ShapeR: Robust Conditional 3D Shape Generation from Casual Captures

Supplementary Material

In this appendix, we provide additional details on the ShapeR evaluation dataset, further experimental results, including results on additional datasets, expanded implementation details of our method, and a discussion of its limitations.

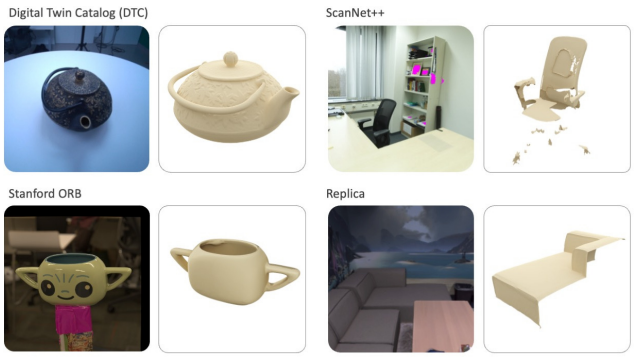

Figure 1. Comparison of 3D reconstruction datasets. DTC [4] and StanfordORB [9] offer controlled studio captures of isolated objects, while ScanNet++ [17] and Replica [14] provide realistic scenes but lack complete ground-truth shapes. The ShapeR evaluation dataset features casually captured sequences with complete meshes for geometric evaluation (see Figs. 2 and 3).

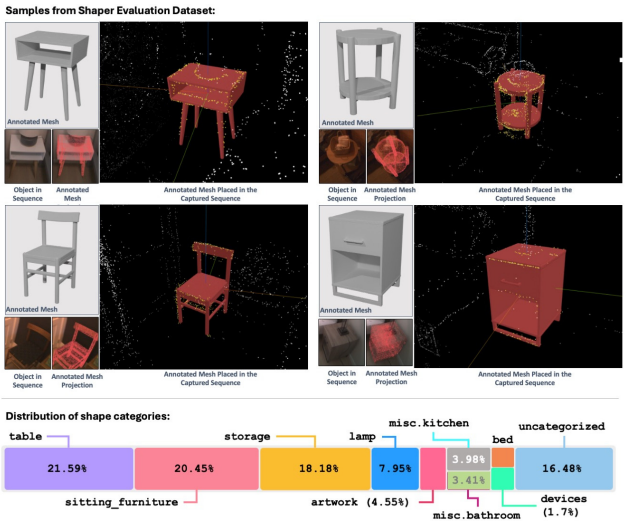

Figure 2. (Top) Examples from the ShapeR evaluation dataset. Each sub-image shows the annotated ground-truth mesh, a representative frame containing the object, the mesh placed within the sequence, and the projection of the mesh onto the image. (Bottom) Distribution of object shapes categories in the ShapeR evaluation set, covering 178 objects across 7 sequences

A. ShapeR Evaluation Dataset

Existing real-world 3D datasets for object reconstruction can be classified into two broad categories. Some, like Digital Twin Catalog [4], StanfordORB [13] and Google Scanned Objects [5] provide complete 3D shape geometry, but only in highly controlled setups. Here, objects are the central focus, placed on uncluttered, disoccluded tabletops, and captured in studio-like conditions (see Fig. 1 left). These datasets typically feature relatively small objects. Others, like ScanNet [3], ScanNet++ [17], Matterport3D [1] offer realistic scene arrangements, with clutter and occlusions captured casually. However, these are not suitable for object-centric evaluation, as the target geometry, usually obtained by 3D scanning, is incomplete in occluded for unobservable regions (see Fig. 1 right). The ShapeR Evaluation Dataset addresses these limitations by providing complete mesh geometry annotations for a selected set of objects, while maintaining casual capture conditions.

As shown in Fig. 3, sequences are recorded using Project Aria [6] Gen 1 or Gen 2 glasses, with the annotator casually walking through the scene and collecting images from the device's RGB and CV cameras. Aria Machine Perception Services [6] are then used to extract SLAM points and camera parameters from the sequence. For a selected set of objects, we obtain 3D shape annotations by moving each object to an area free of clutter and occlusions, capturing a high-resolution image, and manually segmenting it. A state-of-the-art image-to-3D model is then used to generate the 3D geometry. This geometry is manually verified for plausibility and aligned to the object's position in the original casual sequence using a web interface. This interface allows annotators to reposition and rigidly deform the shape in 3D space, guided by SLAM points from the sequence. Annotators further verify placement and dimensions by projecting the mesh into the original sequence images.

In total, we annotate 178 objects across 7 real indoor sequences, spanning a range of categories. Fig. 2 shows sample objects and the distribution of categories in the dataset.

B. Additional Experiments

In this section, we provide additional evaluations of ShapeR across a variety of datasets and tasks. We include comparisons against SegmentAnything 3D Objects [2], assessments on ScanNet++ [17] and Replica [14], results on the Digital Twin Catalog [4] (DTC), analysis of robustness trends, and demonstrations of monocular image-to-3D reconstruction.

Figure 3. To obtain pseudo-ground truth geometry for an object in the sequence (left), we first place the object in isolation to avoid clutter and occlusion, and capture a high-quality, uncluttered image. We then apply segmentation and image-to-3D modeling to generate the object's geometry (mid). This geometry is manually aligned and inserted back into the original casual sequence using a web annotation interface, verified by matching 2D projections to image silhouettes and by checking alignment with the sequence's point cloud (right).

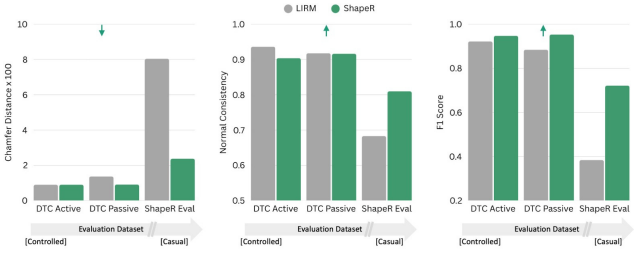

Figure 4. DTC Active, DTC Passive, and ShapeR Evaluation datasets represent a progression from highly controlled capture setups (DTC Active), to slightly less controlled environments (DTC Passive), and finally to casual, real-world scenes (ShapeR Evaluation). As the datasets become more challenging, baseline method metrics deteriorate, while ShapeR remains comparatively stable. Notably, the increase in scene casualness is not linear; ShapeR Evaluation is significantly more challenging than DTC Passive.

Comparison against SegmentAnything 3D Object [2]. SAM 3D Objects was very recently released and addresses the single image-to-3D reconstruction task using *interactive* segmentation. This approach marks a significant improvement in shape quality compared to previous image-to-scene methods like MIDI3D [7] and SceneGen [11], as well as single image-to-3D models such as Hunyuan3D [18], Amodal3R [16], and Direct3DS2 [15].

However, SAM 3D Objects is fundamentally limited by its reliance on single images. As a result, the reconstructed shapes are not metrically accurate. When scenes become more cluttered and contain multiple objects, the method struggles: layout, shape quality, aspect ratios, and relative scales all deteriorate, as shown in Fig. 7. In contrast, ShapeR leverages multiple posed views and additional modalities (such as SLAM points) to *automatically* reconstruct objects with metric accuracy and robust layout, even in casual, cluttered environments, while having only ever been trained on synthetic data. This multimodal approach enables ShapeR to maintain high-quality, metrically consistent reconstructions and object arrangements without interaction, outperforming single image-based methods in challenging real-world scenarios.

Evaluation on Scannet++ [17] and Replica [14]. Fig. 9 and Tab. 2 present a comparison of ShapeR on third-party casually captured datasets. For these experiments, we follow the protocol of DP-Recon [12], using their six ScanNet++ scenes and seven Replica scenes for evaluation. Since these datasets do not provide complete 3D geometry for evaluation (Figs. 1 and 9), we report only recall-based metrics. Notably, ShapeR produces complete reconstructions, often surpassing the ground-truth scans in terms of completeness, as the ground-truth meshes lack geometry in occluded regions.

Evaluation on Digital Twin Catalog (DTC) [4]. Fig. 8 and Tab. 3 show a comparison of ShapeR against LIRM [10] on the controlled capture datasets DTC Active and DTC Passive. Both datasets contain approximately 100 sequences each, with objects placed on a tabletop, free from occlusions and clutter. The passive variant allows for more free user movement, making it more casual compared to the active variant, where the user circles the object. As highlighted in Tab. 3, ShapeR matches state-of-the-art LIRM quality on the highly controlled active set and surpasses it on the more casual passive variant. Additionally, ShapeR produces sharper details on both datasets, as illustrated in Fig. 8.

Robustness Trends. DTC Active, DTC Passive, and the

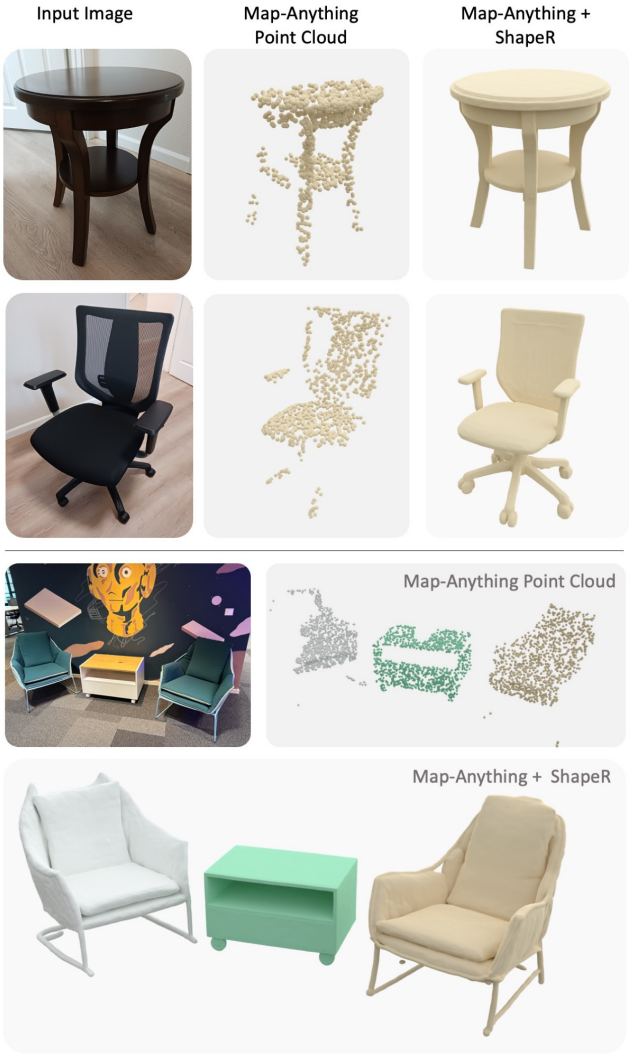

Figure 5. Single image to metric 3D with ShapeR. While ShapeR is trained to leverage posed multi-view signals, it can be configured for single-image 3D reconstruction without retraining by using a metric point cloud and camera estimator such as MapAnything [8]. This enables ShapeR to generate metrically accurate 3D shapes from a monocular image.

ShapeR evaluation dataset represent a non-linear progression from highly controlled to markedly more complex and casual capture setups. As shown in Fig. 4, ShapeR demonstrates significantly greater robustness to increased scene casualness compared to baseline methods such as LIRM, maintaining high reconstruction quality even as the capture conditions become more challenging. We further validate robustness to artificially added point cloud noise in Fig. 11 and Tab. 1.

Monocular Image-to-3D. While ShapeR is trained using multiple posed views and SLAM points extracted from them, it can also be applied to monocular images to produce metric 3D shapes without retraining by leveraging approaches like MapAnything [8]. As illustrated in Fig. 5, ShapeR can condition on a single image and its associated point cloud (obtained from MapAnything) to reconstruct both individual objects and entire scenes. Further improvements are possible by fine-tuning the model on real data collected in this monocular setup, as demonstrated in recent works [2].

Figure 6. ShapeR limitations. (a) Low image fidelity or limited views lead to incomplete or low-detail reconstructions. (b) Closely stacked or attached objects can cause meshes to include parts of adjacent structures, even when the point associated with these structures are not in the input (c) ShapeR relies on upstream 3D detection; missed or inaccurate detections result in unrecoverable objects.

DepthAnythingV3 and MapAnything Experiments. We evaluate our method on the ShapeR evaluation dataset using open-source DepthAnythingV3 points initialized with MapAnything poses (Tab. 1, Fig. 10). Despite the domain shift from the VIO semi-dense points used in training, our method generalizes robustly, keeping 2.1x improvement over SOTA (provided with Aria poses/points) with only marginal reduction.

Variant	CD[norm]↓ $\times 10^2$	NC↑	F1↑	CD[cm]↓
ShapeR	**2.375**	**0.810**	**0.722**	**1.326**
ShapeR w/ DAv3	3.111	0.766	0.615	1.676
ShapeR w/ Noise 0.025m	3.756	0.740	0.555	1.855
ShapeR w/ Noise 0.050m	4.845	0.693	0.448	2.537

Table 1. Results with MapAnything + DAv3 instead of using VIO as a source of SLAM points, and with aritificial noise added to VIO SLAM points.

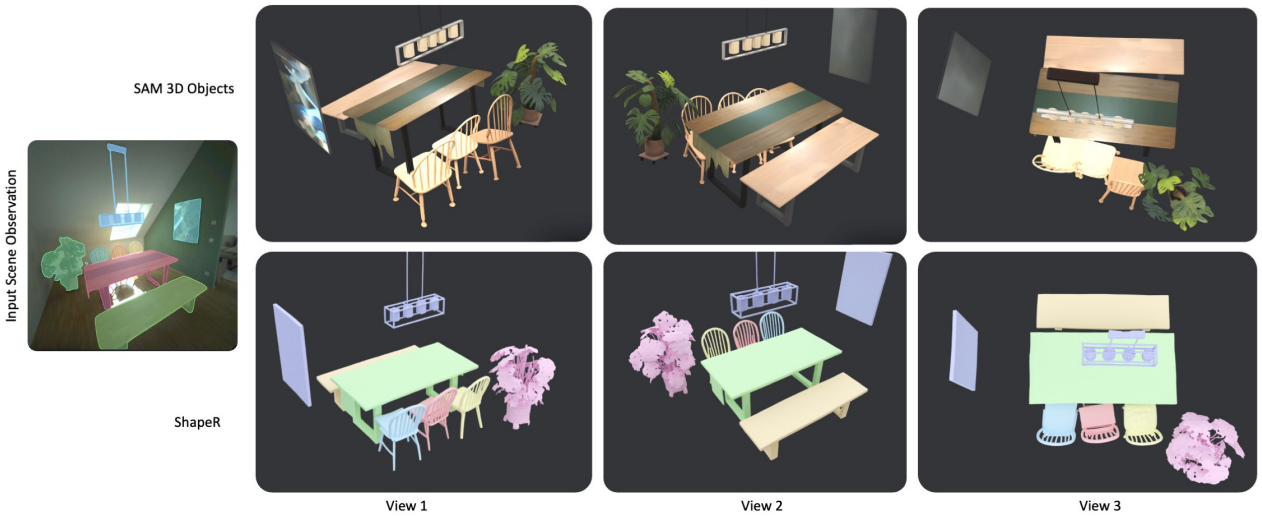

Figure 7. Comparison with SAM 3D Objects [2]. SAM 3D Objects takes a single image and interactive object segments to produce non-metric 3D shapes, which are generally accurate but may exhibit minor hallucinations (e.g., predicting five lamps instead of four) and poor object placement. In contrast, ShapeR leverages posed images from a sequence to generate metrically accurate geometry and consistently well-placed objects.

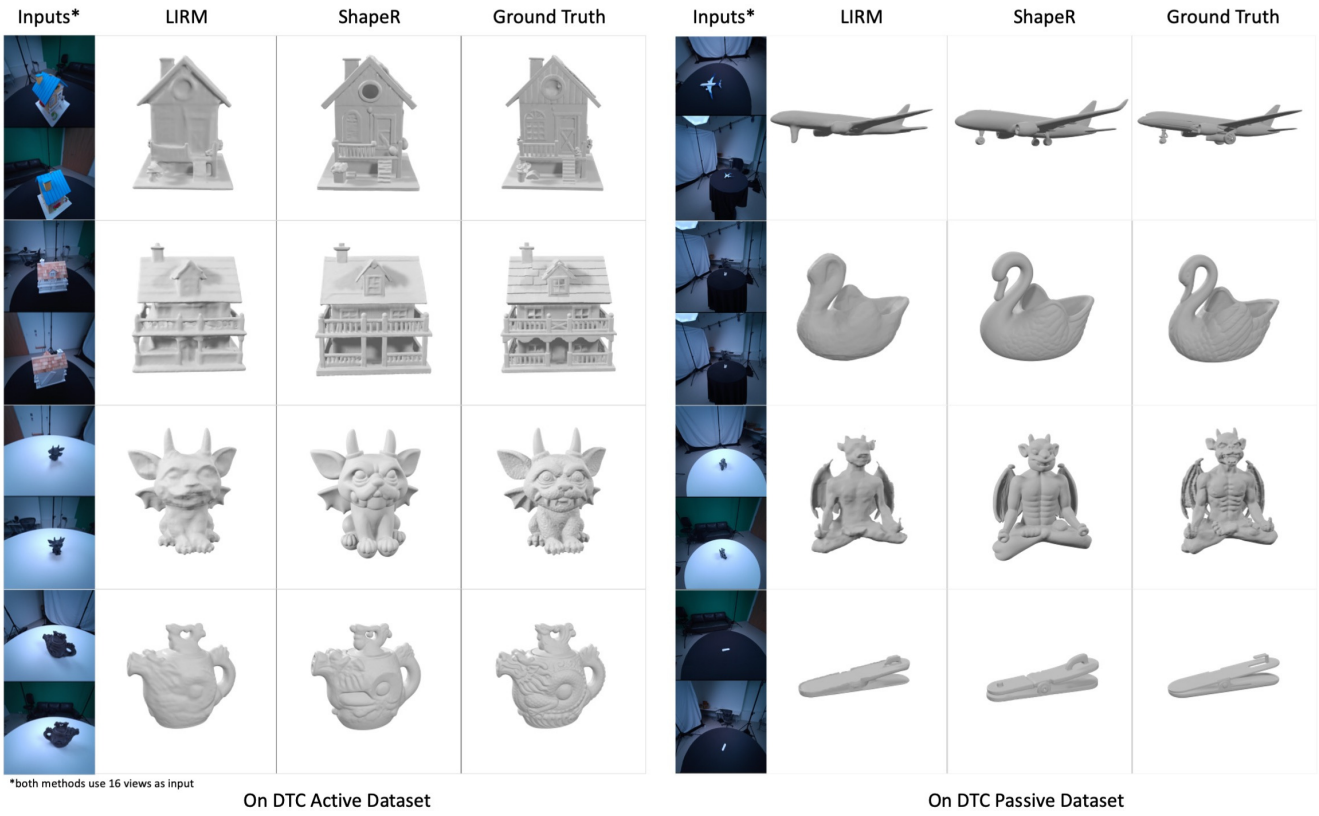

Figure 8. Comparison against LIRM [10] on DTC [4] Active and Passive sequences. Both setups feature tabletop objects without clutter or occlusions; however, Passive sequences allow more free user movement, while Active sequences involve the user circling the object. ShapeR performs competitively on Active sequences and surpasses LIRM on the slightly more casual Passive sequences.

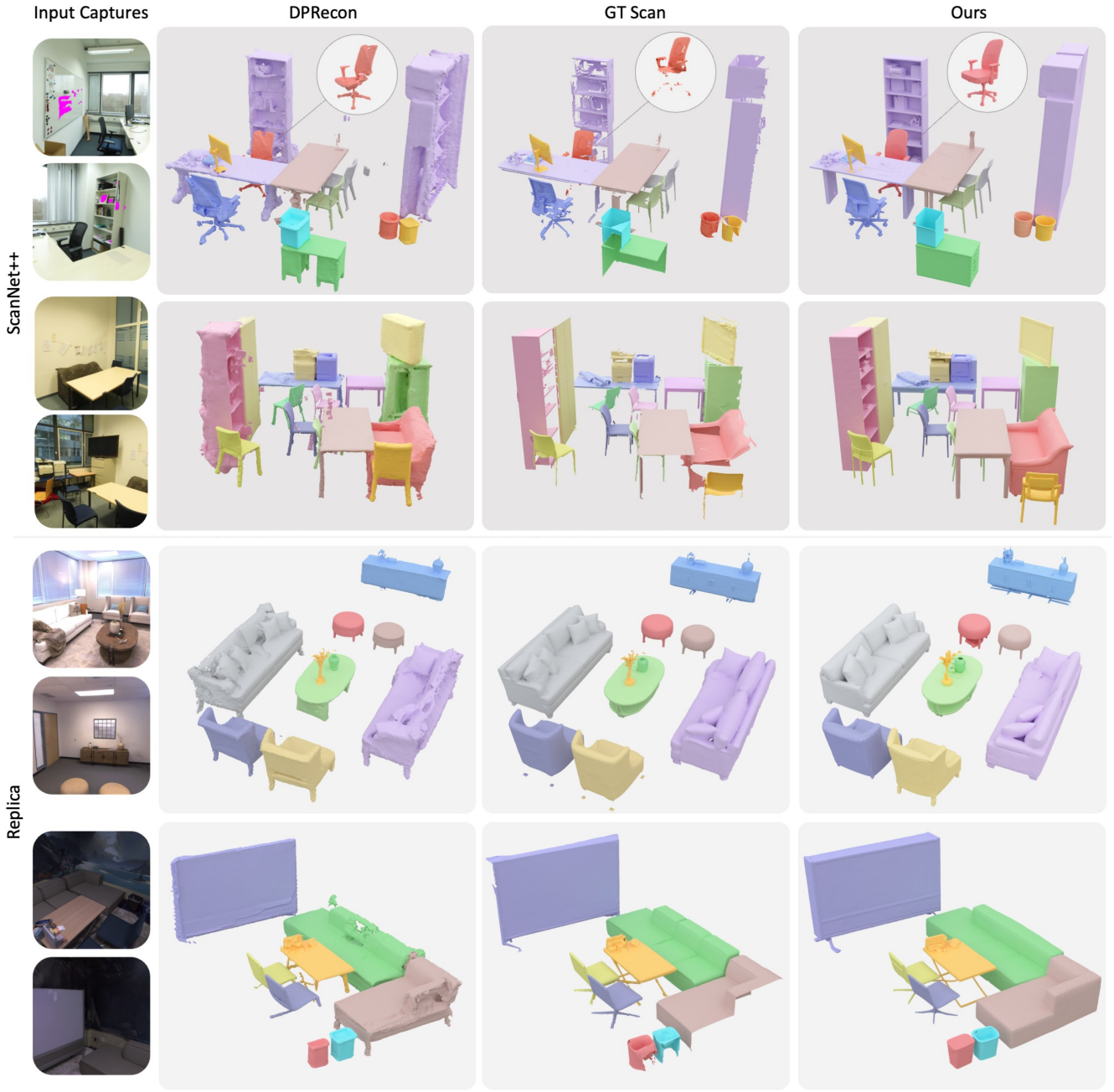

Figure 9. Reconstruction results on ScanNet++ [17] and Replica [14] scenes, compared to DPRecon [12]. ShapeR produces complete reconstructions, often surpassing the ground-truth scans in completeness, as the latter lack geometry in occluded regions.

C. Implementation Details

The 3D VAE encoder consists of 8 transformer layers and the decoder of 16 layers, each with a hidden width of 768, 12 attention heads. The VAE is trained for 200K steps with an effective batch size of 640 across 64 NVIDIA H100 GPUs. The rectified flow transformer comprises 16 dual-stream and 32 single-stream blocks, each with 16 attention heads and a hidden width of 1024. Training is performed for 550K steps using 128 H100 GPUs, progressively increasing the latent sequence length. The effective batch size is 512. Both networks are optimized using Adam with a learning rate of 5×10^{-5}.

D. Limitations

While ShapeR advances 3D shape generation under casual capture scenarios, several limitations remain. First, for ob-

Table 2. Reconstruction performance comparison on ScanNet++ [17] and Replica [14] datasets against DPRecon [12]. We use six scenes from ScanNet++ and seven scenes from Replica as processed by DPRecon. Note that chamfer distance, normal consistency and recall (R) are calculated in one direction, *i.e*. only point present on ground truth meshes are used for evaluation, due to tha lack of incomplete meshes present in these datasets.

Methods	ScanNet++			Replica		
	CD×10^2 ↓	NC↑	R↑	CD×10^2 ↓	NC↑	R↑
DPRecon [12]	7.69	0.73	0.45	4.65	0.75	0.57
ShapeR	**1.09**	**0.84**	**0.91**	**1.77**	**0.84**	**0.82**

Table 3. Reconstruction results on the DTC [4] Active and Passive datasets, each with approximately 100 sequences, compared against LIRM [10]. ShapeR achieves comparable performance to LIRM on the highly controlled Active sequences, and surpasses LIRM on the more challenging Passive sequences.

Methods	DTC Active			DTC Passive		
	CD×10^2 ↓	NC↑	F1↑	CD×10^2 ↓	NC↑	F1↑
LIRM [10]	**0.90**	**0.94**	0.92	1.37	0.91	0.88
ShapeR	0.94	0.91	**0.94**	**0.95**	0.91	**0.95**

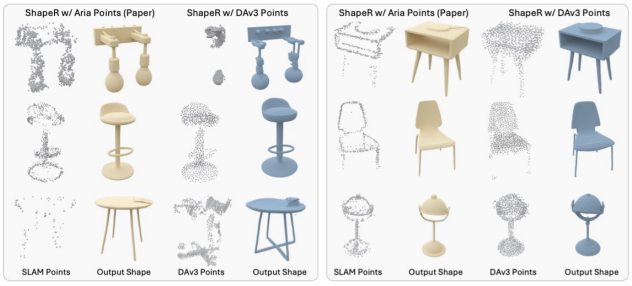

Figure 10. Results with MapAnything + DAv3 instead of using VIO as a source of SLAM points.

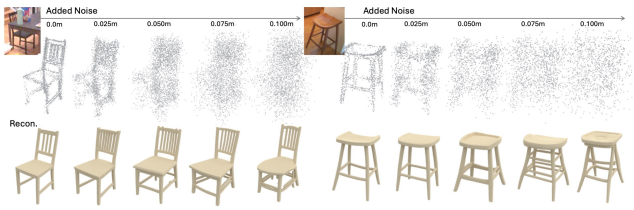

Figure 11. Effect of adding artificial point cloud noise.

jects captured with low image fidelity or observed in very few views, reconstructions can be incomplete or lack fine detail due to insufficient geometric and visual evidence. Second, when objects have other items stacked or closely attached (for example, tables supporting other objects), the reconstructed meshes sometimes include remnants of these adjacent structures instead of cleanly isolating the target object. Finally, ShapeR depends on upstream 3D instance detection; thus, missed detections or inaccurate bounding boxes directly propagate to the reconstruction stage, where missed objects cannot be recovered.

References

[1] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from rgb-d data in indoor environments. *arXiv preprint arXiv:1709.06158*, 2017. 1

[2] Xingyu Chen, Fu-Jen Chu, Pierre Gleize, Kevin J Liang, Alexander Sax, Hao Tang, Weiyao Wang, Michelle Guo, Thibaut Hardin, Xiang Li, Aohan Lin, Jia-Wei Liu, Ziqi Ma, Anushka Sagar, Bowen Song, Xiaodong Wang, Jianing Yang, Bowen Zhang, Piotr Dollár, Georgia Gkioxari, Matt Feiszli, and Jitendra Malik. Sam 3d: 3dfy anything in images, 2025. 1, 2, 3, 4

[3] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5828–5839, 2017. 1

[4] Zhao Dong, Ka Chen, Zhaoyang Lv, Hong-Xing Yu, Yunzhi Zhang, Cheng Zhang, Yufeng Zhu, Stephen Tian, Zhengqin Li, Geordie Moffatt, et al. Digital twin catalog: A large-scale photorealistic 3d object digital twin dataset. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 753–763, 2025. 1, 2, 4, 6

[5] Laura Downs, Anthony Francis, Nate Koenig, Brandon Kinman, Ryan Hickman, Krista Reymann, Thomas B McHugh, and Vincent Vanhoucke. Google scanned objects: A high-quality dataset of 3d scanned household items. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 2553–2560. IEEE, 2022. 1

[6] Jakob Engel, Kiran Somasundaram, Michael Goesele, Albert Sun, Alexander Gamino, Andrew Turner, Arjang Talattof, Arnie Yuan, Bilal Souti, Brighid Meredith, et al. Project aria: A new tool for egocentric multi-modal ai research. *arXiv preprint arXiv:2308.13561*, 2023. 1

[7] Zehuan Huang, Yuan-Chen Guo, Xingqiao An, Yunhan Yang, Yangguang Li, Zi-Xin Zou, Ding Liang, Xihui Liu, Yan-Pei Cao, and Lu Sheng. Midi: Multi-instance diffusion for single image to 3d scene generation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 23646–23657, 2025. 2

[8] Nikhil Keetha, Norman Müller, Johannes Schönberger, Lorenzo Porzi, Yuchen Zhang, Tobias Fischer, Arno Knapitsch, Duncan Zauss, Ethan Weber, Nelson Antunes, et al. Mapanything: Universal feed-forward metric 3d reconstruction. *arXiv preprint arXiv:2509.13414*, 2025. 3

[9] Zhengfei Kuang, Yunzhi Zhang, Hong-Xing Yu, Samir Agarwala, Elliott Wu, Jiajun Wu, et al. Stanford-orb: a real-world 3d object inverse rendering benchmark. *Advances in Neural Information Processing Systems*, 36:46938–46957, 2023. 1

[10] Zhengqin Li, Dilin Wang, Ka Chen, Zhaoyang Lv, Thu Nguyen-Phuoc, Milim Lee, Jia-Bin Huang, Lei Xiao, Yufeng

Zhu, Carl S Marshall, et al. Lirm: Large inverse rendering model for progressive reconstruction of shape, materials and view-dependent radiance fields. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 505–517, 2025. 2, 4, 6

[11] Yanxu Meng, Haoning Wu, Ya Zhang, and Weidi Xie. Scenegen: Single-image 3d scene generation in one feedforward pass. *arXiv preprint arXiv:2508.15769*, 2025. 2

[12] Junfeng Ni, Yu Liu, Ruijie Lu, Zirui Zhou, Song-Chun Zhu, Yixin Chen, and Siyuan Huang. Decompositional neural scene reconstruction with generative diffusion prior. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 6022–6033, 2025. 2, 5, 6

[13] Yinyu Nie, Xiaoguang Han, Shihui Guo, Yujian Zheng, Jian Chang, and Jian Jun Zhang. Total3dunderstanding: Joint layout, object pose and mesh reconstruction for indoor scenes from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 55–64, 2020. 1

[14] Julian Straub, Thomas Whelan, Lingni Ma, Yufan Chen, Erik Wijmans, Simon Green, Jakob J Engel, Raul Mur-Artal, Carl Ren, Shobhit Verma, et al. The replica dataset: A digital replica of indoor spaces. *arXiv preprint arXiv:1906.05797*, 2019. 1, 2, 5, 6

[15] Shuang Wu, Youtian Lin, Feihu Zhang, Yifei Zeng, Yikang Yang, Yajie Bao, Jiachen Qian, Siyu Zhu, Xun Cao, Philip Torr, et al. Direct3d-s2: Gigascale 3d generation made easy with spatial sparse attention. *arXiv preprint arXiv:2505.17412*, 2025. 2

[16] Tianhao Wu, Chuanxia Zheng, Frank Guan, Andrea Vedaldi, and Tat-Jen Cham. Amodal3r: Amodal 3d reconstruction from occluded 2d images. *arXiv preprint arXiv:2503.13439*, 2025. 2

[17] Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. Scannet++: A high-fidelity dataset of 3d indoor scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12–22, 2023. 1, 2, 5, 6

[18] Zibo Zhao, Zeqiang Lai, Qingxiang Lin, Yunfei Zhao, Haolin Liu, Shuhui Yang, Yifei Feng, Mingxin Yang, Sheng Zhang, Xianghui Yang, et al. Hunyuan3d 2.0: Scaling diffusion models for high resolution textured 3d assets generation. *arXiv preprint arXiv:2501.12202*, 2025. 2