

Decompositional Neural Scene Reconstruction with Generative Diffusion Prior

Junfeng Ni^{1,2} Yu Liu^{1,2} Ruijie Lu^{2,3} Zirui Zhou¹
Song-Chun Zhu^{1,2,3} Yixin Chen^{2†} Siyuan Huang^{2†}

[†] Corresponding author ¹ Tsinghua University

² State Key Laboratory of General Artificial Intelligence, BIGAI ³ Peking University

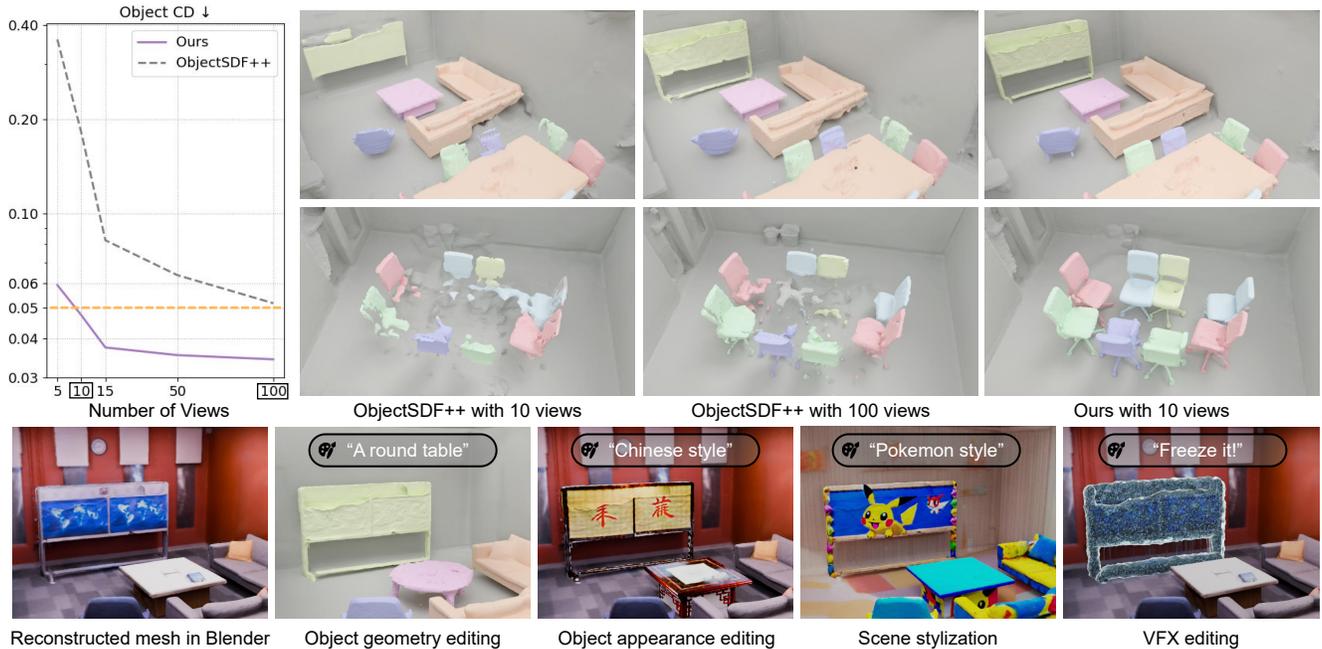


Figure 1. We propose **DP-RECON**, which capitalizes on pre-trained diffusion models for complete and decompositional neural scene reconstruction. This approach significantly improves reconstruction quality in less captured regions, where previous methods often struggle. Additionally, our method enables flexible text-based editing of geometry and appearance, as well as photorealistic VFX editing.

Abstract

Decompositional reconstruction of 3D scenes, with complete shapes and detailed texture of all objects within, is intriguing for downstream applications but remains challenging, particularly with sparse views as input. Recent approaches incorporate semantic or geometric regularization to address this issue, but they suffer significant degradation in underconstrained areas and fail to recover occluded regions. We argue that the key to solving this problem lies in supplementing missing information for these areas. To this end, we propose **DP-RECON**, which employs diffusion priors in the form of Score Distillation Sampling (SDS) to optimize the neural representation of each individual object under novel views. This provides additional information for the underconstrained areas, but directly incorporating diffusion prior raises potential conflicts between the reconstruction and generative guidance. Therefore, we further introduce a visibility-guided approach to dynamically adjust the per-pixel SDS loss weights. Together these com-

ponents enhance both geometry and appearance recovery while remaining faithful to input images. Extensive experiments across Replica and ScanNet++ demonstrate that our method significantly outperforms state-of-the-art methods. Notably, it achieves better object reconstruction under 10 views than the baselines under 100 views. Our method enables seamless text-based editing for geometry and appearance through SDS optimization and produces decomposed object meshes with detailed UV maps that support photorealistic Visual effects (VFX) editing.

1. Introduction

3D scene reconstruction from multi-view images is a long-standing topic in computer vision [8, 15]. Traditional methods typically represent the entire scene holistically, limiting flexibility and downstream usability. In contrast, decompositional reconstruction [10, 24] aims to break down the implicit 3D representation into individual objects in the scene

and facilitate broader applications in embodied AI [1, 5], robotics [4, 7], and more [3]. However, existing methods [13, 16, 25] in decompositional neural reconstruction still fall short of expectations in downstream applications to reconstruct complete 3D geometry and accurate appearance (see Fig. 1), especially in less densely captured or heavily occluded areas with sparse inputs. To address the challenge of sparse-view reconstruction, many approaches propose to incorporate semantic or geometric regularizations [6, 9, 17, 26]. Still, they often demonstrate significant degradation in non-observable regions since they fail to provide additional information for the underconstrained areas. Thus, we believe the key is to introduce supplementary information for these areas based on the observation from known views.

In this paper, we propose **DP-RECON** to facilitate the decompositional neural reconstruction with generative diffusion prior. Given multiple posed images, the neural implicit representation is optimized to represent both individual objects and the background within the scene. Besides the reconstruction loss, we employ a 2D diffusion model as a critic to supervise the optimization of each object through SDS [18], which iteratively refines the 3D representation by evaluating the quality of novel views from differentiable rendering. We use the pretrained Stable Diffusion [20], a more general diffusion model without fine-tuning on specific datasets. We meticulously design the optimization pipeline so that the generative prior optimizes both the geometry and appearance of each object alongside the reconstruction loss, filling in the missing information in unobserved and occluded regions.

However, directly integrating the diffusion prior into the reconstruction pipeline may compromise the overall consistency, particularly in observed regions, due to their potential conflicts. Ideally, we want to preserve the visible area in the input images while the diffusion prior completes the rest. To alleviate this problem, we propose a novel visibility approach that models the visibility of 3D points across the input views using a learnable grid. The visibility information is derived from the accumulated transmittance in volume rendering, enabling us to optimize the visibility grid without introducing computationally intensive external visibility priors [21]. For each novel view, the visibility map can be rendered from this grid, which can dynamically adjust the per-pixel SDS and rendering loss weights, benefiting both geometry and appearance optimization stages.

Extensive experiment results on Replica [22] and ScanNet++ [27] demonstrate that our method significantly surpasses all state-of-the-art methods in both geometry and appearance reconstruction, particularly in heavily occluded regions. *Remarkably, with only 10 input views, our method achieves object reconstruction quality superior to baseline methods that rely on 100 input views for heavily occluded*

scenes in Fig. 1. Our method enables seamless text-based editing, e.g., geometry and appearance stylization, using SDS optimization. It produces decomposed object meshes with detailed UV maps, enabling photorealistic rendering and VFX editing in common 3D software, thereby supporting various downstream applications.

In summary, our main contributions are three-fold:

- We introduce a novel method **DP-RECON** that incorporates generative prior into decompositional scene reconstruction, significantly improving geometry and appearance recovery, particularly in heavily occluded regions.
- We propose a visibility-guided approach to dynamically adjust the SDS loss, alleviating the conflict between the reconstruction objective and generative prior guidance.
- Extensive experiments demonstrate that our model significantly enhances both geometry and appearance. Our method enables seamless geometry and appearance editing, yielding decomposed object meshes with detailed UV maps for broad downstream applications.

2. Method

Given a set of posed RGB images and corresponding instance masks, we aim to reconstruct the geometry and appearance of objects and the background in the scene. Fig. 2 presents an overview of our proposed **DP-RECON**.

2.1. 3D Reconstruction with Generative Priors

The latent neural representation of the 3D scene is primarily optimized by the reconstruction loss \mathcal{L}_{recon} derived from volume rendering, following prior work [13, 24, 25]. However, regions with sparse capture or heavy occlusions often lead to suboptimal geometry and appearance recovery due to insufficient information as reconstruction guidance. To mitigate this gap, we introduce diffusion prior to optimize the the 3D model, both in geometry and appearance, so that it looks realistic at novel unobserved views.

Prior-guided Geometry Optimization We adopt the decompositional neural implicit surface as our 3D representation, which is parameterized with a series of multi-layer perceptrons (MLPs) with parameter θ . The rendering functions serve as the image generator $g(\theta)$. At each training iteration, we sample the j -th object and render its normal map and mask map at a randomly sampled camera pose. Following previous work [2, 19], we use a concatenated map \tilde{n}_j of the normal and mask maps as the input for the diffusion model to improve geometric optimization stability. We then employ the SDS loss to compute the gradient for updating θ as follows:

$$\nabla_{\theta} \mathcal{L}_{SDS}^g = \mathbb{E}_{t, \epsilon} \left[w(t) (\hat{\epsilon}_{\phi}(z_t; y, t) - \epsilon) \frac{\partial z}{\partial \tilde{n}_j} \frac{\partial \tilde{n}_j}{\partial \theta} \right], \quad (1)$$

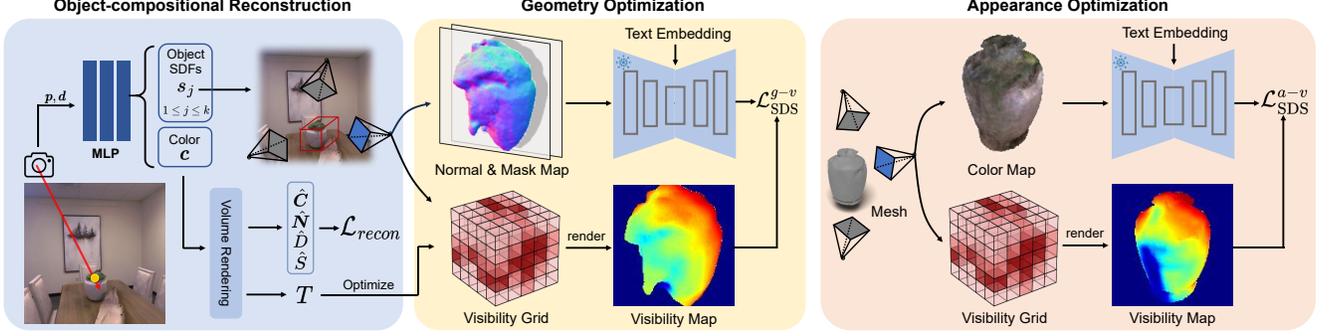


Figure 2. **Overview of DP-RECON.** We first use reconstruction loss \mathcal{L}_{recon} for decompositional neural reconstruction, followed by the prior-guided geometry optimization stage that incorporates SDS loss \mathcal{L}_{SDS}^{g-v} . We finally export the object meshes and optimize their appearance with \mathcal{L}_{SDS}^{a-v} . The visibility balances the guidance from prior and reconstruction by dynamically adjusting per-pixel SDS loss.

where z is the latent code of \tilde{n}_j . The background is also treated as one object for geometry optimization.

Prior-guided Appearance Optimization To produce object meshes with detailed UV maps, which are friendly for photorealistic rendering in common 3D software and enable more downstream applications, we directly optimize the mesh appearance rather than Neural Radiance Field (NeRF)’s appearance field. More specifically, we export the mesh for each object after the geometry optimization stage. Using NVDiffrast [11] for differentiable mesh rendering, we employ another small network ψ to predict color for the mesh surface points. At each training iteration, the color map c_j for j -th is rendered at a randomly selected camera view, and the appearance SDS loss is used to compute the gradient for updating ψ :

$$\nabla_{\psi} \mathcal{L}_{SDS}^a = \mathbb{E}_{t,\epsilon} \left[w(t) (\hat{\epsilon}_{\phi}(z_t; y, t) - \epsilon) \frac{\partial z}{\partial c_j} \frac{\partial c_j}{\partial \psi} \right], \quad (2)$$

where z is the latent code of c_j . Note that the color rendering loss from input views is also used to optimize ψ .

2.2. Visibility-guided Optimization

Score Distillation Sampling (SDS), despite its wide application, has been shown to suffer from significant artifacts [12, 28], such as oversaturation, oversmoothing, and low-diversity, and optimization instability [14, 23]. They become even more significant when optimizing the latent 3D representation through both reconstruction and SDS guidance, due to their potential conflict, leading to inconsistencies with the observations. We address this problem by proposing a visibility-guided approach, which adjusts geometry and appearance SDS loss based on pixel visibility in the input view when rendered from a novel view.

Visibility Modeling We introduce a learnable visibility grid G to model the visibility v of a 3D point p in the input views. We employ a view-independent modeling for visi-

bility, *i.e.*, $v = G(p)$, as it only depends on the input views and is independent of the ray direction from novel views.

Ideally, points observed in more input views should have higher visibility. The accumulated transmittance T for a 3D point p represents the probability that the corresponding ray reaches p without hitting any other particles - higher transmittance T means greater visibility probability in the input views. Therefore, we initialize G as zero and utilize the T from input views to optimize the visibility grid G via:

$$\mathcal{L}_v = \sum_{i=0}^n \max(T_i - G(p_i), 0). \quad (3)$$

We detach the gradient of T_i to avoid the influence on the reconstruction network. We optimize G after finishing the decompositional reconstruction stage to ensure the accuracy of the transmittance and freeze G in the geometry and appearance optimization stage with generative diffusion prior.

Visibility-guided SDS We obtain the visibility map V under novel view by volume rendering. V for a ray r is calculated as $V(r) = \sum_{i=0}^{n-1} T_i \alpha_i v_i$. The visibility weighting function $w^v(z)$ is calculated as:

$$w^v(z) = \begin{cases} w_0 + m_0 V(z) & \text{if } V(z) \leq \tau \\ w_1 + m_1 V(z) & \text{if } V(z) > \tau \end{cases}, \quad (4)$$

where w and m are piecewise linear coefficients, $V(z)$ denotes the pixel-wise visibility associated with latent z , and τ a threshold separating high and low visibility area. We reduce the SDS loss weight in high visibility regions to enhance reconstruction guidance while increasing SDS loss weight in low visibility regions for higher generative prior guidance. Then we rewrite Eq. (1) and Eq. (2) as:

$$\begin{aligned} \nabla_{\theta} \mathcal{L}_{SDS}^{g-v} &= \mathbb{E}_{t,\epsilon} \left[w^v(z) w(t) (\hat{\epsilon}_{\phi}(z_t; y, t) - \epsilon) \frac{\partial z}{\partial n_j} \frac{\partial n_j}{\partial \theta} \right] \\ \nabla_{\psi} \mathcal{L}_{SDS}^{a-v} &= \mathbb{E}_{t,\epsilon} \left[w^v(z) w(t) (\hat{\epsilon}_{\phi}(z_t; y, t) - \epsilon) \frac{\partial z}{\partial c_j} \frac{\partial c_j}{\partial \psi} \right] \end{aligned} \quad (5)$$

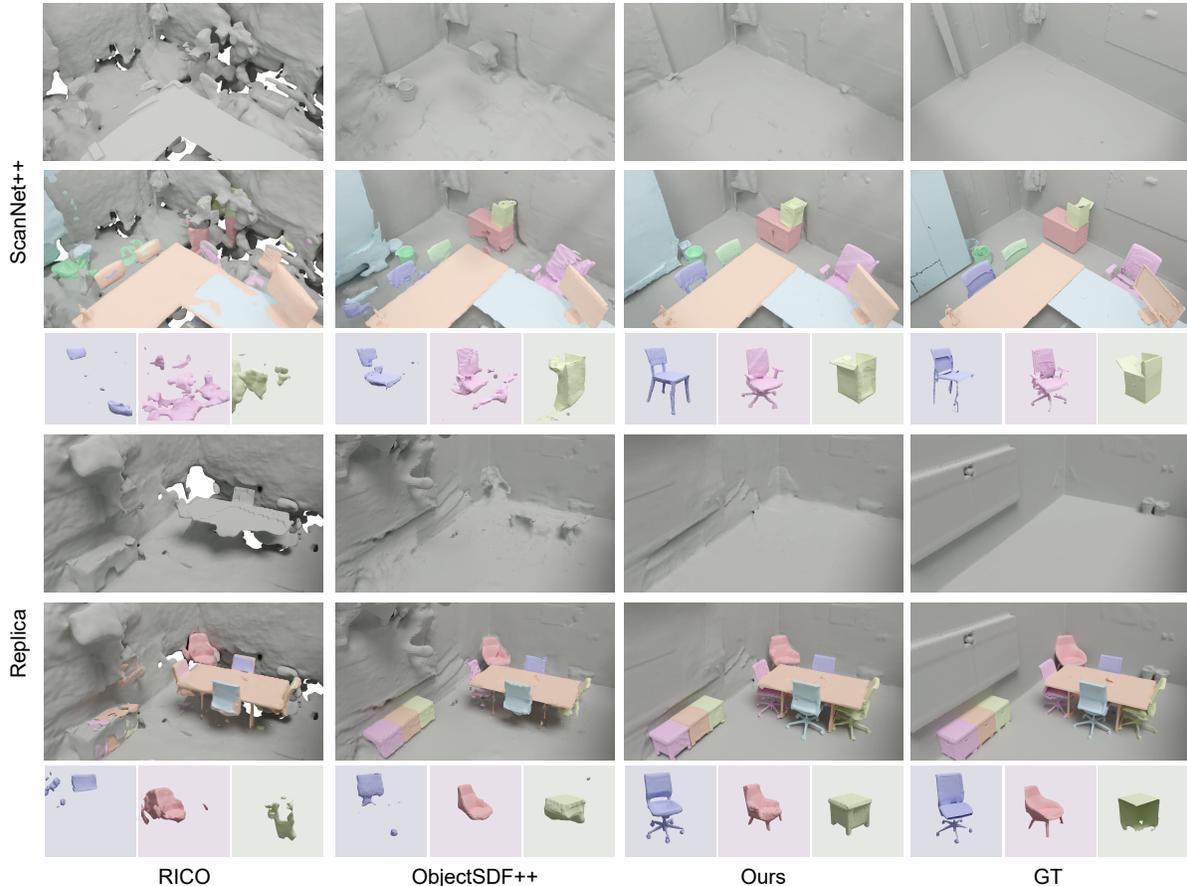


Figure 3. Qualitative comparison of 10-view reconstruction.

Table 1. Compositional object reconstruction.

Method	Object Reconstruction				BG Reconstruction		
	CD↓	F-Score ↑	NC↑	mIoU↑	CD↓	F-Score ↑	NC↑
Replica							
RICO	10.32	49.26	61.27	71.21	13.35	39.73	85.32
ObjectSDF++	7.49	56.69	64.75	71.72	10.33	44.19	86.34
Ours	5.54	67.71	73.50	88.21	9.39	46.14	92.83
ScanNet++							
RICO	24.09	39.26	58.26	42.25	18.37	34.72	78.26
ObjectSDF++	14.52	46.87	61.57	45.73	13.20	38.92	80.47
Ours	5.03	66.55	72.91	70.01	11.51	40.12	86.24

3. Experiments

We compare **DP-RECON** with decompositional reconstruction baselines RICO [13] and ObjectSDF++ [25] on sparse-view 3D reconstruction using 10 input views. Key findings are summarized in Tab. 1, Fig. 3 and Fig. 4:

1. Our method significantly outperforms all baselines. By integrating generative priors, it achieves more accurate reconstructions in less captured areas, more precise object structures, smoother background reconstruction, and fewer floating artifacts, as illustrated in Fig. 3.

- Generative priors notably improve reconstruction in occluded regions, yielding better object structure and fewer artifacts (e.g., the chair behind the table or background occlusion in Fig. 3). Our visibility-guided strategy also preserves consistency with input images in visible areas, mitigating conflicts between the priors and observations.
- As shown in Fig. 4, our method enables seamless text-based editing of geometry and appearance for each object. It also produces high-fidelity decomposed meshes with detailed UV maps, enabling VFX workflows in standard 3D software such as Blender.



Figure 4. Examples of scene editing. Our model seamlessly supports flexible text-guided editing, as well as VFX editing.

References

- [1] Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton Van Den Hengel. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2
- [2] Rui Chen, Yongwei Chen, Ningxin Jiao, and Kui Jia. Fantasia3d: Disentangling geometry and appearance for high-quality text-to-3d content creation. In *International Conference on Computer Vision (ICCV)*, 2023. 2
- [3] Yiwen Chen, Zilong Chen, Chi Zhang, Feng Wang, Xiaofeng Yang, Yikai Wang, Zhongang Cai, Lei Yang, Huaping Liu, and Guosheng Lin. Gaussianeditor: Swift and controllable 3d editing with gaussian splatting. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2
- [4] Ran Gong, Jiangyong Huang, Yizhou Zhao, Haoran Geng, Xiaofeng Gao, Qingyang Wu, Wensi Ai, Ziheng Zhou, Demetri Terzopoulos, Song-Chun Zhu, et al. Arnold: A benchmark for language-grounded task learning with continuous states in realistic 3d scenes. *arXiv preprint arXiv:2304.04321*, 2023. 2
- [5] Jiayuan Gu, Fanbo Xiang, Xuanlin Li, Zhan Ling, Xiqiang Liu, Tongzhou Mu, Yihe Tang, Stone Tao, Xinyue Wei, Yunchao Yao, et al. Maniskill2: A unified benchmark for generalizable manipulation skills. *arXiv preprint arXiv:2302.04659*, 2023. 2
- [6] Ajay Jain, Matthew Tancik, and Pieter Abbeel. Putting nerf on a diet: Semantically consistent few-shot view synthesis. In *International Conference on Computer Vision (ICCV)*, 2021. 2
- [7] Yunfan Jiang, Agrim Gupta, Zichen Zhang, Guanzhi Wang, Yongqiang Dou, Yanjun Chen, Li Fei-Fei, Anima Anandkumar, Yuke Zhu, and Linxi Fan. Vima: General robot manipulation with multimodal prompts. *arXiv preprint arXiv:2210.03094*, 2022. 2
- [8] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. In *ACM SIGGRAPH / Eurographics Symposium on Computer Animation (SCA)*, 2023. 1
- [9] Mijeong Kim, Seonguk Seo, and Bohyung Han. Infonerf: Ray entropy minimization for few-shot neural volume rendering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2
- [10] Xin Kong, Shikun Liu, Marwan Taher, and Andrew J Davison. vmap: Vectorised object mapping for neural field slam. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 1
- [11] Samuli Laine, Janne Hellsten, Tero Karras, Yeongho Seol, Jaakko Lehtinen, and Timo Aila. Modular primitives for high-performance differentiable rendering. In *ACM SIGGRAPH / Eurographics Symposium on Computer Animation (SCA)*, 2020. 3
- [12] Kyungmin Lee, Kihyuk Sohn, and Jinwoo Shin. Dreamflow: High-quality text-to-3d generation by approximating probability flow. *arXiv preprint arXiv:2403.14966*, 2024. 3
- [13] Zizhang Li, Xiaoyang Lyu, Yuanyuan Ding, Mengmeng Wang, Yiyi Liao, and Yong Liu. Rico: Regularizing the unobservable for indoor compositional reconstruction. In *International Conference on Computer Vision (ICCV)*, 2023. 2, 4
- [14] David McAllister, Songwei Ge, Jia-Bin Huang, David W Jacobs, Alexei A Efros, Aleksander Holynski, and Angjoo Kanazawa. Rethinking score distillation as a bridge between image distributions. *arXiv preprint arXiv:2406.09417*, 2024. 3
- [15] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *European Conference on Computer Vision (ECCV)*, 2020. 1
- [16] Junfeng Ni, Yixin Chen, Bohan Jing, Nan Jiang, Bin Wang, Bo Dai, Puhao Li, Yixin Zhu, Song-Chun Zhu, and Siyuan Huang. Phyrecon: Physically plausible neural scene reconstruction. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. 2
- [17] Michael Niemeyer, Jonathan T. Barron, Ben Mildenhall, Mehdi S. M. Sajjadi, Andreas Geiger, and Noha Radwan. Regnerf: Regularizing neural radiance fields for view synthesis from sparse inputs. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2
- [18] Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. In *International Conference on Learning Representations (ICLR)*, 2022. 2
- [19] Lingteng Qiu, Guanying Chen, Xiaodong Gu, Qi Zuo, Mutian Xu, Yushuang Wu, Weihao Yuan, Zilong Dong, Liefeng Bo, and Xiaoguang Han. Richdreamer: A generalizable normal-depth diffusion model for detail richness in text-to-3d. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2
- [20] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2
- [21] Nagabhushan Somraj and Rajiv Soundararajan. ViP-NeRF: Visibility prior for sparse input neural radiance fields. In *ACM SIGGRAPH / Eurographics Symposium on Computer Animation (SCA)*, 2023. 2
- [22] Julian Straub, Thomas Whelan, Lingni Ma, Yufan Chen, Erik Wijmans, Simon Green, Jakob J. Engel, Raul Mur-Artal, Carl Ren, Shobhit Verma, Anton Clarkson, Mingfei Yan, Brian Budge, Yajie Yan, Xiaqing Pan, June Yon, Yuyang Zou, Kimberly Leon, Nigel Carter, Jesus Briales, Tyler Gillingham, Elias Mueggler, Luis Pesqueira, Manolis Savva, Dhruv Batra, Hauke M. Strasdat, Renzo De Nardi, Michael Goesele, Steven Lovegrove, and Richard Newcombe. The Replica dataset: A digital replica of indoor spaces. *arXiv preprint arXiv:1906.05797*, 2019. 2
- [23] Zhengyi Wang, Cheng Lu, Yikai Wang, Fan Bao, Chongxuan Li, Hang Su, and Jun Zhu. Prolificdreamer: High-fidelity and diverse text-to-3d generation with variational score distillation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. 3

- [24] Qianyi Wu, Xian Liu, Yuedong Chen, Kejie Li, Chuanxia Zheng, Jianfei Cai, and Jianmin Zheng. Object-compositional neural implicit surfaces. In *European Conference on Computer Vision (ECCV)*, 2022. [1](#), [2](#)
- [25] Qianyi Wu, Kaisiyuan Wang, Kejie Li, Jianmin Zheng, and Jianfei Cai. Objectsdf++: Improved object-compositional neural implicit surfaces. In *International Conference on Computer Vision (ICCV)*, 2023. [2](#), [4](#)
- [26] Jiawei Yang, Marco Pavone, and Yue Wang. Freenerf: Improving few-shot neural rendering with free frequency regularization. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. [2](#)
- [27] Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. Scannet++: A high-fidelity dataset of 3d indoor scenes. In *International Conference on Computer Vision (ICCV)*, 2023. [2](#)
- [28] Xin Yu, Yuan-Chen Guo, Yangguang Li, Ding Liang, Song-Hai Zhang, and Xiaojuan Qi. Text-to-3d with classifier score distillation. *arXiv preprint arXiv:2310.19415*, 2023. [3](#)