This ICCV paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

DepR: Depth Guided Single-view Scene Reconstruction with Instance-level Diffusion

Qingcheng Zhao*,1,† Xiang Zhang*,✉,2 Haiyang Xu2 Zeyuan Chen2
Jianwen Xie3 Yuan Gao4 Zhuowen Tu2
1ShanghaiTech University 2UC San Diego 3Lambda, Inc. 4Stanford University
*equal contribution ✉corresponding author

Abstract

We propose DepR, a depth-guided single-view scene reconstruction framework that integrates instance-level diffusion within a compositional paradigm. Instead of reconstructing the entire scene holistically, DepR generates individual objects and subsequently composes them into a coherent 3D layout. Unlike previous methods that use depth solely for object layout estimation during inference and therefore fail to fully exploit its rich geometric information, DepR leverages depth throughout both training and inference. Specifically, we introduce depth-guided conditioning to effectively encode shape priors into diffusion models. During inference, depth further guides DDIM sampling and layout optimization, enhancing alignment between the reconstruction and the input image. Despite being trained on limited synthetic data, DepR achieves state-of-the-art performance and demonstrates strong generalization in single-view scene reconstruction, as shown through evaluations on both synthetic and real-world datasets.

1. Introduction

Single-view scene reconstruction is a long-standing challenge in computer vision, requiring the estimation of 3D object shapes and scene structure from a single 2D image. The task is inherently difficult due to the limited spatial cues available in a single-view observation, making it heavily reliant on prior knowledge to infer plausible 3D geometry.

Previous approaches performing holistic scene reconstruction [5, 6, 65] have successfully leveraged depth to back-project features into 3D volumes, followed by an encoder-decoder architecture to reconstruct scene-level geometries. Depth plays a crucial role in these methods, facilitating effective shape reconstruction. However, the limited resolution and scarcity of scene-level training data hinder their ability to generalize effectively to real images.

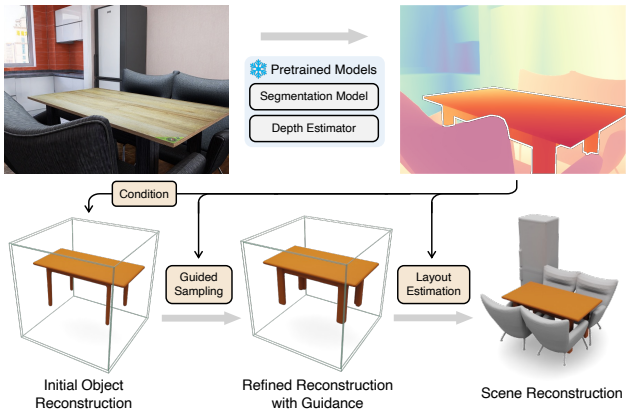

Figure 1. Illustration of DepR. Given a single 2D input image, we leverage depth predicted by a pretrained model to condition object reconstruction and guide sampling, followed by layout estimation to assemble the reconstructed instances into a coherent scene.

Recently, with the advancement of single-view object reconstruction/generation models [19, 24, 31–33] alongside the emergence of large-scale object datasets [7, 8], stage-wise compositional scene generation has gained increasing popularity. Notably, methods such as DeepPriorAssembly [67] and Gen3DSR [9] utilize off-the-shelf 2D segmentation models to extract individual objects from images, then pass the masked images into a pre-trained image-to-shape or single-image-to-multi-view generative model. A monocular depth estimator is subsequently used to optimize object poses, ensuring proper spatial arrangement.

While these methods benefit from the generalizability of pre-trained object reconstruction models, they have two key limitations: (1) they overlook the strong shape priors provided by depth; and (2) they rely on reconstruction models pre-trained on complete object images, neglecting occlusions and scene-level context, which can cause inconsistencies between the reconstruction and the input image.

To overcome these limitations, we propose DepR, a depth-guided single-view scene reconstruction pipeline fol-

†Work done during internship at UC San Diego.
Project page: https://mlpc-ucsd.github.io/DepR

lowing a compositional generation paradigm. Our method begins with off-the-shelf pre-trained segmentation and monocular depth estimation models, but rather than using depth solely for object pose alignment, we incorporate depth priors directly into the object reconstruction network via depth-guided conditioning. To handle partial observations caused by object occlusions, we introduce local-global attention, which leverages scene-level context to enhance the conditioning signal. The depth guidance enables our image-conditioned shape generative model to generalize effectively to real images, even when trained on limited synthetic data. At inference time, we further utilize depth for guided DDIM sampling and layout estimation, ensuring accurate alignment between the reconstruction and the input.

Our contributions can be summarized as follows:

- We introduce a novel framework for single-view scene reconstruction that effectively leverages a pre-trained monocular depth estimator to facilitate both object reconstruction and layout estimation.
- We formulate object reconstruction as an image-conditioned generation task, incorporating *depth-guided conditioning* and *local-global attention* to leverage depth priors while integrating global scene context.
- To improve alignment between the generated reconstructions and input images, we introduce depth-guided sampling, which enhances geometric consistency. Experiments demonstrate that DepR produces accurate reconstructions that closely align with input images.

2. Related Work

2.1. Single-view Object Reconstruction

Recent advancements in large-scale datasets [7, 8] have significantly progressed 3D object reconstruction/generation. Several approaches [22, 31–33, 35, 36, 52, 53] leverage priors from image [41, 44] or video generative models [1] to synthesize multi-view consistent images from a single view. These generated views are processed by feed-forward reconstruction frameworks [50, 51, 59] to construct an underlying 3D representation, such as Gaussian splatting [25] or neural radiance fields (NeRFs) [37]. Other methods directly generate 3D representations, such as triplanes [47], using large Transformer-based architectures [19, 57, 58, 64, 68] or diffusion models [54, 56, 60, 66], eliminating the need for intermediate multi-view image synthesis.

2.2. Single-view Scene Reconstruction

Reconstructing a 3D scene from a single image is a challenging problem due to the inherent depth ambiguity and occlusions. Existing approaches can be broadly categorized into two paradigms based on how they handle scene composition and object representation: holistic scene reconstruction and compositional scene reconstruction.

Holistic scene reconstruction. Holistic scene reconstruction methods [5, 6, 65] reconstruct the entire scene as a whole, grouping individual instances based on predicted instance labels. Monocular depth estimation [15, 20] is often employed as a strong shape prior, facilitating the back-projection of 2D features into a 3D volume. This is typically followed by an encoder-decoder architecture [45] to recover the scene geometry. While these approaches naturally capture object poses and overall scene layout, they often suffer from limited resolution and availability of high-quality training data, leading to poor generalization to real-world images and reduced reconstruction fidelity.

Compositional scene reconstruction. Compositional scene reconstruction approaches decompose a scene into individual objects, reconstruct them separately, and subsequently estimate the scene layout and object poses to reintegrate them into the scene space. Earlier feed-forward methods utilize either explicit mesh representations [16, 21, 30, 39], or implicit signed distance functions (SDFs) [3, 30, 63], trained end-to-end with 3D supervision. While these methods are relatively efficient to train and infer, they exhibit limited generalizability to out-of-domain images, particularly in complex real-world scenarios with variations not seen during training.

Another line of research [14, 17, 23, 27–29, 55, 61] focuses on retrieval-based approaches, where 3D models are selected from a database and aligned to the scene. While these methods benefit from high-quality pre-existing 3D assets, their reconstruction accuracy is inherently constrained by the diversity and completeness of the database. Additionally, insufficient visual cues may lead to misalignment between retrieved models and the actual scene content.

More recently, generation-based methods [4, 9, 18, 67] have emerged as a promising direction, leveraging large-scale generative models for both images [26, 34, 38, 41–44, 46] and 3D shapes [10, 24, 33, 64, 66]. These approaches typically follow a pipeline where objects are first segmented using pre-trained segmentation models [26, 43], then reconstructed from their segmented images using large-scale object reconstruction models [24, 33], and finally refined through pose optimization to ensure spatial coherence within the scene. While the use of large-scale pre-trained object reconstruction models enhances reconstruction quality, these methods lack holistic scene context. Furthermore, since object reconstruction models are generally trained on complete object images, they struggle with occluded or incomplete segmented objects, necessitating amodal completion to infer missing regions.

3. Method

Our method, illustrated in Fig. 2, reconstructs a complete 3D scene from a single RGB image using a multi-stage

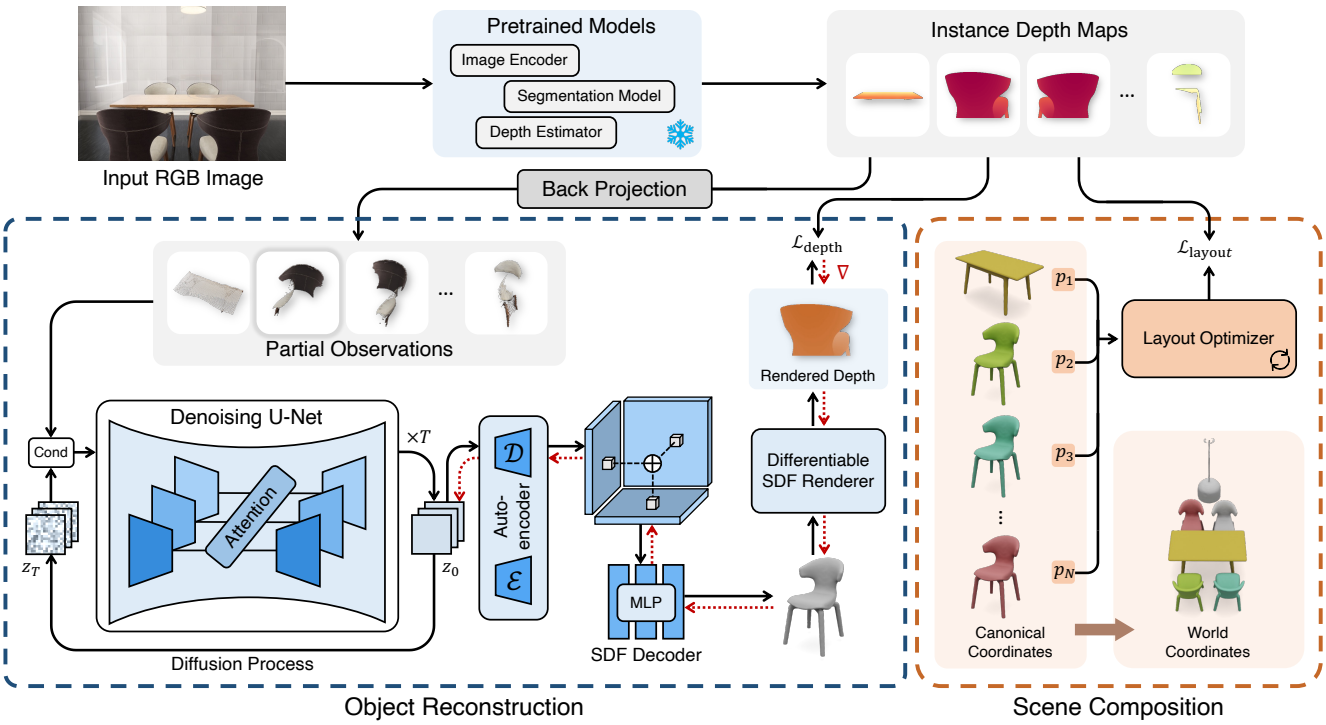

Figure 2. Overview of our method. Depth is utilized in three key stages: 1) to back-project features to condition the latent tri-plane diffusion model to generate complete 3D shapes; 2) to guide the diffusion sampling process via gradients from a depth loss $\mathcal{L}_{\text{depth}}$; and 3) to optimize object poses $\{p_i\}_{i=1}^N$ via $\mathcal{L}_{\text{layout}}$ for accurate scene composition.

pipeline that leverages conditional generation. We begin by formulating the problem addressed by DepR in Sec. 3.1, followed by a description of our object-level reconstruction approach in Sec. 3.2. Finally, Sec. 3.3 details our layout optimization strategy, which composes the reconstructed objects back into a coherent scene.

3.1. Problem Formulation

Single-view scene reconstruction pipelines aim to recover the geometry of a scene from a single RGB image. Following the setting of InstPIFu [30] and DeepPriorAssembly [67], we focus exclusively on object instances (*e.g.* sofas, chairs, *etc.*), while excluding background elements such as walls. Unlike holistic reconstruction approaches such as Uni-3D [65] and PanoRe [6], which treat the entire scene as a single mesh, our stage-wise pipeline reconstructs each object instance individually.

Given an RGB image $I \in \mathbb{R}^{H\times W\times 3}$, we leverage off-the-shelf models to obtain the depth map D, instance segmentation masks $\mathcal{M} = \{M_i\}_{i=1}^N$, and image features F. For each object instance, we extract its depth $D_i = D \odot M_i$ and corresponding image features $F_i = F \odot M_i$, where $\odot$ represents element-wise multiplication. These serve as inputs to the object reconstruction model F_{recon}:

$$O_i = F_{\text{recon}}(F_i, D_i, F) \tag{1}$$

where O_i denotes the reconstructed object shape in its canonical coordinate frame. Due to self-occlusion, the depth D_i captures only the visible front surface of the object. Furthermore, inter-object occlusion often results in incomplete segmentation masks M_i. These limitations make it difficult for feedforward methods such as InstPIFu to accurately reconstruct full object geometry. To address this, we build F_{recon} as a generative model conditioned on partial observations.

To reassemble the full scene, a layout optimizer F_{layout} can be employed to estimate the pose p_i of each object:

$$p_i = F_{\text{layout}}(O_i, D_i) \tag{2}$$

The final scene reconstruction is given by the set $\{O_i, p_i\}_{i=1}^N$, which specifies the geometry and spatial configuration of all object instances.

3.2. Object Reconstruction from Partial Observations

In this section, we introduce our object reconstruction model. Inspired by works such as BlockFusion [56] and Frankenstein [60], we adopt the latent tri-plane representation to enable compact and efficient encoding of 3D geometry. We begin with tri-plane fitting, followed by a variational autoencoder (VAE) that compresses the raw tri-planes into a latent space (Sec. 3.2.1). Latent codes are then generated by

a diffusion model conditioned on partial observations, leveraging depth priors (Sec. 3.2.2). To further improve alignment with the input image, we apply depth-guided sampling during inference (Sec. 3.2.3).

3.2.1. Tri-plane Fitting and VAE Training

We represent 3D shapes using signed distance fields (SDFs), which are decoded from triplane features via a multilayer perceptron (MLP) parameterized by θ. The tri-plane $\mathcal{F} \in \mathbb{R}^{3\times C_{\text{raw}}\times N_{\text{raw}}^2}$ is a tensor built on three axis-aligned 2D planes: XY, YZ, and XZ, to factorize the dense 3D volume grid. Here, N_{raw} denotes the spatial resolution and C_{raw} is the number of feature channels in the raw tri-plane. Given a query point $q \in \mathbb{R}^3$, the function Φ: $\mathbb{R}^3 \mapsto \mathbb{R}$ outputs its signed distance value:

$$\Phi_\theta(q) = \text{MLP}_\theta\left(\bigoplus_{\Pi\in\text{XY,YZ,XZ}} \mathcal{F}_\Pi\Big(\text{Proj}_\Pi(q)\Big)\right) \tag{3}$$

where $\text{Proj}_\Pi(\cdot)$ orthogonally projects the 3D point onto plane Π to retrieve plane features via bilinear interpolation, and $\oplus$ denotes feature concatenation.

To convert all meshes in the dataset into tri-plane representations, we follow the procedure in BlockFusion [56]. For each object, we first convert the geometry into a watertight mesh, then uniformly sample on-surface points Ω_0 and off-surface points Ω. We jointly optimize the MLP parameters θ and the tri-plane features $\mathcal{F}$ using the following geometry loss:

$$\mathcal{L}_{\text{geometry}} = \mathcal{L}_{\text{normal}} + \mathcal{L}_{\text{eikonal}} + \mathcal{L}_{\text{SDF}}. \tag{4}$$

where each term is defined as:

$$\mathcal{L}_{\text{normal}} = \sum_{p\in\Omega_0} \mathbf{n}_p \cdot \hat{\mathbf{n}}_p + \|\mathbf{n}_p - \hat{\mathbf{n}}_p\|_2 \tag{5}$$

$$\mathcal{L}_{\text{eikonal}} = \sum_{p\in\Omega} \|\hat{\mathbf{n}}_p - 1\|_2 \tag{6}$$

$$\mathcal{L}_{\text{SDF}} = \sum_{p\in\Omega\cup\Omega_0} \|\Phi_\theta(p) - s_p\| \tag{7}$$

where $\hat{\mathbf{n}}_p = \nabla\Phi_\theta(p)$ is the predicted surface normal, $\mathbf{n}_p$ is the ground truth surface normal, and s_p is the ground truth signed distance value at point p. For points on the surface ($p \in \Omega_0$), $s_p = 0$. For brevity, we omit the weighting coefficients of individual loss terms.

To compress raw tri-planes into a lower-dimensional latent space, we employ a variational autoencoder (VAE). The encoder $\mathcal{E}$ maps $\mathcal{F}$ to a latent code $z \in \mathbb{R}^{3\times C_{\text{latent}}\times N_{\text{latent}}^2}$, where $C_{\text{latent}} < C_{\text{raw}}$ and $N_{\text{latent}} < N_{\text{raw}}$. The decoder $\mathcal{D}$ reconstructs the raw tri-planes from the latent representation. The VAE is trained to minimize the following objective:

$$\mathcal{L}_{\text{VAE}} = \mathcal{L}_{\text{recon}} + \mathcal{L}_{\text{KL}} + \mathcal{L}_{\text{geometry}} \tag{8}$$

where $\mathcal{L}_{\text{recon}}$ is the L1 loss between $\mathcal{F}$ and $\mathcal{D}(\mathcal{E}(\mathcal{F}))$, $\mathcal{L}_{\text{KL}}$ denotes KL divergence loss, and $\mathcal{L}_{\text{geometry}}$ is computed using the decoded latent.

3.2.2. Depth-conditioned Denoising

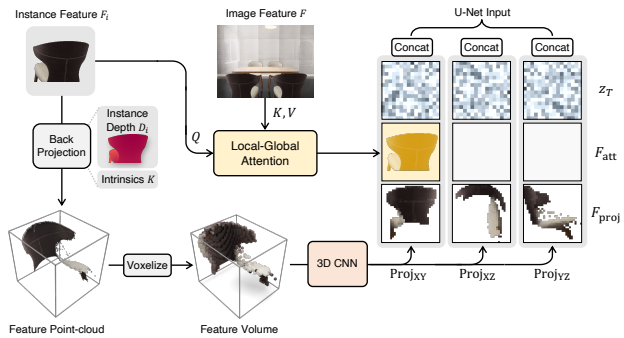

Figure 3. Illustration of depth-guided conditioning. `Concat` denotes channel-wise concatenation. The attention-enhanced feature F_{att} is applied to the XY plane, while zero padding is used for the remaining two planes.

After obtaining latent tri-plane representations for all objects, we design a diffusion model that generates latents for complete shapes, conditioned on partial observations. Rather than injecting image features through naive concatenation or cross-attention, we leverage strong geometric priors provided by depth.

Specifically, we back-project the instance features F_i using the estimated depth D_i and camera intrinsics K, and voxelize the resulting 3D points into a feature volume of size N_{latent}^3, where the spatial resolution N_{latent} is chosen to match that of the latent tri-planes. Features falling into the same voxel are aggregated via average pooling. The voxelized feature volume V is then processed by a 3D CNN and a projection layer: $V_{\text{conv}} = \text{Conv3D}(V)$. Instead of reconstructing shapes directly from these features, as in previous works [6, 65], we use them as conditioning inputs for our latent diffusion model. This generative formulation leverages shape priors from depth to improve shape fidelity, particularly in the presence of occlusion. After volumetric feature extraction, we project V_{conv} onto the three orthogonal planes (XY, YZ, XZ) using $\text{Proj}_\Pi(\cdot)$ as defined in Eq. (3), followed by average pooling along the projection axis. The resulting 3-view feature maps $F_{\text{proj}} = \{\text{Proj}_\Pi(V_{\text{conv}}) \mid \Pi \in \{\text{XY}, \text{YZ}, \text{XZ}\}\}$ provide a structured and spatially aligned representation of the partial observations, enriched with shape priors from the depth.

To incorporate global scene context, *e.g.*, cues from other visible instances of the same category or shape, we introduce a local-global attention mechanism. It enhances the local object feature with global scene-level information. Given the instance features F_i and the image features F, we

compute cross-attention as:

$$F_{\text{att}} = \text{CrossAttention}(q = F_i, k = F, v = F) \tag{9}$$

This local-global attention allows our model to incorporate broader visual context when reconstructing occluded or ambiguous regions.

The 3-view features F_{proj}, together with the attention-enhanced features F_{att}, are passed to a 3D-Aware U-Net [56] denoiser $\mathcal{G}$. These features are concatenated with the noised latent z_t along the feature dimension and used as input to the network. An overview of the conditioning process is illustrated in Fig. 3. The model predicts the clean latent z_0 as

$$\hat{z}_0 = \mathcal{G}\left(\text{Concat}(z_t, F_{\text{proj}}, F_{\text{att}}), t\right) \tag{10}$$

We train the denoiser using the standard diffusion objective:

$$\mathcal{L}_{\text{diffusion}} = \mathbb{E}_{z_0,\epsilon,t}\left[\|z_0 - \hat{z}_0\|_2^2\right] \tag{11}$$

where $z_t = \sqrt{\bar{\alpha}_t} z_0 + \sqrt{1-\bar{\alpha}_t}\epsilon$, with $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and $\bar{\alpha}_t$ denoting the noise schedule. To further preserve geometric fidelity, we incorporate the geometry loss $\mathcal{L}_{\text{geometry}}$ from Sec. 3.2.1. The final training objective is

$$\mathcal{L}_{\text{recon}} = \mathcal{L}_{\text{diffusion}} + \mathcal{L}_{\text{geometry}} \tag{12}$$

This conditioning framework enables our model to effectively integrate partial 3D observations with image-based contextual cues, producing accurate and high-fidelity 3D shapes. By combining depth guidance with local-global attention, our method remains robust to occlusions and ambiguities, resulting in consistent and detailed reconstructions.

3.2.3. Depth-guided Sampling

Generative models may produce inaccurate shapes due to insufficient observations. However, depth estimators provide valuable priors for object shapes. To incorporate this information during inference, we adopt classifier-guided DDIM [48] sampling, using depth to guide SDF generation. Specifically, we render the predicted sample $\hat{z}_0$ obtained during the diffusion process into a depth map $\hat{D}_i$ using the differentiable SDF depth renderer from MonoSDF [62]. We then compute the depth loss $\mathcal{L}_{\text{depth}}$ against the estimated instance depth D_i, and use the gradient $\nabla_{\hat{z}_0}\mathcal{L}_{\text{depth}}$ to guide the sampling trajectory.

To handle depth scale ambiguity in depth supervision, we employ scale-invariant loss [11], defined as

$$\mathcal{L}_{\text{depth}} = \frac{1}{m}\sum_{j=1}^{m}\left\|\left(\log \hat{D}_i^{(j)} - \log D_i^{(j)} + \alpha(\hat{D}_i, D_i)\right)\right\|_2 \tag{13}$$

where

$$\alpha(\hat{D}_i, D) = \frac{1}{m}\sum_{j=1}^{m}\left(\log D_i^{(j)} - \log \hat{D}_i^{(j)}\right) \tag{14}$$

m is the number of valid pixels for instance depth D_i.

During diffusion sampling, we apply depth guidance to improve the geometric accuracy of the generated shape. At each timestep t, after obtaining the predicted clean latent $\hat{z}_0$, we compute the gradient of the depth loss and use it to update the predicted noise ϵ_t as:

$$\epsilon_t' = \epsilon_t - \lambda_{\text{guide}}\sqrt{1-\bar{\alpha}_t}\nabla_{\hat{z}_0}\mathcal{L}_{\text{depth}} \tag{15}$$

where λ_{guide} is a guidance strength parameter that controls the influence of the depth prior, and $\sqrt{1-\bar{\alpha}_t}$ scales the gradient according to the current noise level. The adjusted noise ϵ_t' is then used in the DDIM update rule to compute the next latent z_{t-1}.

This depth-guided sampling process effectively steers generation toward geometry consistent with observed depth, leading to more accurate reconstruction, particularly for occluded objects.

To ensure proper alignment with the estimated depth map, we adopt a two-stage sampling strategy. In the first stage, we generate an initial 3D shape without applying depth guidance. Based on this initial output, we estimate the object pose using the layout estimator F_{layout}, which will be described in Sec. 3.3. Once the transformation parameters are obtained, we perform a second sampling pass with depth guidance. In this pass, the intermediate shape is transformed from canonical to camera space to render the depth map, and the scale-invariant loss is computed to guide further refinement.

3.3. Scene Composition

As described in Sec. 3.2, objects are reconstructed in their canonical coordinate systems, which enables flexibility in handling variations in pose and scale commonly found in indoor environments. Following DeepPriorAssembly [67], we formulate scene composition as a shape registration problem, where each object's transformation is parameterized by translation (3 parameters), rotation (1 parameter) around the vertical axis, and scale (1 parameter). To account for camera roll and pitch, we additionally apply a global rotation that is shared across all objects.

We randomly initialize these learnable transformation parameters, denoted as ϕ. The sampled point cloud P_i' from the reconstructed surface O_i of object i is transformed into the camera coordinate system as $P_i = f_\phi(P_i')$, where f_ϕ is the parameterized transformation function. We optimize ϕ using gradient descent. The optimization objective is defined as:

$$\mathcal{L}_{\text{layout}} = \mathcal{L}_{\text{CD}}(\tilde{P}_i, P_i) + \mathcal{L}_{\text{CD}}(\mathcal{P}(\tilde{P}_i), \mathcal{P}(P_i)) \tag{16}$$

where $\tilde{P}_i$ is the partial point cloud from the depth map. $\mathcal{L}_{\text{CD}}$ represents the Chamfer distance loss. The operator $\mathcal{P}$ denotes projection from 3D camera coordinates to 2D pixel

Method	Scene-level			Object-level		Runtime
	CD ($\times 10^{-3}, \downarrow$)	CD-S ($\times 10^{-3}, \downarrow$)	F-Score (%, $\uparrow$)	CD ($\times 10^{-3}, \downarrow$)	F-Score (%, $\uparrow$)	
InstPIFu [30]	213.4	124.9	13.72	44.74	29.63	19.8 s
Uni-3D [65]	218.3	113.3	12.99	—	—	3.1 s
BUOL [5]	282.7	154.7	12.57	—	—	6.5 s
Gen3DSR [9]	222.4	137.5	13.52	9.74	31.42	15.5 min
DeepPriorAssembly [67]	191.8	76.2	16.72	20.13	27.83	6.7 min
DepR (Ours)	**153.2**	**56.4**	**25.00**	**2.57**	**89.66**	1.2 min

Table 1. Quantitative comparison with state-of-the-art methods on the 3D-FRONT [12] test set. We report Chamfer distance (CD; lower is better) and F-score (higher is better) at scene and object levels, along with single-direction Chamfer distance (CD-S) at the scene level.

space, where the Chamfer distance is computed between the corresponding 2D point sets.

We select the optimized parameters ϕ that yield the minimum 3D Chamfer distance (the first term in Eq. (16)) as the final object poses for scene reconstruction.

4. Experiments

4.1. Datasets

We conduct our experiments using the 3D-FRONT [12] dataset, specifically the version processed by InstPIFu [30]. In addition to the original object meshes from the 3D-FUTURE [13] dataset and scene layouts, it provides realistic rendered images and rich annotations such as depth maps, instance segmentation masks, and object labels. We use 22,673 scene images for training, and randomly sample 969 images for object-level evaluation and 156 images for scene-level evaluation.

4.2. Implementation Details

We use off-the-shelf models to ensure robustness and broad applicability: Grounded-SAM [43] for instance segmentation, Depth Pro [2] for depth estimation, and DINOv2 [40] as the image encoder. Our raw tri-plane representation has a resolution of 128×128 with 32 feature channels, while the latent triplane is compressed to 32×32 with 2 channels. Prior to training, we perform a one-time tri-plane fitting across the entire 3D-FUTURE [13] dataset, which contains approximately 16,000 shapes. This procedure is parallelized and takes around 24 hours on CPUs. The VAE and diffusion models are trained on four NVIDIA A6000 GPUs, for roughly 168 hours and 96 hours, respectively. During guided sampling, we set the gradient scale s to 100.

4.3. Metrics

We evaluate our method using Chamfer distance (CD) and F-score (with a threshold of 0.002), following InstPIFu. Each reconstructed mesh is uniformly sampled with 10k points for metric computation.

We report metrics at two levels:

- **Object-level:** To assess object quality, we decompose the predicted scene into individual instances and compare each against its corresponding ground truth mesh. For fair comparison, both predicted and ground-truth meshes are normalized to a unit sphere.
- **Scene-level:** We evaluate the full reconstructed scene, capturing both object geometry and their spatial layout. Following DeepPriorAssembly [67], we additionally report the single-direction Chamfer distance (CD-S) from the generated scene to the ground truth.

4.4. Results

We benchmark our method against two categories of approaches: (1) generative-model-based methods, including Gen3DSR [9] and DeepPriorAssembly [67]; and (2) feed-forward reconstruction methods, such as InstPIFu, Uni-3D [65], and BUOL [5]. Since Uni-3D and BUOL perform holistic scene reconstruction without generating individual object meshes, object-level evaluation is not applicable for these methods.

Tab. 1 presents a quantitative comparison between our method and state-of-the-art approaches on the 3D-FRONT test set. DepR outperforms all baselines across both object-level and scene-level metrics. Notably, at the object level, it achieves a substantial performance gain, highlighting the effectiveness of our depth-guided latent diffusion in generating high-fidelity geometries, even under occlusion. At the scene level, our method consistently surpasses existing approaches, showcasing its ability to more accurately recover spatial relationships and maintain global consistency. Furthermore, our method achieves significantly lower inference time per scene (measured on a single NVIDIA A100 GPU) compared to other generative-model-based compositional approaches, highlighting its efficiency alongside reconstruction quality.

We present qualitative comparisons in Fig. 4 on the synthetic 3D-FRONT dataset, and in Fig. 5 on real-world images from Pix3D [49] and our own collected samples. In both settings, DepR yields more accurate and visually coherent reconstructions. Its depth-guided design enhances object geometries and spatial arrangements, resulting in

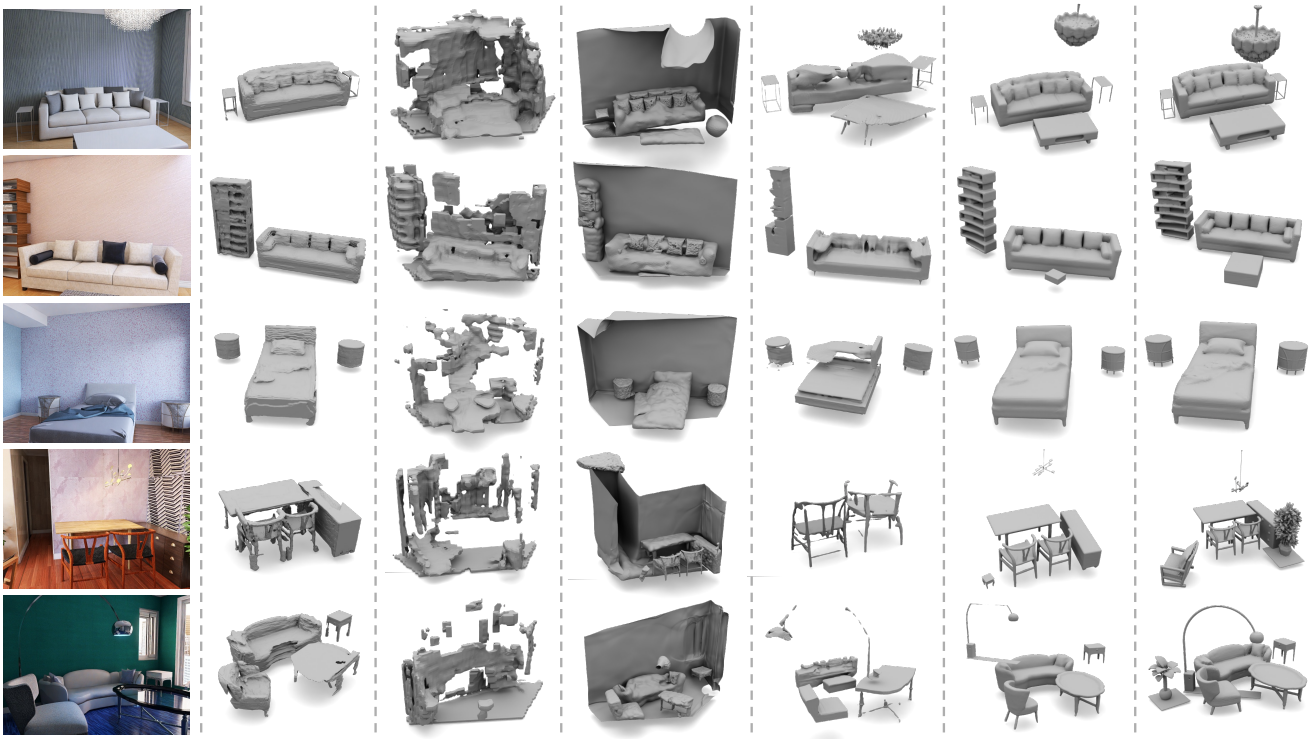

Figure 4. Qualitative comparisons the on the 3D-FRONT [12] synthetic dataset against state-of-the-art methods.

outputs that closely align with the input images. Notably, our approach generalizes well to real-world scenarios, successfully reconstructing scenes from diverse environments and demonstrating strong robustness and practicality.

4.5. Ablation Study

Framework design. To evaluate the efficacy of each proposed component, we conduct ablation studies and report object reconstruction results in Tab. 2. Depth-guided conditioning yields the most significant improvement, boosting F-score by 27.76. Local-global attention provides a 2.3-point gain, while depth-guided sampling offers a smaller but noticeable enhancement. Qualitative results in Fig. 6 further illustrate that guided sampling improves alignment of fine details with the input image.

Method	Object-level	
	CD ($\times 10^{-3}$, ↓)	F-Score (↑)
Full model	**2.57**	**89.66**
w/o Depth-guided Conditioning	11.87	61.90
w/o Local-global Attention	3.14	87.36
w/o Depth-guided Sampling	2.61	89.64

Table 2. Ablation studies on the proposed components.

Effects of depth and layout in object reconstruction. Depth is used both for conditioning object reconstruction and for layout estimation during guided sampling. In Tab. 3, we present ablations using ground-truth depth for conditioning and layout estimation, and additionally comparing with ground-truth layout to estimate the theoretical upper bound of model performance. Results show that ground-truth depth further improves object reconstruction quality. Moreover, guided sampling with ground-truth layout yields a more notable improvement (approximately +0.3 F-score) over the non-guided variant, regardless of whether ground-truth depth is used for conditioning. This suggests that improved layout estimation can further benefit the model's performance.

GT Depth	GT Layout	Object-level	
		CD ($\times 10^{-3}$, ↓)	F-Score (↑)
w/o Depth-guided Sampling			
	–	2.61	89.64
✓	–	2.42	90.33
with Depth-guided Sampling			
		2.57	89.66
	✓	2.55	89.92
✓		2.37	90.24
✓	✓	**2.29**	**90.62**

Table 3. Effects of depth and layout in object reconstruction.

Effects of upstream errors in scene reconstruction. DepR relies on external models for depth estimation and segmentation, making it susceptible to error propagation. In Tab. 4, we show ablations using ground-truth depth, segmentation, and layout to assess their individual impact on

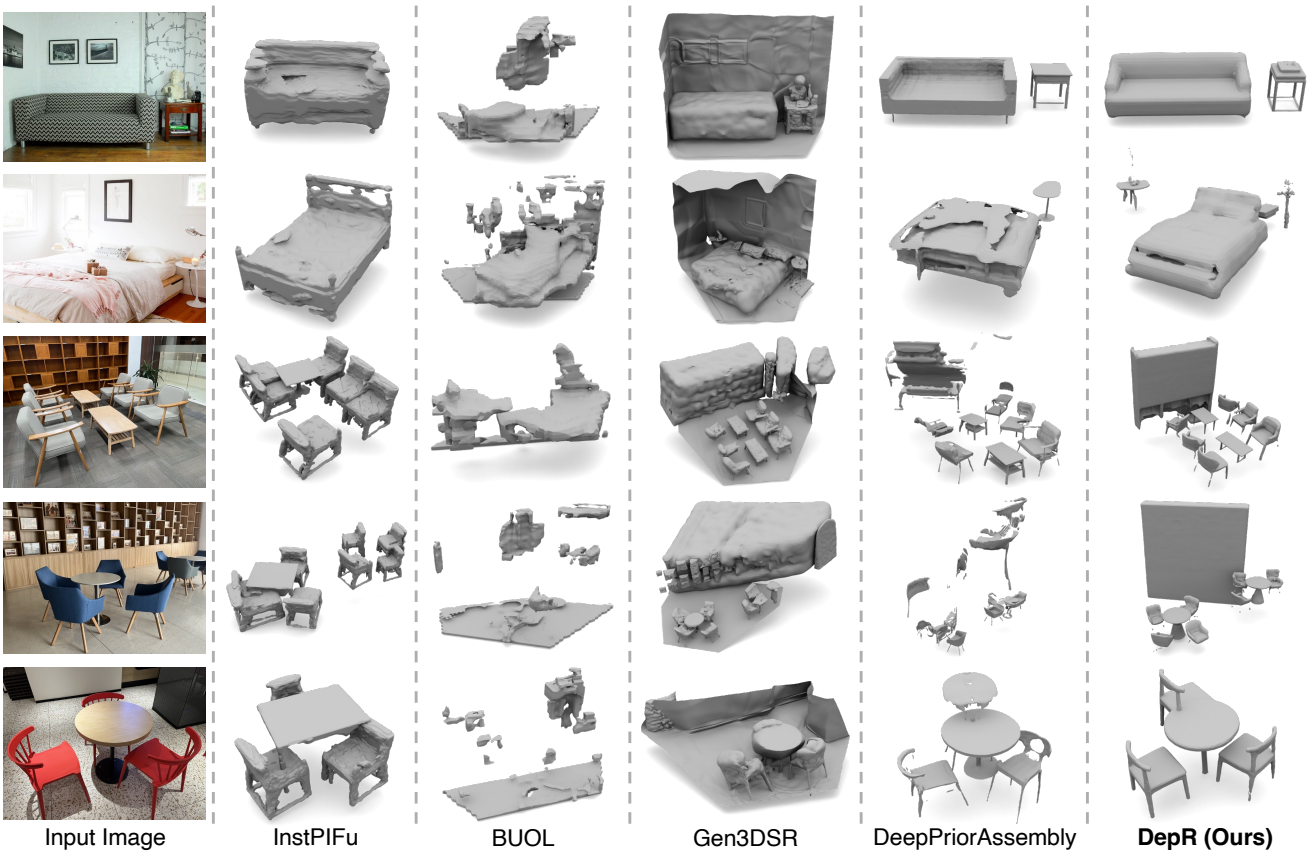

Figure 5. Qualitative comparisons on real images, including Pix3D [49] (top two rows) and our own collected images (bottom three rows).

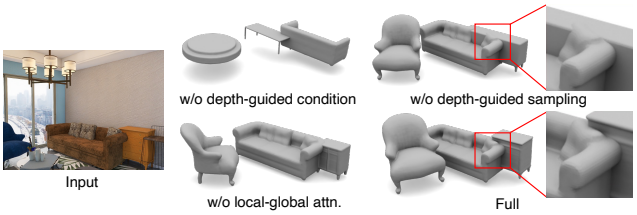

Figure 6. Qualitative comparisons on proposed components.

GT Inputs			Scene-level		
Depth	Segm	Layout	CD ($\times 10^{-3}$, ↓)	CD-S ($\times 10^{-3}$, ↓)	F-Score (↑)
			155.0	57.3	24.63
✓			152.4	61.4	26.09
	✓		103.4	34.1	30.47
	✓	✓	<u>20.2</u>	<u>10.5</u>	**67.65**
✓	✓		91.6	28.0	33.14
✓	✓	✓	**18.8**	**9.5**	<u>67.42</u>

Table 4. Effects of upstream (depth, segmentation, and layout) errors in scene reconstruction. Note that ground-truth layout implies ground-truth segmentation.

scene-level reconstruction. We disable depth-guided sampling to isolate the effect of layout from object reconstruction. Note that ground-truth layout implies ground-truth segmentation due to their one-to-one correspondence. Results show that while ground-truth depth improves layout estimation, segmentation quality has a more pronounced effect as missing or incorrectly segmented objects directly degrade scene composition. The main limitation stems from optimization-based layout estimation; with ground-truth layout, scene-level F-score exceeds 67.

5. Conclusion

In this paper, we present DepR, a depth-guided single-view scene reconstruction framework that integrates depth priors into both object reconstruction and scene layout estimation. By leveraging depth-guided conditioning, our model achieves higher shape fidelity and improved robustness to occlusions. A local-global attention mechanism further incorporates scene-level context to refine object reconstructions. Additionally, we introduce depth-guided DDIM sampling to enhance geometric consistency between the reconstructed 3D scene and the input image. Extensive quantitative and qualitative evaluations demonstrate that DepR outperforms state-of-the-art methods at both the object and scene levels, while generalizing effectively to real-world images. Our approach highlights the critical role of depth priors in single-view reconstruction and lays the groundwork for more robust and scalable 3D scene understanding.

Acknowledgment

This work is supported by NSF award IIS-2127544 and NSF award IIS-2433768. We thank Lambda, Inc. for their compute resource help and travel grant.

References

[1] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023. 2

[2] Alexey Bochkovskiy, Amaël Delaunoy, Hugo Germain, Marcel Santos, Yichao Zhou, Stephan Richter, and Vladlen Koltun. Depth pro: Sharp monocular metric depth in less than a second. In *The Thirteenth International Conference on Learning Representations*, 2025. 6

[3] Yixin Chen, Junfeng Ni, Nan Jiang, Yaowei Zhang, Yixin Zhu, and Siyuan Huang. Single-view 3d scene reconstruction with high-fidelity shape and texture. In *2024 International Conference on 3D Vision (3DV)*, pages 1456–1467. IEEE, 2024. 2

[4] Yongwei Chen, Tengfei Wang, Tong Wu, Xingang Pan, Kui Jia, and Ziwei Liu. Comboverse: Compositional 3d assets creation using spatially-aware diffusion guidance. In *European Conference on Computer Vision*, pages 128–146. Springer, 2024. 2

[5] Tao Chu, Pan Zhang, Qiong Liu, and Jiaqi Wang. Buol: A bottom-up framework with occupancy-aware lifting for panoptic 3d scene reconstruction from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4937–4946, 2023. 1, 2, 6

[6] Manuel Dahnert, Ji Hou, Matthias Niessner, and Angela Dai. Panoptic 3d scene reconstruction from a single rgb image. In *Advances in Neural Information Processing Systems*, pages 8282–8293. Curran Associates, Inc., 2021. 1, 2, 3, 4

[7] Matt Deitke, Ruoshi Liu, Matthew Wallingford, Huong Ngo, Oscar Michel, Aditya Kusupati, Alan Fan, Christian Laforte, Vikram Voleti, Samir Yitzhak Gadre, Eli VanderBilt, Aniruddha Kembhavi, Carl Vondrick, Georgia Gkioxari, Kiana Ehsani, Ludwig Schmidt, and Ali Farhadi. Objaverse-xl: A universe of 10m+ 3d objects. In *Advances in Neural Information Processing Systems*, pages 35799–35813. Curran Associates, Inc., 2023. 1, 2

[8] Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3d objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13142–13153, 2023. 1, 2

[9] Andreea Dogaru, Mert Özer, and Bernhard Egger. Gen3DSR: Generalizable 3d scene reconstruction via divide and conquer from a single view. In *International Conference on 3D Vision 2025*, 2025. 1, 2, 6

[10] Ainaz Eftekhar, Alexander Sax, Jitendra Malik, and Amir Zamir. Omnidata: A scalable pipeline for making multi-task mid-level vision datasets from 3d scans. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10786–10796, 2021. 2

[11] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2014. 5

[12] Huan Fu, Bowen Cai, Lin Gao, Ling-Xiao Zhang, Jiaming Wang, Cao Li, Qixun Zeng, Chengyue Sun, Rongfei Jia, Binqiang Zhao, and Hao Zhang. 3d-front: 3d furnished rooms with layouts and semantics. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10933–10942, 2021. 6, 7

[13] Huan Fu, Rongfei Jia, Lin Gao, Mingming Gong, Binqiang Zhao, Steve Maybank, and Dacheng Tao. 3d-future: 3d furniture shape with texture. *International Journal of Computer Vision*, 129:3313–3337, 2021. 6

[14] Daoyi Gao, Dávid Rozenberszki, Stefan Leutenegger, and Angela Dai. Diffcad: Weakly-supervised probabilistic cad model retrieval and alignment from an rgb image. *ACM Transactions on Graphics (TOG)*, 43(4):1–15, 2024. 2

[15] Naiyu Gao, Fei He, Jian Jia, Yanhu Shan, Haoyang Zhang, Xin Zhao, and Kaiqi Huang. Panopticdepth: A unified framework for depth-aware panoptic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1632–1642, 2022. 2

[16] Georgia Gkioxari, Nikhila Ravi, and Justin Johnson. Learning 3d object shape and layout without 3d supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1695–1704, 2022. 2

[17] Can Gümeli, Angela Dai, and Matthias Nießner. Roca: Robust cad model retrieval and alignment from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4022–4031, 2022. 2

[18] Haonan Han, Rui Yang, Huan Liao, Jiankai Xing, Zunnan Xu, Xiaoming Yu, Junwei Zha, Xiu Li, and Wanhua Li. Reparo: Compositional 3d assets generation with differentiable 3d layout alignment. *arXiv preprint arXiv:2405.18525*, 2024. 2

[19] Yicong Hong, Kai Zhang, Jiuxiang Gu, Sai Bi, Yang Zhou, Difan Liu, Feng Liu, Kalyan Sunkavalli, Trung Bui, and Hao Tan. LRM: Large reconstruction model for single image to 3d. In *The Twelfth International Conference on Learning Representations*, 2024. 1, 2

[20] Junjie Hu, Mete Ozay, Yan Zhang, and Takayuki Okatani. Revisiting single image depth estimation: Toward higher resolution maps with accurate object boundaries. In *2019 IEEE winter conference on applications of computer vision (WACV)*, pages 1043–1051. IEEE, 2019. 2

[21] Siyuan Huang, Siyuan Qi, Yixin Zhu, Yinxue Xiao, Yuanlu Xu, and Song-Chun Zhu. Holistic 3d scene parsing and reconstruction from a single rgb image. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 187–203, 2018. 2

[22] Zehuan Huang, Hao Wen, Junting Dong, Yaohui Wang, Yangguang Li, Xinyuan Chen, Yan-Pei Cao, Ding Liang, Yu

Qiao, Bo Dai, and Lu Sheng. Epidiff: Enhancing multi-view synthesis via localized epipolar-constrained diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9784–9794, 2024. 2

[23] Hamid Izadinia, Qi Shan, and Steven M. Seitz. Im2cad. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2

[24] Heewoo Jun and Alex Nichol. Shap-e: Generating conditional 3d implicit functions. *arXiv preprint arXiv:2305.02463*, 2023. 1, 2

[25] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkuehler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics (TOG)*, 42(4):1–14, 2023. 2

[26] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollar, and Ross Girshick. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4015–4026, 2023. 2

[27] Weicheng Kuo, Anelia Angelova, Tsung-Yi Lin, and Angela Dai. Mask2cad: 3d shape prediction by learning to segment and retrieve. In *European Conference on Computer Vision*, pages 260–277. Springer, 2020. 2

[28] Weicheng Kuo, Anelia Angelova, Tsung-Yi Lin, and Angela Dai. Patch2cad: Patchwise embedding learning for in-the-wild shape retrieval from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 12589–12599, 2021.

[29] Florian Langer, Gwangbin Bae, Ignas Budvytis, and Roberto Cipolla. Sparc: Sparse render-and-compare for cad model alignment in a single rgb image. *arXiv preprint arXiv:2210.01044*, 2022. 2

[30] Haolin Liu, Yujian Zheng, Guanying Chen, Shuguang Cui, and Xiaoguang Han. Towards high-fidelity single-view holistic reconstruction of indoor scenes. In *European Conference on Computer Vision*, pages 429–446. Springer, 2022. 2, 3, 6

[31] Minghua Liu, Chao Xu, Haian Jin, Linghao Chen, Mukund Varma T, Zexiang Xu, and Hao Su. One-2-3-45: Any single image to 3d mesh in 45 seconds without per-shape optimization. In *Advances in Neural Information Processing Systems*, pages 22226–22246. Curran Associates, Inc., 2023. 1, 2

[32] Minghua Liu, Ruoxi Shi, Linghao Chen, Zhuoyang Zhang, Chao Xu, Xinyue Wei, Hansheng Chen, Chong Zeng, Jiayuan Gu, and Hao Su. One-2-3-45++: Fast single image to 3d objects with consistent multi-view generation and 3d diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10072–10083, 2024.

[33] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9298–9309, 2023. 1, 2

[34] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European Conference on Computer Vision*, pages 38–55. Springer, 2024. 2

[35] Yuan Liu, Cheng Lin, Zijiao Zeng, Xiaoxiao Long, Lingjie Liu, Taku Komura, and Wenping Wang. Syncdreamer: Generating multiview-consistent images from a single-view image. In *The Twelfth International Conference on Learning Representations*, 2024. 2

[36] Xiaoxiao Long, Yuan-Chen Guo, Cheng Lin, Yuan Liu, Zhiyang Dou, Lingjie Liu, Yuexin Ma, Song-Hai Zhang, Marc Habermann, Christian Theobalt, and Wenping Wang. Wonder3d: Single image to 3d using cross-domain diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9970–9980, 2024. 2

[37] 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. *Communications of the ACM*, 65(1):99–106, 2021. 2

[38] Alexander Quinn Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob Mcgrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. In *International Conference on Machine Learning*, pages 16784–16804. PMLR, 2022. 2

[39] 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 (CVPR)*, 2020. 2

[40] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel HAZIZA, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. Featured Certification. 6

[41] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: Improving latent diffusion models for high-resolution image synthesis. In *The Twelfth International Conference on Learning Representations*, 2024. 2

[42] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1 (2):3, 2022.

[43] Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, et al. Grounded sam: Assembling open-world models for diverse visual tasks. *arXiv preprint arXiv:2401.14159*, 2024. 2, 6

[44] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, 2022. 2

[45] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241. Springer, 2015. 2

[46] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. In *Advances in Neural Information Processing Systems*, pages 36479–36494. Curran Associates, Inc., 2022. 2

[47] J. Ryan Shue, Eric Ryan Chan, Ryan Po, Zachary Ankner, Jiajun Wu, and Gordon Wetzstein. 3d neural field generation using triplane diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20875–20886, 2023. 2

[48] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021. 5

[49] Xingyuan Sun, Jiajun Wu, Xiuming Zhang, Zhoutong Zhang, Chengkai Zhang, Tianfan Xue, Joshua B. Tenenbaum, and William T. Freeman. Pix3d: Dataset and methods for single-image 3d shape modeling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 6, 8

[50] Jiaxiang Tang, Zhaoxi Chen, Xiaokang Chen, Tengfei Wang, Gang Zeng, and Ziwei Liu. Lgm: Large multi-view gaussian model for high-resolution 3d content creation. In *European Conference on Computer Vision*, pages 1–18. Springer, 2024. 2

[51] Vikram Voleti, Chun-Han Yao, Mark Boss, Adam Letts, David Pankratz, Dmitry Tochilkin, Christian Laforte, Robin Rombach, and Varun Jampani. Sv3d: Novel multi-view synthesis and 3d generation from a single image using latent video diffusion. In *European Conference on Computer Vision*, pages 439–457. Springer, 2024. 2

[52] Zhengyi Wang, Yikai Wang, Yifei Chen, Chendong Xiang, Shuo Chen, Dajiang Yu, Chongxuan Li, Hang Su, and Jun Zhu. Crm: Single image to 3d textured mesh with convolutional reconstruction model. In *European Conference on Computer Vision*, pages 57–74. Springer, 2024. 2

[53] Kailu Wu, Fangfu Liu, Zhihan Cai, Runjie Yan, Hanyang Wang, Yating Hu, Yueqi Duan, and Kaisheng Ma. Unique3d: High-quality and efficient 3d mesh generation from a single image. In *Advances in Neural Information Processing Systems*, pages 125116–125141. Curran Associates, Inc., 2024. 2

[54] Shuang Wu, Youtian Lin, Feihu Zhang, Yifei Zeng, Jingxi Xu, Philip Torr, Xun Cao, and Yao Yao. Direct3d: Scalable image-to-3d generation via 3d latent diffusion transformer. In *Advances in Neural Information Processing Systems*, pages 121859–121881. Curran Associates, Inc., 2024. 2

[55] Yuefan Wu, Zeyuan Chen, Shaowei Liu, Zhongzheng Ren, and Shenlong Wang. Casa: Category-agnostic skeletal animal reconstruction. In *Advances in Neural Information Processing Systems*, pages 28559–28574. Curran Associates, Inc., 2022. 2

[56] Zhennan Wu, Yang Li, Han Yan, Taizhang Shang, Weixuan Sun, Senbo Wang, Ruikai Cui, Weizhe Liu, Hiroyuki Sato, Hongdong Li, et al. Blockfusion: Expandable 3d scene generation using latent tri-plane extrapolation. *ACM Transactions on Graphics (TOG)*, 43(4):1–17, 2024. 2, 3, 4, 5

[57] Haiyang Xu, Yu Lei, Zeyuan Chen, Xiang Zhang, Yue Zhao, Yilin Wang, and Zhuowen Tu. Bayesian diffusion models for 3d shape reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10628–10638, 2024. 2

[58] Jiale Xu, Weihao Cheng, Yiming Gao, Xintao Wang, Shenghua Gao, and Ying Shan. Instantmesh: Efficient 3d mesh generation from a single image with sparse-view large reconstruction models. *arXiv preprint arXiv:2404.07191*, 2024. 2

[59] Yinghao Xu, Zifan Shi, Wang Yifan, Hansheng Chen, Ceyuan Yang, Sida Peng, Yujun Shen, and Gordon Wetzstein. Grm: Large gaussian reconstruction model for efficient 3d reconstruction and generation. In *European Conference on Computer Vision*, pages 1–20. Springer, 2024. 2

[60] Han Yan, Yang Li, Zhennan Wu, Shenzhou Chen, Weixuan Sun, Taizhang Shang, Weizhe Liu, Tian Chen, Xiaqiang Dai, Chao Ma, et al. Frankenstein: Generating semantic-compositional 3d scenes in one tri-plane. In *SIGGRAPH Asia 2024 Conference Papers*, pages 1–11, 2024. 2, 3

[61] Kai Yan, Fujun Luan, Miloš Hašan, Thibault Groueix, Valentin Deschaintre, and Shuang Zhao. Psdr-room: Single photo to scene using differentiable rendering. In *SIGGRAPH Asia 2023 Conference Papers*, pages 1–11, 2023. 2

[62] Zehao Yu, Songyou Peng, Michael Niemeyer, Torsten Sattler, and Andreas Geiger. Monosdf: Exploring monocular geometric cues for neural implicit surface reconstruction. In *Advances in Neural Information Processing Systems*, pages 25018–25032. Curran Associates, Inc., 2022. 5

[63] Cheng Zhang, Zhaopeng Cui, Yinda Zhang, Bing Zeng, Marc Pollefeys, and Shuaicheng Liu. Holistic 3d scene understanding from a single image with implicit representation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8833–8842, 2021. 2

[64] Longwen Zhang, Ziyu Wang, Qixuan Zhang, Qiwei Qiu, Anqi Pang, Haoran Jiang, Wei Yang, Lan Xu, and Jingyi Yu. Clay: A controllable large-scale generative model for creating high-quality 3d assets. *ACM Transactions on Graphics (TOG)*, 43(4):1–20, 2024. 2

[65] Xiang Zhang, Zeyuan Chen, Fangyin Wei, and Zhuowen Tu. Uni-3d: A universal model for panoptic 3d scene reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9256–9266, 2023. 1, 2, 3, 4, 6

[66] Zibo Zhao, Wen Liu, Xin Chen, Xianfang Zeng, Rui Wang, Pei Cheng, BIN FU, Tao Chen, Gang Yu, and Shenghua Gao. Michelangelo: Conditional 3d shape generation based on shape-image-text aligned latent representation. In *Advances in Neural Information Processing Systems*, pages 73969–73982. Curran Associates, Inc., 2023. 2

[67] Junsheng Zhou, Yu-Shen Liu, and Zhizhong Han. Zero-shot scene reconstruction from single images with deep prior assembly. *Advances in Neural Information Processing Systems*, 37:39104–39127, 2024. 1, 2, 3, 5, 6

[68] Zi-Xin Zou, Zhipeng Yu, Yuan-Chen Guo, Yangguang Li, Ding Liang, Yan-Pei Cao, and Song-Hai Zhang. Triplane meets gaussian splatting: Fast and generalizable single-view 3d reconstruction with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10324–10335, 2024. 2