and depth maps both in the resolution of 240×180. Specifically, we uniformly sample 18 scenes whose number of views is more than 50 from the testing set following [9].
ArkitSences. Similarly, we prepare the ArkitSences [2] dataset by preprocessing the RGB images and depth maps into the resolution of 240×180 as the same experiment setting shown in [9]. Then, we uniformly sample 100 views per scene, so the number of percentages in the experiments will be the number of views we used for initializing a mesh (e.g., 5% means 5 views). For sampling the scenes for evaluation, we consider the official testing set which contains 549 scenes, and select the 189 scenes that were captured without hand-held camera rotation. Finally, we uniformly sample 20 scenes from them for evaluation.

1.2 Metric Calculation

We follow [9] to compute evaluation metrics shown on the tables. For each scene, a certain percentage of images (5%, 10%, 20%, or 50%) will be uniformly sampled for all the images from this scene as input sparse observations, and the rest of them will be used as testing images for computing metrics. Note that the in-order camera poses of testing images will serve as prefined camera trajectories for two baseline methods to inpaint their meshes. Even if there exist holes on the generated meshes of T2R+RGBD and RGBD2, these holes are outside the trajectories which won't influence the quantitative results.

1.3 Geometric Metrics

We evaluate the geometric quality by comparing depth renderings from the generated meshes with ground-truth depth maps in the testing sets (see Sec. 4.1 in the main paper). In addition, the depth MSE measures the geometric consistency of a scene since ground-truth depth maps come from the same scene and should be geometry-consistent. In [9], bi-directional Chamfer Distance is used as an alternative for evaluating the geometric quality. However, we observed that the ground-truth mesh generated by back-projecting ground-truth images may not

Table 1: Additional quantitative results of geometry quality on ScanNet. We report one-directional Chamfer Distance (one-directional CD), which shows the distance between each point in the ground-truth mesh and its nearest neighbor in the generated mesh. GenRC can generate more geometrically correct meshes when sparse observations are given (i.e., 5%, 10%, 20%).

Methods	One-directional CD(↓)			
	5%	10%	20%	50%
T2R+RGBD [7]	0.091	0.031	0.018	0.014
RGBD2 [9]	0.073	**0.018**	0.011	**0.005**
Ours	**0.050**	**0.018**	**0.009**	0.007

capture the whole scene. As a result, the bi-directional Chamfer Distance metric will also penalize our method when our generated meshes are more complete. Hence, we report one-directional Chamfer Distance (the distance between each point in the ground-truth mesh and its nearest neighbor in the generated mesh) in Tab. 1. Our method outperforms other state-of-the-art methods, especially when input observations are sparse.

1.4 Baseline Implementation

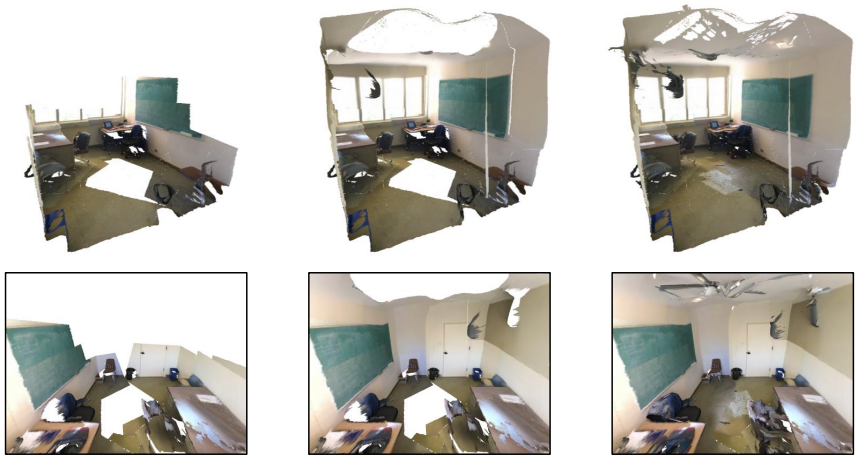
We utilized the official codebases of RGBD2 [9] and Text2Room [7] and conducted experiments on them for evaluation. In particular, we modify Text2Room [7] as T2R+RGBD to complete the mesh from sparse RGBD inputs by initializing the mesh for Text2Room and utilizing the camera poses of testing images as the predefined camera trajectory. For T2R+RGBD, the text prompt to Stable Diffusion [10] is "a simple and clean room in the style of S^*", where S^* is the textual token from *textual inversion*.

2 Method Details

2.1 Text Prompt

We utilize *textual inversion* [6] to extract the token S^* to represent the style of a room and use it in text prompts for Stable Diffusion, as described at Sec. 3.3 in the main paper. The following templates are used to extract the token S^* that represents the style of a room:

- "a S^* room",
- "the S^* room",
- "one S^* room",
- "a room in the style of S^*",
- "the room in the style of S^*",
- "one room in the style of S^*".



(a) Initialized mesh **(b)** After panorama **(c)** After completion

Fig. 1: Mesh completion. As described in Sec. 3.7 in the main paper, our mesh completion method can further complete the temporal mesh after RGB and depth panorama inpainting by sampling additional camera poses facing existing holes on the mesh.

Then, we utilize the extracted S^* in text prompts for Stable Diffusion. Note that we use a fixed input prompt: "a simple and clean room in the style of S^*". for all image inpainting, so GenRC doesn't require any scene-specific or detailed prompts shown in previous works [5, 7].

2.2 E-Diffusion

For E-Diffusion, we consider 8 rectangular views with the field of view as 98 degrees, which ensures the stitched panorama is fully covered by these views. The noise ϵ used to obtain x_{t-1}^i in Eq. (3) in the main paper is sampled randomly from the Gaussian distribution of unit variance every two steps.

For the input of MultiDiffusion [1] used in texture refinement, we stitch the eight perspective views together as one equirectangular panorama of 2048 $\times$ 1024 pixels and only keep the region with latitude between -45 and 45 degrees, resulting in a panoramic image of 2048 $\times$ 512 pixels. For MultiDiffusion, we consider 16 sliding windows with window size of 512 $\times$ 512 pixels and step size of 128 pixels.

2.3 Mesh Completion

We demonstrate our mesh completion method for the generation of a complete room-scale mesh. To this end, we will iteratively select 30 camera poses to patch

Table 2: Sensitivity Analysis of Camera Trajectories. The metrics of RGBD2 dramatically decline without in-order camera trajectories which are composed of closely adjacent camera poses. While T2R+RGBD performs a higher depth MSE, given shuffled trajectories, the decreasing visual metric reflects it generates structures that are not similar to the ground truth. However, GenRC can still effectively generate cross-view consistent room structures even if given camera trajectories are not in order.

Methods	$\text{PSNR}_{\text{color}}(\uparrow)$		$\text{MSE}_{\text{depth}}(\downarrow)$		$\text{CS}_{\text{input}}(\uparrow)$	
	in-order	shuffled	in-order	shuffled	in-order	shuffled
T2R+RGBD [7]	11.6	11.2	0.88	**0.76**	0.72	**0.75**
RGBD2 [9]	12.2	10.7	0.72	1.72	0.69	0.66
Ours	**12.9**	**12.7**	**0.59**	0.77	**0.74**	**0.75**

up the remaining holes in the mesh. To select an optimal camera pose in each iteration, we first find the bounding box that covers the scene mesh and randomly sample 200 camera poses centered within the central 80% of the bounding box in the horizontal direction and central 10% of the bounding box in the vertical direction. The elevation angles are between 15 and -15 degrees. Given the inpainting ratio as the ratio of unobserved pixels to total pixels in the rendered image and the back-face ratio as the ratio of pixels that are rendered from the back of the mesh, we filter out the camera poses with inpainting ratio greater than 50%, back-face ratio greater than 1%, or minimum depth less than 1m. Then, we select the camera poses with the highest product of inpainting ratio and minimum depth. Finally, we move the selected camera poses backward as long as the criteria we use to filter out camera poses are satisfied, which helps include as much information as possible into the field of view. We showcase the effectiveness of our proposed mesh completion method (mentioned in Sec. 3.7 in the main paper) qualitatively in Fig. 1.

3 Sensitivity Analysis of Camera Trajectories

In comparison to GenRC which aims to generate a panorama covering most parts of the scene, RGBD2 and T2R+RGBD iteratively generate the scene following a pre-designed camera trajectory composed of closely adjacent camera poses. In addition, these methods should start from a viewpoint where a certain portion of the mesh exists to ensure appearance and geometric consistency.

In this analysis, we test each method on the extremely sparse observations as 3% on the ScanNet dataset to analyze the sensitivity when the given camera trajectories are not composed of closely adjacent camera poses. To this end, we randomly shuffle the originally "in-order" camera trajectories, which are the sequences of camera poses of testing images, to "shuffled" camera trajectories. In Tab. 2, we can observe that, without "in-order" camera trajectories composed of closely adjacent camera poses, the visual metrics of both RGBD2 and T2R+RGBD decline while GenRC's performance almost remains the same. Es-

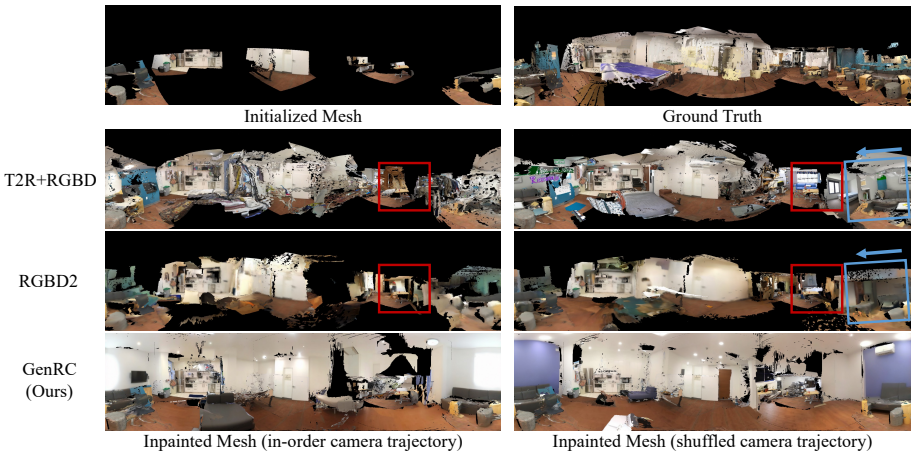


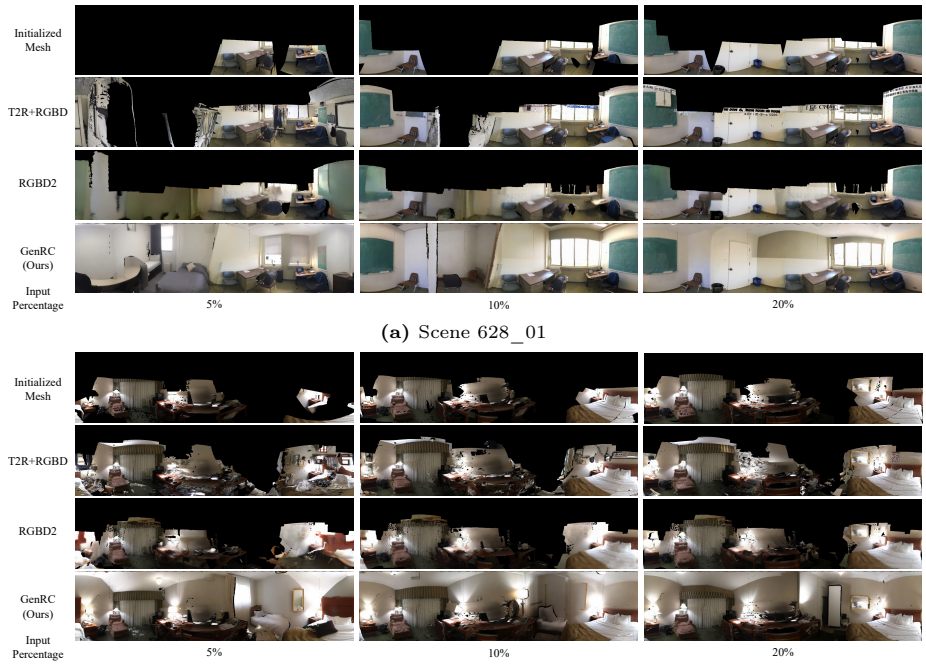
Fig. 2: Sensitivity Analysis of Camera Trajectories. In comparison to GenRC which aims to generate a panorama covering most parts of the scene, T2R+RGBD and RGBD2 iteratively generate the scene and require pre-designed camera trajectories composed of closely adjacent camera poses to ensure reasonable cross-view geometry. In addition, these methods should start from a viewpoint where a certain portion of the mesh exists to ensure appearance and geometric consistency. For instance, T2R+RGBD and RGBD2 may produce inconsistent geometries that cannot close the scene (as shown in the red boxes) and produce unreasonable room structures that are not perpendicular to the ground (as shown in the blue boxes). In contrast, GenRC can still generate cross-view consistent and complete room structures even if given camera trajectories are arbitrary.

pecially, the decreasing PSNR and increasing depth mean square error reflect that RGBD2 fails to reconstruct the reasonable appearance and geometry of a scene. Even if T2R+RGBD shows the higher mean square of depth while using shuffled trajectories, the decreasing visual metric reflects it generates structures that are not similar to the ground truth. For instance, as shown in Fig. 2, we can observe that T2R+RGBD and RGBD2 may generate inconsistent geometries that cannot close the scene and produce unreasonable room structures that are not perpendicular to the ground. Thanks to the continuous geometry provided by panorama inpainting (Sec 3.4 and Sec 3.5 in the main paper), the big portion of a mesh has been completed after the RGBD inpainting and therefore GenRC can effectively generate cross-view consistent room structures even if camera trajectories are arbitrary.

4 More Results

4.1 Qualitative Results on ScanNet

Refer to Sec. 4.3 in the main paper, GenRC outperforms two baseline methods on the ScanNet dataset when sparse observations are provided, which can be



(a) Scene 628_01

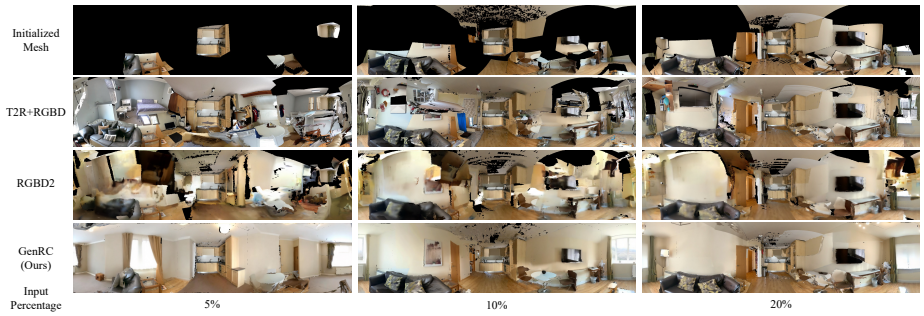
(b) Scene 679_01

Fig. 3: Comparison with baselines on ScanNet. We project generated meshes to panoramas to demonstrate the portions of meshes that are completed. In comparison to the prior method RGBD2 [9], GenRC can generate more complete meshes and high-fidelity images due to RGB and depth inpainting of GenRC. Besides, while T2R+RGBD [7] achieves high-fidelity texture, it may generate cross-view inconsistent geometry and artifacts. Please refer to Sec. 4.3 in the main paper for more quantitative discussions.

attributed to our proposed panorama inpainting technique that generates cross-view consistent panoramas, as described in Sec. 3.4 in the main paper. We project generated meshes to panoramas in Fig. 3 to demonstrate that the big portions of meshes are completed through our panorama inpainting.

4.2 Qualitative Results on ArkitScenes

Refer to Sec. 4.4 in the main paper, GenRC demonstrates superior performance in both visual and geometric metrics on the ArkitScenes dataset. In Fig. 4, we can observe that when it comes to cross-domain data, RGBD2 [9] and T2R+RGBD [7] cannot successfully produce reasonable room structures when the input observation is sparse (i.e., 5%). Refer to the 5% results of RGBD2 and T2R+RGBD in Fig. 4, the 3D geometries are inconsistent along with unreasonable room structures that are not perpendicular to the ground. These point out the limitation of iterative methods which could still fail even given predefined trajectories. In



(a) scene 41069021

(b) scene 47334360

Fig. 4: Cross-domain results on ArkitScenes. We project generated meshes to panoramas. When it comes to cross-domain data, RGBD2 [9] and T2R+RGBD [7] may generate unreasonable room structures, especially when the input observations are sparse (5%). In contrast, GenRC can still generate cross-view consistent room meshes. Please refer to Sec. 4.4 in the main paper for more quantitative discussions.

contrast, GenRC can still generate visually pleasing room appearance and 3D consistent room structures even if the input observations are sparse and without predefined trajectories.

4.3 Qualitative Results of Ablation Studies

GenRC generates high-fidelity and cross-view consistent panoramas via E-Diffusion, texture refinement, and textual inversion (refer to Sec. 3.3 and 3.4 in the main paper). We demonstrate the importance of these components by removing one of them at a time and the results are shown in Fig. 5.

4.4 More Ablations

We provide additional ablation studies on hyperparameter selection of E-Diffusion (refer to Sec. 3.4 in the main paper): (1) the number of views used while E-Diffusion and (2) the number of denoising steps for texture refinement. As shown

Fig. 5: Ablation studies on panorama inpainting and textual inversion. Directly applying MultiDiffusion [1] for panorama inpainting (referred to as w/o E-Diffusion) can produce high-resolution panoramas. However, they don't satisfy the geometry of equirectangular projection (e.g., the edges between walls and ceilings appear straight in the red boxes). In addition, directly performing our proposed E-Diffusion without texture refinement causes blurry results. Without textual inversion, the Stable Diffusion [10] model may generate objects (e.g., beds in the blue boxes) that are irrelevant to input images. Our method with all components can produce more detailed, stylistically coherent, and geometrically correct results. Please refer to Sec. 4.5 in the main paper for more quantitative discussions.

in Tab. 3 and Tab. 4, we consider 8 views in E-Diffusion and set the number of denoising steps for texture refinement as 20 out of 50 denoising steps in the reverse diffusion process.

5 Future Works

When a scene is cluttered with many objects, our approach may not complete the geometry of all 3D objects. We can further complete 3D objects through 3D object completion techniques such as [3, 8, 11, 12] in future works.

References

1. Bar-Tal, O., Yariv, L., Lipman, Y., Dekel, T.: Multidiffusion: Fusing diffusion paths for controlled image generation. ICML (2023) 3, 8, 9
2. 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) 1
3. 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) 8

Table 3: Ablation studies on the number of views for E-Diffusion. To inpaint a panorama with equirectangular geometry, E-Diffusion considers a set of overlapping views and denoise them concurrently (refer to Sec. 3.4 in the main paper). We finally select 8 views in E-Diffusion because it results in the lowest depth means square error and highest CLIP score when sparse observations are given (i.e., 5% and 10%).

Number of views	$\text{PSNR}_{\text{color}}(\uparrow)$			$\text{MSE}_{\text{depth}}(\downarrow)$			$\text{CS}_{\text{input}}(\uparrow)$		
	5%	10%	20%	5%	10%	20%	5%	10%	20%
6	**14.5**	16.4	**17.6**	0.29	0.16	**0.09**	0.771	0.801	**0.811**
8 (Ours)	14.4	**16.7**	**17.6**	**0.27**	**0.13**	**0.09**	**0.794**	**0.812**	0.809
16	**14.5**	**16.7**	**17.6**	0.32	0.19	0.11	0.772	0.801	**0.811**

Table 4: Ablation studies on the number of denoising steps for texture refinement. We utilize MultiDiffusion [1] for the last F denoising steps to refine the high-frequency texture (refer to Sec. 3.4 in the main paper). Out of 50 denoising steps in the reverse diffusion process, we set the number of denoising steps for texture refinement as 20, which results in the lowest depth mean square error. Note that $F = 50$ is equivalent to directly applying MultiDiffusion [1] for panorama inpainting (as w/o E-Diffusion described in Sec. 4.5 of the main paper) and $F = 0$ is equivalent to applying E-Diffusion without texture refinement.

Number of Steps (F)	$\text{PSNR}_{\text{color}}(\uparrow)$			$\text{MSE}_{\text{depth}}(\downarrow)$			$\text{CS}_{\text{input}}(\uparrow)$		
	5%	10%	20%	5%	10%	20%	5%	10%	20%
50 (w/o E-Diffusion)	13.8	16.4	17.5	0.32	0.16	**0.09**	0.765	0.799	0.808
40	14.4	16.6	**17.7**	0.31	**0.13**	0.10	0.770	0.801	**0.812**
30	14.4	16.6	17.6	0.35	0.14	0.10	0.772	0.801	0.811
20 (Ours)	14.4	**16.7**	17.6	**0.27**	**0.13**	**0.09**	**0.794**	**0.812**	0.809
10	14.3	16.6	17.5	0.37	0.15	0.12	0.764	0.801	0.811
0 (w/o texture refinement)	**14.6**	16.4	**17.7**	0.35	0.21	0.13	0.785	0.808	0.807

4. 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) 1
5. Fridman, R., Abecasis, A., Kasten, Y., Dekel, T.: Scenescape: Text-driven consistent scene generation. NeurIPS (2023) 3
6. 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) 2
7. Höllein, L., Cao, A., Owens, A., Johnson, J., Nießner, M.: Text2room: Extracting textured 3d meshes from 2d text-to-image models. ICCV (2023) 2, 3, 4, 6, 7
8. Kasten, Y., Rahamim, O., Chechik, G.: Point-cloud completion with pretrained text-to-image diffusion models. NeurIPS (2023) 8
9. Lei, J., Tang, J., Jia, K.: Rgbd2: Generative scene synthesis via incremental view inpainting using rgbd diffusion models. In: CVPR. pp. 8422–8434 (2023) 1, 2, 4, 6, 7

10. 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, 8
11. Wu, C.Y., Johnson, J., Malik, J., Feichtenhofer, C., Gkioxari, G.: Multiview compressive coding for 3d reconstruction. In: CVPR. pp. 9065–9075 (2023) 8
12. Yan, X., Lin, L., Mitra, N.J., Lischinski, D., Cohen-Or, D., Huang, H.: Shapeformer: Transformer-based shape completion via sparse representation. In: CVPR. pp. 6239–6249 (2022) 8