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

Supplementary Material

1. Detailed Runtime Analysis

Tab. 1 reports the per-module inference time for DepR, averaged per scene and measured on a single NVIDIA A100 GPU. The 2D preprocessing stage, which includes segmentation and depth estimation using pre-trained models, takes 1.6 s in total. Since diffusion is performed over latent tri-planes of size 2×32^2, as opposed to the raw tri-plane of size 32×128^2, the process remains efficient, requiring only 1.2 s for 50 DDIM sampling steps.

The most computationally expensive operation is guided sampling, which involves depth map rendering at each sampling step and takes 35.9 s. Layout optimization requires 16.1 s and is performed twice during guided sampling.

Overall, the full pipeline takes approximately 1.2 min per scene with guided sampling enabled, and about 20 s without it.

Table 1. Inference runtime of different modules for DepR.

Module	**Runtime**
Segmentation	0.8 s
Depth Estimation	0.8 s
Latent Triplane Diffusion	1.2 s
VAE + SDF Decoding	1.0 s
Layout Optimization	16.1 s
Guided Sampling	35.9 s

2. Additional Implementation Details

2.1. SDF Depth Rendering

Depth-guided sampling requires rendering a depth map for the object being reconstructed. Following MonoSDF [5], we render the predicted SDF field into a depth map via differentiable volumetric rendering. SDF values s are first converted into density values using the following equation:

$$\sigma_\beta(s) = \begin{cases} \frac{1}{2\beta} \exp\left(\frac{s}{\beta}\right) & \text{if } s \leq 0 \\ \frac{1}{\beta}\left(1 - \frac{1}{2}\exp\left(-\frac{s}{\beta}\right)\right) & \text{if } s > 0 \end{cases} \tag{1}$$

where β is a hyper-parameter, set to 0.001 in our experiments.

To compute the depth $\hat{D}(\mathbf{r})$ of the surface intersecting the current ray $\mathbf{r}$, we sample M points of the form $\mathbf{o} + t_\mathbf{r}^i \mathbf{v}$, where $\mathbf{o}$ and $\mathbf{v}$ are the camera origin and viewing direction, respectively. The expected depth is computed as:

$$\hat{D}(\mathbf{r}) = \sum_{i=1}^{M} T_\mathbf{r}^i \alpha_\mathbf{r}^i t_\mathbf{r}^i \tag{2}$$

where the transmittance $T_\mathbf{r}^i$ is defined as:

$$T_\mathbf{r}^i = \prod_{j=1}^{i-1} (1 - \alpha_\mathbf{r}^j) \tag{3}$$

and the alpha value $\alpha_\mathbf{r}^i$ is given by:

$$\alpha_\mathbf{r}^i = 1 - \exp(-\sigma_\mathbf{r}^i \delta_\mathbf{r}^i) \tag{4}$$

where $\delta_\mathbf{r}^i$ is the distance between adjacent sample points. The rendered depth $\hat{D}$ is then used to compute the scale-invariant loss, which guides the DDIM sampling process for improved alignment with the actual depth.

2.2. Network Architecture

For the conditioning input, we construct the feature volume V by back-projecting image features into 3D space, followed by a 3D CNN and a linear projection to obtain a tensor of shape 9×32^3. After orthogonal projection onto the three planes (XY, YZ, XZ), we obtain 3-view feature maps $F_\text{proj} \in \mathbb{R}^{3 \times 9 \times 32^2}$.

We then concatenate the noised latent $z_t \in \mathbb{R}^{3 \times 2 \times 32^2}$, the 3-view features F_proj, and the attention-enhanced 2D feature $F_\text{att} \in \mathbb{R}^{1 \times 32^2}$ (zero-padded for XZ and YZ planes) along the feature dimension. The resulting tensor has shape $3 \times 12 \times 32^2$ and is passed to the diffusion U-Net.

The diffusion U-Net follows the architecture proposed in BlockFusion [4], which adapts a standard 2D U-Net to operate over tri-plane inputs. All 2D convolution layers are replaced with a specialized `GroupConv` operator, enabling parallel processing of the three planes. To allow information exchange across planes, the feature maps are flattened into 1D tokens and passed through six self-attention layers in the U-Net's middle block.

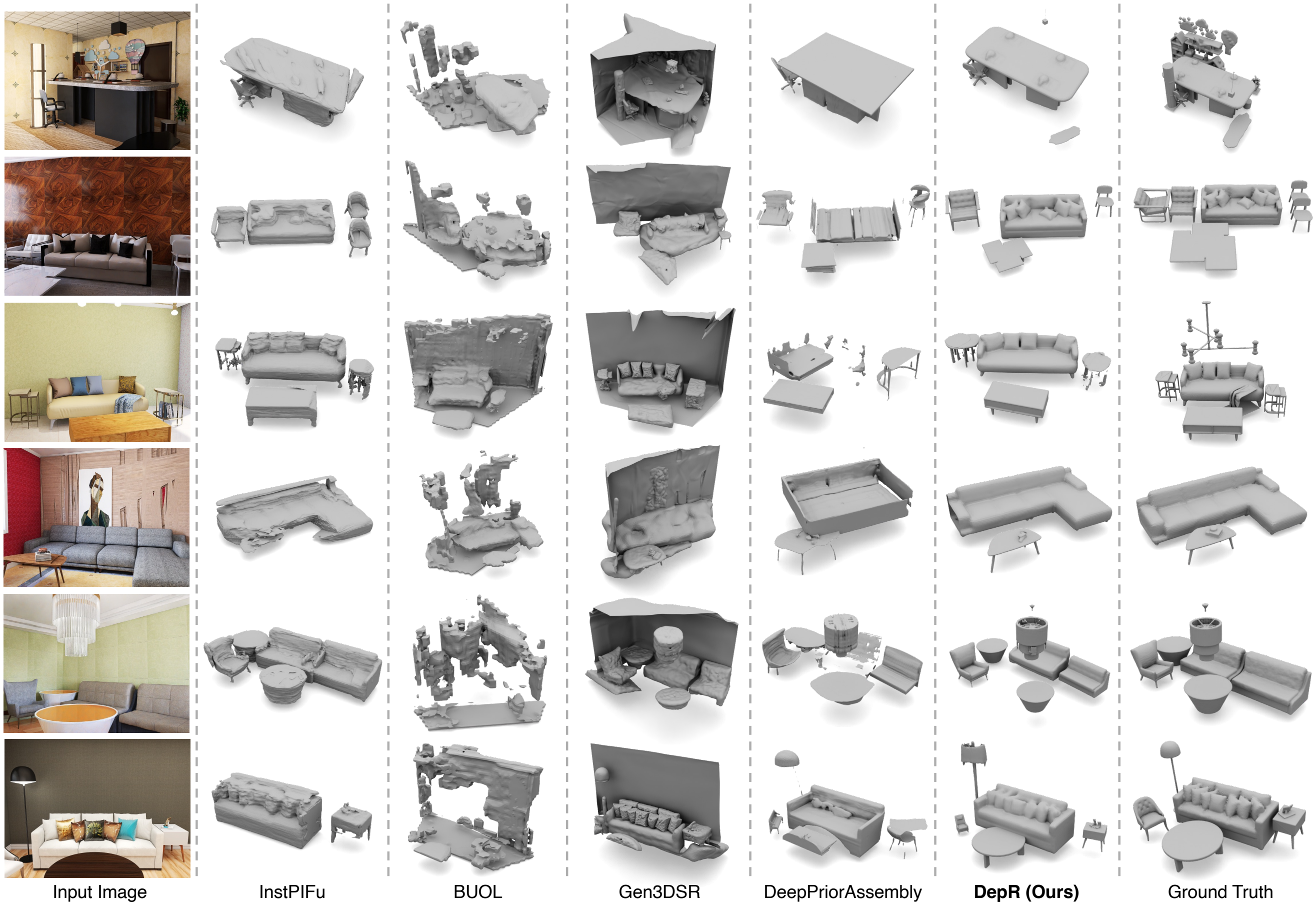

Figure 1. Additional qualitative results on the 3D-FRONT [2] dataset.

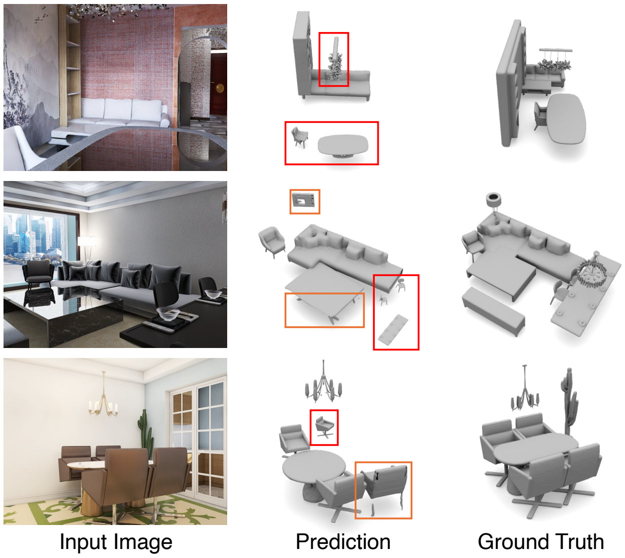

Figure 2. Failure cases of DepR. Red rectangles indicate incorrect layout estimation; orange rectangles highlight incorrect object reconstruction.

3. Additional Qualitative Results

Fig. 1 presents additional qualitative examples from the 3D-FRONT [2] dataset. Feed-forward methods such as InstPIFu [3] and BUOL [1] generally demonstrate limited generalizability. BUOL, which is trained on less realistic renderings from 3D-FRONT, fails to reconstruct meaningful geometry in most scenes, except for the scene in row 3. Among compositional methods, DepR consistently produces more visually coherent surfaces and superior overall geometry.

4. Limitations

Fig. 2 illustrates several failure cases of DepR. Due to its generative nature, DepR may incorrectly reconstruct or "hallucinate" objects when observations are severely limited by extreme occlusions (highlighted in orange rectangles). In the scene composition stage, the primary limitation arises from the optimization-based layout estimation (highlighted in red rectangles), which is susceptible to local minima. This issue becomes particularly pronounced when only a small portion of an object is visible, leading to highly incomplete depth point clouds that hinder accurate pose estimation.

Future work could address this limitation by integrating a learned, feed-forward pose regression module into the reconstruction framework. Such an approach may reduce the sensitivity to local minima and significantly improve overall inference efficiency.

References

[1] 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.

[2] 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.

[3] 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.

[4] 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.

[5] 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.