Single-view 3D Scene Reconstruction with High-fidelity Shape and Texture

Yixin Chen[1⋆], Junfeng Ni[2⋆†], Nan Jiang[3†], Yaowei Zhang[1], Yixin Zhu[3], Siyuan Huang[1]

⋆Equal contributors † Work done during an internship at BIGAI

[1] National Key Laboratory of General Artificial Intelligence, BIGAI [2] Tsinghua University [3] Peking University

https://dali-jack.github.io/SSR

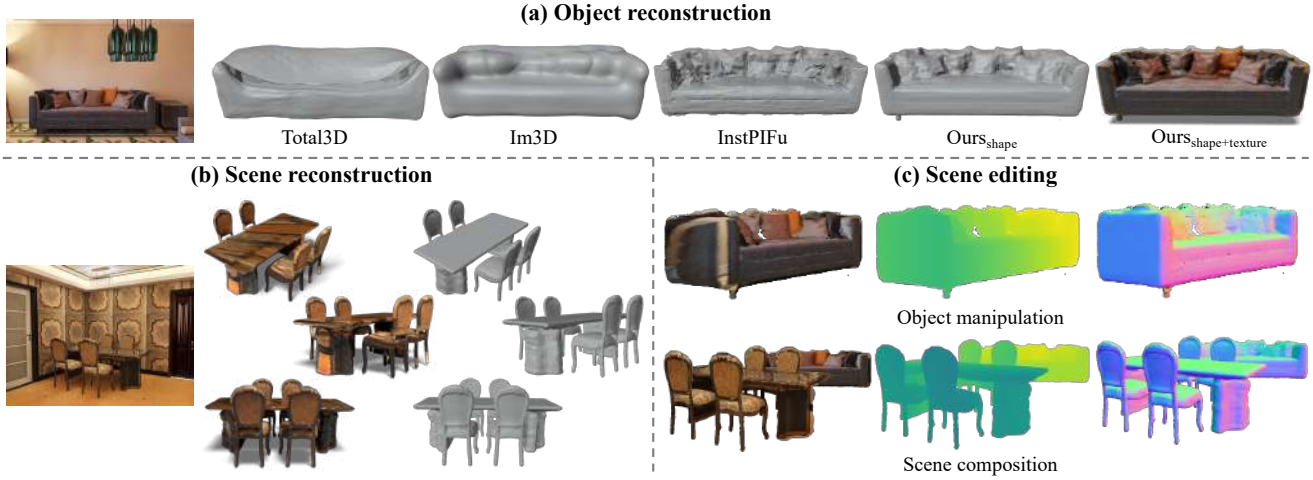

Figure 1. **Single-view 3D scene reconstruction with high-fidelity shape and texture.** (a) Object-level and (b) scene-level reconstruction. Rendering of **color**, **depth**, and **normal** images from the original and novel viewpoints enables 3D scene editing. (c) Object manipulation by rotating the object in (a) and scene composition of (a) and (b).

Abstract

Reconstructing detailed 3D scenes from single-view images remains a challenging task due to limitations in existing approaches, which primarily focus on geometric shape recovery, overlooking object appearances and fine shape details. To address these challenges, we propose a novel framework for simultaneous high-fidelity recovery of object shapes and textures from single-view images. Our approach utilizes the proposed Single-view neural implicit Shape and Radiance field (SSR) representations to leverage both explicit 3D shape supervision and volume rendering of color, depth, and surface normal images. To overcome shape-appearance ambiguity under partial observations, we introduce a two-stage learning curriculum incorporating both 3D and 2D supervisions. A distinctive feature of our framework is its ability to generate fine-grained textured meshes while seamlessly integrating rendering capabilities into the single-view 3D reconstruction model. This integration enables not only improved textured 3D object reconstruction by 27.7% *and* 11.6% *on the 3D-FRONT and Pix3D datasets, respectively, but also supports the rendering of images from novel viewpoints. Beyond individual objects, our approach facilitates composing object-level representations into flexible scene representations, thereby enabling applications such as holistic scene understanding and 3D scene editing. We conduct extensive experiments to demonstrate the effectiveness of our method.*

1. Introduction

Single-view 3D reconstruction is a challenging task in computer vision that aims to recover a scene's 3D geometry and appearance from a single monocular image. This task holds immense importance as it allows machines to understand and interact with the real 3D world, enabling various applications in virtual reality, augmented reality, and robotics.

The primary obstacle in single-view reconstruction lies in the inherent uncertainties and ambiguities resulting from the limited observations provided by a single image. A model must be able to infer the 3D object shape accurately based on visible regions while also generating a plausible representation of unseen object parts present in the image.

Over the years, various representations and methods have been proposed to tackle this challenge. Early methods in this field utilize 3D bounding boxes to parameterize 3D objects and estimate their size, rotation, and translation [9, 14, 16, 26, 29]. Recent advancements have focused on recovering detailed object shapes using either explicit [21, 52] or implicit [40, 85] representations. However, these approaches suffer from two notable drawbacks. First, they neglect the importance of object textures, which contain essential geometric and semantic details for embodied tasks [20, 28, 45] and 3D vision-language reasoning [2, 3, 8, 11, 86]. Second, they often rely solely on image inputs for feature extraction without taking direct textural supervision from them [52, 85]. Consequently, these models tend to focus insufficiently on geometric subtleties and may learn mean shapes for each object category, leading to challenges in generating smooth and instance-specific details, even when instance and pixel-aligned features are utilized for implicit representation learning [40].

To address the aforementioned limitations and improve single-view 3D reconstruction, we propose a novel framework that **simultaneously recovers shapes and textures** from single-view images. Our framework leverages the Single-view neural implicit Shape and Radiance field (SSR) representations. Conditioned on the input image, we extract pixel-aligned and instance-aligned features to predict the signed distance function (SDF) value using an implicit network and the color value using a rendering network for each query 3D point. By expressing volume density as a function of the SDF, our model can be trained end-to-end with **both 3D shape supervision and volume rendering** of color, depth, and surface normal images.

However, due to shape-appearance ambiguity, simply incorporating rendering supervision may lead to generating realistic textured images but inconsistent underlying geometry [12, 51], especially under partial observations. To tackle this issue and achieve improved coordination between 2D and 3D supervision, we propose a carefully designed two-stage learning curriculum. This curriculum balances the rendering and reconstruction losses, allowing our framework to learn a 3D object prior that reconstructs unseen parts from partial observations while capturing pixel-level fine-grained details from the images.

We extensively evaluate our proposed model for single-view object reconstruction on both synthetic 3D-FRONT dataset [19] and real Pix3D dataset [67]. The experimental results demonstrate that our method excels in recovering high-fidelity object shapes and textures, significantly outperforming state-of-the-art methods by **27.7%** and **11.6%** on 3D-FRONT and Pix3D, respectively. Through thorough ablation studies, we demonstrate the benefits of introducing textural supervision and the importance of the learning curriculum. We show that our model is capable of rendering images from novel viewpoints given single-view inputs, and the quality of the rendered depth and normal is comparable with existing depth and normal estimators [17, 84]. Finally, we showcase our model's capability in holistic scene understanding and 3D scene editing, allowing for object translation, rotation, and composition of objects in 3D space.

In summary, our work represents a significant advancement in single-view 3D reconstruction, and our contributions are three-fold:

1. We propose a novel framework that simultaneously recovers high-fidelity object shapes and textures from single-view images. Our framework leverages the strengths of neural implicit surfaces in shape learning and radiance fields in texture modeling, and seamlessly introduces rendering capabilities into a single-view 3D reconstruction model.
2. To effectively employ supervision from both 3D shapes and volume rendering, we conduct a thorough analysis and propose a carefully designed two-stage learning curriculum that improves 2D-3D supervision coordination and addresses shape-appearance ambiguity.
3. Extensive experiments and ablations demonstrate that our proposed method significantly enhances the details of textured 3D object reconstruction, outperforming all state-of-the-art methods. We demonstrate its ability to render color, depth, and normal images from novel viewpoints and its potential to facilitate applications such as holistic scene understanding and 3D scene editing.

2. Related work

3D reconstruction from a single Image Reconstructing 3D shapes from single images remains a challenging task in indoor scene understanding, and it has spurred the development of relevant datasets [13, 19, 66, 67] and models [14, 24, 27, 39]. Early approaches utilized 3D bounding boxes [9, 16, 26, 29, 31] or retrieved CAD models [30, 32, 50] to represent objects, but they lacked instance-specific 3D object geometries. Recent methods explored explicit [21, 52] or implicit [40, 85] representations to address these limitations. However, they overlooked object textures, a crucial aspect for semantic-demanding tasks that require pixel-level details. This is often addressed through generative approaches given the 3D shapes [5, 65]. In this paper, we propose a novel approach that simultaneously recovers detailed 3D geometry and object textures using neural implicit shape and radiance field representation.

Generative methods have also been proposed for single-view 3D reconstruction, utilizing priors learned from large-scale datasets. 2D prior-guided models [42, 47, 64, 69] generate images from novel views and reconstruct objects under a multi-view setting, while the 3D counterpart employs millions of 3D-text pairs to train a conditional generative model [33]. In comparison, our approach leverages benefits

from both 3D and 2D supervision in a discriminative way and demonstrates superior capture of high-fidelity 3D structures and details by explicitly modeling the object shapes and textures together.

Neural implicit surfaces representation Implicit representations model 3D geometry with neural networks in a parametrized manner [18, 41, 54, 57]. Unlike explicit representations (such as point cloud [1, 61], mesh [21, 59], voxels [34, 75]), implicit representations are continuous, high spatial resolution, and have constant memory usage. However, most existing work [10, 48, 52, 55, 57] conditions neural implicit representation on global image features, which improves memory efficiency but compromises on preserving details, leading to retrieval-like results. Even when instance and pixel-aligned features are utilized for implicit representation learning from a single view [40, 63, 77], the model may fail to capture higher-order relationships between 3D points and lack incentives to capture geometric details reflecting pixel-level image details. In this paper, we address this limitation by employing neural implicit shape and radiance field representation, which benefits from both 3D shape supervision and volume rendering, allowing the model to learn geometric and appearance details jointly.

Surface representation learning Recent advances in implicit volume rendering (*e.g.*, neural radiance fields (NeRF) [35, 46, 70]) have spurred new efforts in surface representation learning. However, extracting high-fidelity surfaces from learned radiance fields is challenging due to insufficient constraints on the level sets in density-based scene representation. To overcome this limitation, recent methods have combined the benefits of implicit surface and volume rendering-based methods by converting the SDF to density and applying volume rendering to train this representation with robustness [56, 74, 80]. Nevertheless, rendering-based approaches often yield unsatisfactory results in 3D geometry recovery, especially when provided with sparse input views, such as in cluttered indoor scenes. Such failure is rooted in the shape-appearance ambiguity with photo-realistic losses, where an infinite number of photo-consistent explanations exist for the same input image. In this work, we propose an approach that leverages both 3D shape and volume rendering supervision for single-view reconstruction. Moreover, we make the first attempt to investigate how to coordinate these two sources of supervision effectively. To this end, we introduce a two-stage learning curriculum with an incremental increase for the rendering loss weights to achieve improved coordination between the 3D and 2D supervisions and better capture geometric details for textured 3D object reconstruction.

3. Method

Given a single image of an indoor scene, our objective is to simultaneously reconstruct the 3D geometry and appearance of all objects present. We build upon existing methods [40, 52, 85] for 3D object detection and camera pose estimation, focusing on the reconstruction of fine-grained textured meshes.

3.1. Background

Neural implicit surfaces with SDF We utilize neural implicit surfaces with SDF to represent 3D geometry. The SDF provides a continuous function that yields the distance s to the closest surface for a given point $\mathbf{x}$, with the sign indicating whether the point lies inside (negative) or outside (positive) the surface:

$$\text{SDF}(\mathbf{x}) \;=\; s : \mathbf{x} \in \mathbb{R}^3,\; s \in \mathbb{R}. \tag{1}$$

The zero-level set of the SDF function $\Omega = \{\mathbf{x} \in \mathbb{R}^3 \mid \text{SDF}(\mathbf{x}) = 0\}$ implicitly represents the surface.

Volume rendering of implicit surfaces To enable optimization with differentiable volume rendering, we convert the neural implicit surface representation SDF to density σ [56, 74, 80]. The conversion is performed using a learnable parameter β as follows:

$$\sigma_\beta(s) = \begin{cases} \frac{1}{2\beta}\exp(\frac{s}{\beta}) & s \leqslant 0 \\ \frac{1}{\beta}(1-\exp(-\frac{s}{\beta})) & s \geqslant 0 \end{cases}. \tag{2}$$

Following the concept of NeRF [49], we sample M points on the ray $\mathbf{r}$ from the camera center $\mathbf{o}$ to the pixel along the viewing direction $\mathbf{d}$:

$$\mathbf{x}_\mathbf{r}^i = \mathbf{o} + t_\mathbf{r}^i \mathbf{d}, \quad i = 1, \ldots, M, \tag{3}$$

where $t_\mathbf{r}^i$ is the distance from the sample point to the camera center. We predict the SDF value $\hat{s}$ and color value $\hat{\mathbf{c}}_\mathbf{r}^i$ for each sample point on the ray.

The predicted color $\hat{C}(\mathbf{r})$ for the ray $\mathbf{r}$ can be computed using transmittance $T_\mathbf{r}^i$ and alpha values $\alpha_\mathbf{r}^i$:

$$\hat{C}(\mathbf{r}) = \sum_{i=1}^{M} T_\mathbf{r}^i \alpha_\mathbf{r}^i \hat{\mathbf{c}}_\mathbf{r}^i, \tag{4}$$

where $T_\mathbf{r}^i = \prod_{j=1}^{i-1}(1-\alpha_\mathbf{r}^j)$. The alpha value is calculated as $\alpha_\mathbf{r}^i = 1 - \exp(-\sigma_\mathbf{r}^i \delta_\mathbf{r}^i)$, and $\delta_\mathbf{r}^i$ represents the distance between neighboring sample points along the ray. Additionally, we can compute the depth $\hat{D}(\mathbf{r})$ and normal $\hat{N}(\mathbf{r})$ of the surface intersecting the current ray $\mathbf{r}$ as:

$$\hat{D}(\mathbf{r}) = \sum_{i=1}^{M} T_\mathbf{r}^i \alpha_\mathbf{r}^i t_\mathbf{r}^i, \quad \hat{N}(\mathbf{r}) = \sum_{i=1}^{M} T_\mathbf{r}^i \alpha_\mathbf{r}^i \mathbf{n}_\mathbf{r}^i. \tag{5}$$

3.2. 3D object reconstruction

Given the input image I of the scene, we aim to recover the 3D shape of the object $\mathcal{O}$, identified by its 2D bounding box, 3D bounding box, and category class.

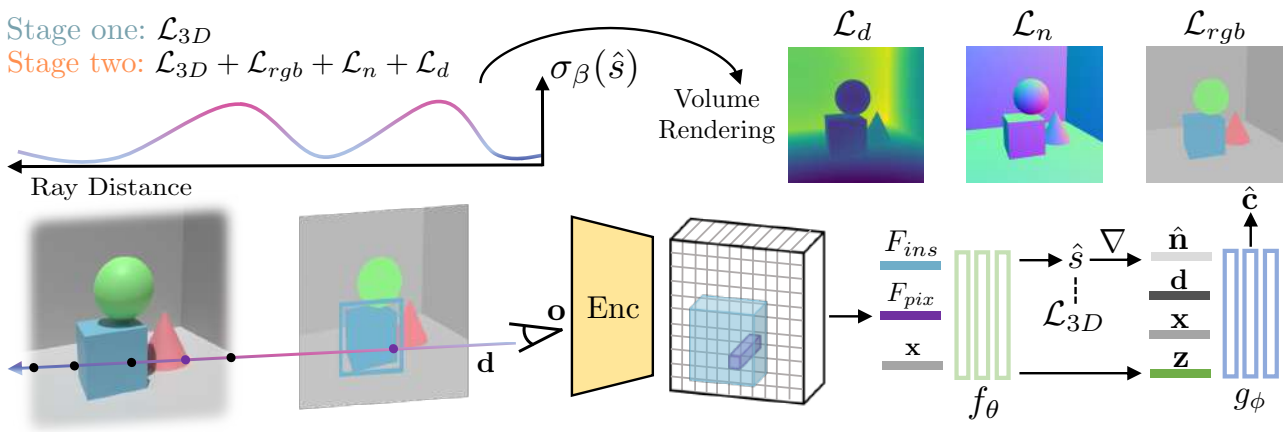

Figure 2. **Framework overview.** Our framework jointly recovers 3D object shapes and textures from single-view images. Given a query point $\mathbf{x}$ along a camera ray with direction $\mathbf{d}$, we extract pixel-aligned and instance-aligned features using an image encoder Enc. The implicit network f_θ predicts the geometry feature $\hat{\mathbf{z}}$ and SDF value $\hat{s}$, which is then transformed to volume density σ. The rendering network g_θ takes the normal $\hat{\mathbf{n}}$ and viewing direction $\mathbf{d}$ to predict the color value $\hat{\mathbf{c}}$. Our learning curriculum consists of two stages: Stage One, which only employs explicit SDF supervision, and Stage Two, where volume rendering supervision is incrementally added.

Feature extraction We extract image features $F = \text{Enc}(I)$ using a CNN-based encoder Enc and utilize both instance-aligned feature F_{ins} and pixel-aligned feature F_{pix} for recovering detailed shapes and textures. $F_{ins}(\mathcal{O})$ is obtained by cropping out the region-of-interest (ROI) features from F based on the 2D bounding box of object $\mathcal{O}$ following He *et al*. [25] and Liu *et al*. [40]. To obtain the pixel-aligned feature $F_{pix}(\mathbf{x})$ for a 3D point $\mathbf{x}$, we project $\mathbf{x}$ onto the image plane to obtain the corresponding image coordinates $\pi(\mathbf{x})$ using the camera intrinsics. The pixel-aligned feature is then obtained through linear interpolation on the feature map, *i.e*., $F_{pix}(\mathbf{x}) = \text{Interp}(F(\pi(\mathbf{x})))$.

Implicit and rendering networks We parameterize the SDF function with an implicit network f_θ, which is a single MLP [57, 81] taking the instance-aligned feature, pixel-aligned feature, and the position $\mathbf{x}$ as input to predict the SDF value $\hat{s}$:

$$\hat{s} = f_\theta(\gamma(\mathbf{x}), F_{ins}(\mathcal{O}), F_{pix}(\mathbf{x})). \tag{6}$$

Here, $\gamma(\cdot)$ represents a positional encoding with 6 exponentially increasing frequencies. The rendering network g_ϕ predicts RGB color values $\hat{\mathbf{c}}$ for each 3D point using the 3D point $\mathbf{x}$, normal $\hat{\mathbf{n}}$, viewing direction $\mathbf{d}$, and a global geometry feature $\hat{\mathbf{z}}$ as input, following Yariv *et al*. [79]:

$$\hat{\mathbf{c}} = g_\phi(\mathbf{x}, \mathbf{d}, \hat{\mathbf{n}}, \hat{\mathbf{z}}). \tag{7}$$

The 3D normal $\hat{\mathbf{n}}$ is calculated as the analytical gradient of the SDF function, *i.e*., $\hat{\mathbf{n}} = \nabla f_\theta(\cdot)$. The feature vector $\hat{\mathbf{z}}$ is the output of a second linear head of the implicit network, as in Yariv *et al*. [79] and Yu *et al*. [82].

3.3. Supervision and learning curriculum

We employ neural implicit shape and radiance field representation to effectively learn the 3D shape prior and to capture pixel-level details, benefiting from both explicit 3D and volume rendering supervision.

3D supervision We apply direct 3D supervision by using the following loss between the predicted and real SDF values:

$$\mathcal{L}_{3D} = \sum_{\mathbf{x} \in \{\mathcal{X} \bigcup \mathbf{r}\}} ||s(\mathbf{x}) - \hat{s}(\mathbf{x})||_1. \tag{8}$$

This loss is computed for points along the rays $\mathbf{r}$ and from the point set $\mathcal{X}$, which contains uniformly sampled points and near-surface points.

Photometric reconstruction loss For all rays $\mathbf{r}$ in the minibatch, we render each pixel with the predicted SDF values $\hat{s}$ and color value $\hat{\mathbf{c}}$ for all sampled points on the ray; the volume rendering formulations are detailed in Sec. 3.1. The photometric reconstruction loss is defined as:

$$\mathcal{L}_{rgb} = \sum_{\mathbf{r}} ||C(\mathbf{r}) - \hat{C}(\mathbf{r})||_1, \tag{9}$$

where $C(\mathbf{r})$ denotes the color value in the input image.

Exploiting monocular geometric cues To further alleviate ambiguities in recovering 3D shapes from single-view inputs, we follow Yu *et al*. [82] and exploit monocular depth and normal cues to facilitate the training process. The depth and normal consistency losses are defined as follows:

$$\begin{aligned} \mathcal{L}_d &= \sum_{\mathbf{r}} ||D(\mathbf{r}) - \hat{D}(\mathbf{r})||^2 \\ \mathcal{L}_n &= \sum_{\mathbf{r}} ||N(\mathbf{r}) - \hat{N}(\mathbf{r})||_1 + ||1 - N(\mathbf{r})^\top \hat{N}(\mathbf{r})||_1 \end{aligned} \tag{10}$$

Compared to the photometric reconstruction loss, the depth and normal consistency losses can directly supervise the SDF prediction in the implicit network f_θ through back-propagation without the rendering network g_ϕ; please refer to Fig. 2 for detailed illustration.

Overall loss The overall loss used to optimize the implicit and rendering networks jointly is given by:

$$\mathcal{L} = \alpha_{3D}\mathcal{L}_{3D} + \alpha_{rgb}\mathcal{L}_{rgb} + \alpha_d\mathcal{L}_d + \alpha_n\mathcal{L}_n, \quad (11)$$

where α denotes the respective loss weight. $\mathcal{L}_{rgb}$, $\mathcal{L}_d$, and $\mathcal{L}_n$ are applied to the visible pixels for the object $\mathcal{O}$, indicated by its visible mask segmentation. Note that depth, normal, and segmentation are only used during the training stage, and none are required during the inference stage, preserving the flexibility and applicability of our model.

Learning curriculum To address the limitations of naively incorporating 3D and rendering supervision in the single-view setting, which may result in realistic images but inconsistent 3D geometry due to shape-appearance ambiguity, we introduce a learning curriculum based on two empirical observations: 1) the rendering supervision should serve as an auxiliary to 3D supervision, and 2) it is more effective to first learn the overall object shape before delving into finer details. Following these, our learning curriculum comprises two stages: Stage One, which only employs 3D supervision $\mathcal{L}_{3D}$, and Stage Two, which incorporates $\mathcal{L}_{rgb}$, $\mathcal{L}_d$, and $\mathcal{L}_\mathbf{n}$ with linearly increasing loss weights:

$$\alpha = \eta(\lambda - \lambda_0), \lambda > \lambda_0. \quad (12)$$

λ denotes the epoch number, λ_0 is the starting epoch of Stage Two, and η is the linear coefficient. λ_0 is empirically selected by observing the shape learning curves and our curriculum is crucial for performance improvement (Sec. 4.1).

3.4. 3D scene composition

A scene can be represented by the composition of objects $\{\mathcal{O}_i, i = 1, \cdots, k\}$ within it. We obtain both the 3D reconstructed geometry and the photometric rendering of the scene by composing the implicit representations of the individual objects given their 2D and 3D bounding boxes.

3D scene geometry To compose the 3D geometry of the scene, we transform each object's implicit surfaces into explicit meshes using the marching cube algorithm [44]. The object meshes are then combined using the camera's extrinsic parameters and 3D object bounding boxes.

3D scene rendering To render an image of the scene, we sample points along the rays and estimate their density and color values for each individual object. Sampled points on the same ray from different objects are then grouped together to composite the colors and densities for volume rendering. This distance-aware integration ensures that only visible objects appear in the final images, as the accumulated transmittance along the ray reflects visibility.

The object composition operation offers flexibility for both reconstruction and rendering, making it applicable for holistic scene understanding and novel view synthesis with 3D scene editing, such as object rotation, translation, and compositions from different scenes.

4. Experiment

We evaluate single-view object reconstruction in indoor scenes using synthetic dataset 3D-FRONT [19] and real dataset Pix3D [67]. Our model's capabilities are tested in novel view synthesis, depth estimation, and normal estimation tasks, leveraging its rendering capabilities. Additionally, we showcase potential applications of our model, including holistic scene understanding and 3D scene editing.

4.1. Indoor object reconstruction

Datasets We evaluate our single-view object reconstruction on synthetic dataset 3D-FRONT [19] and real dataset Pix3D [67]. We adopt the same splits as Liu *et al*. [40] for both datasets. Data preparation details, including monocular cues and SDF generation, are in **Sup. Mat.**.

Evaluation metrics To evaluate 3D object reconstruction, we use Chamfer Distance (CD), F-Score, and Normal Consistency (NC) following Wang *et al*. [73] and Mescheder *et al*. [48]. CD measures the sum of squared distances between the nearest neighbor correspondences of two point clouds after mesh alignment. F-Score [37] is the harmonic mean of precision and recall of points in the prediction and ground truth within the nearest neighbor. NC quantifies how well methods capture higher-order information by computing the mean absolute dot product of the normals between meshes after alignment.

Results In the single-view object reconstruction task, we compare our method with MGN of Total3D [52], the LIEN of Im3D [85], and InstPIFu [40]. Our model surpasses state-of-the-art methods across all three metrics in both synthetic (Tab. 1) and real (Tab. 2) datasets. Notably, on 3D-FRONT, our model achieves 27.7% and 16.4% performance gain in mean CD and F-Score, respectively, as well as the best NC on all object categories. This highlights our model's proficiency in predicting highly detailed object shapes and smoother surfaces (Fig. 3). Moreover, our model predicts high-fidelity textures, a capability lacking in previous models. It represents significant progress in single-view 3D object reconstruction, enabling the recovery of both fine-grained shapes and intricate textures. Further details, results, and failure cases are in **Sup. Mat.**.

Ablations To analyze the effects of various supervisions and learning curricula, we conduct ablative studies on SDF, color (C), depth (D), and normal (N) supervisions, along with different loss weights and curriculum strategies ($Curr.$). Key findings from Tab. 3 and Fig. 4 are as follows:

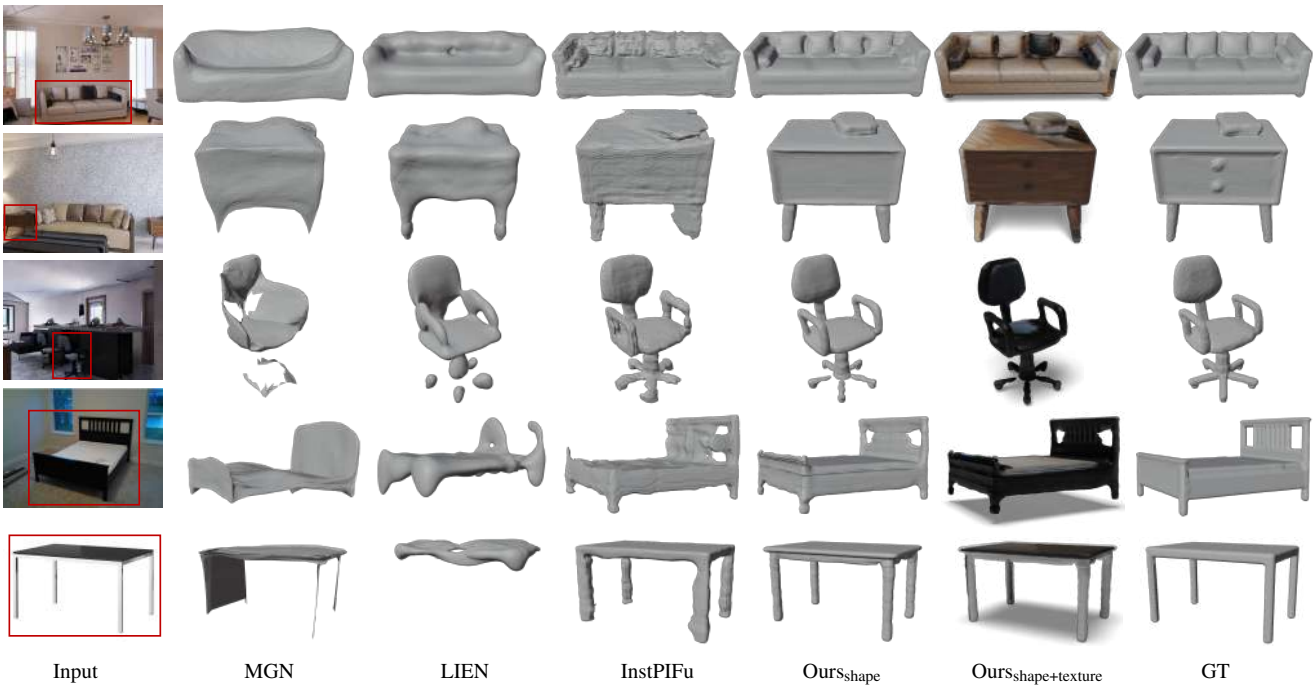

Figure 3. **Qualitative results of indoor object reconstruction.** Examples from 3D-FRONT [19] (top three rows) and Pix3D [67] (bottom two rows) datasets. Our model produces textured 3D objects with smoother surfaces and finer details than previous methods.

1. Incorporating color, depth, and normal supervision in our framework significantly enhances 3D object reconstruction, especially in capturing finer details.
2. 2D supervision should act as an auxiliary to 3D supervision, as simply increasing 2D loss weights (*e.g.*, $\checkmark_{\times 5}$ or $\checkmark_{\times 10}$) negatively impacts 3D reconstruction. Fig. 4(c) shows clear artifacts along the ray directions, indicating the importance of a suitable learning curriculum.
3. Our proposed curriculum, gradually increasing 2D loss weights after the 3D shape prior learning phase (Stage Two starting epoch $\lambda_0 = 150$), outperforms early injection of 2D supervision ($\lambda_0 = 0$ or $\lambda_0 = 70$).

Table 1. **Object reconstruction on the 3D-FRONT [19] dataset.** Our model achieves the best performance on mean CD and F-Score, as well as the best NC on all object categories, outperforming MGN [52], LIEN [85], and InstPIFu [40]. †: Results reproduced from the official repository.

Category		bed	chair	sofa	table	desk	nightstand	cabinet	bookshelf	mean
CD ↓	MGN	15.48	11.67	8.72	20.90	17.59	17.11	13.13	10.21	14.07
	LIEN	16.81	41.40	9.51	35.65	26.63	16.78	7.44	11.70	28.52
	InstPIFu	18.17	14.06	7.66	23.25	33.33	**11.73**	**6.04**	8.03	14.46
	Ours	**4.96**	**10.52**	**4.53**	**16.12**	**25.86**	17.90	6.79	**3.89**	**10.45**
F-Score ↑	MGN	46.81	57.49	64.61	49.80	46.82	47.91	54.18	54.55	55.64
	LIEN	44.28	31.61	61.40	43.22	37.04	50.76	69.21	55.33	45.63
	InstPIFu	47.85	59.08	67.60	56.43	**48.49**	57.14	**73.32**	66.13	61.32
	Ours	**76.34**	**69.17**	**80.06**	**67.29**	47.12	**58.48**	70.45	**85.93**	**71.36**
NC ↑	MGN†	0.829	0.758	0.819	0.785	0.711	0.833	0.802	0.719	0.787
	LIEN†	0.822	0.793	0.803	0.755	0.701	0.814	0.801	0.747	0.786
	InstPIFu†	0.799	0.782	0.846	0.804	0.708	0.844	0.841	0.790	0.810
	Ours	**0.896**	**0.833**	**0.894**	**0.838**	**0.764**	**0.897**	**0.856**	**0.862**	**0.854**

Comparison with prior-guided models We compare our model with generative models demonstrating potential zero-shot generalizability by leveraging 2D or 3D geometric priors learned from *large-scale* datasets. Specifically, we choose two representative works: (1) Zero-1-to-3 [42], which uses Objaverse [15] to learn a 2D diffusion prior for novel view synthesis under specified camera transformation and reconstructs objects under a multi-view setting; (2) Shap·E [33], which directly generates textured meshes given images and category prompts, trained on millions of paired 3D and text data. For a fair evaluation, we compare them with our model on a subset of the test split in 3D-FRONT with ground truth object scale. Results in Fig. 5 and Tab. 4 show that while Zero-1-to-3 produces reasonable

Table 2. **Object Reconstruction on the Pix3D [67] dataset.** On the non-overlapped split [40], our model outperforms the state-of-the-art methods by significant margins. †: Results reproduced from the official repository.

Category		bed	bookcase	chair	desk	sofa	table	tool	wardrobe	misc	mean
CD ↓	MGN	22.91	33.61	56.47	33.95	9.27	81.19	94.70	10.43	137.50	44.32
	LIEN	11.18	29.61	40.01	65.36	10.54	146.13	29.63	4.88	144.06	51.31
	InstPIFu	10.90	7.55	32.44	**22.09**	8.13	45.82	10.29	**1.29**	47.31	24.65
	Ours	**6.31**	**7.21**	**26.23**	28.63	**5.68**	**43.87**	**8.29**	2.07	**35.03**	**21.79**
F-Score ↑	MGN	34.69	28.42	35.67	65.36	51.15	17.05	57.16	52.04	10.41	36.20
	LIEN	37.13	15.51	25.70	26.01	49.71	21.16	5.85	59.46	11.04	31.45
	InstPIFu	54.99	62.26	35.30	**47.30**	56.54	37.51	64.24	**94.62**	27.03	45.62
	Ours	**68.78**	**66.69**	**55.18**	42.49	**71.22**	**51.93**	**65.38**	91.84	**46.92**	**59.71**
NC ↑	MGN†	0.737	0.592	0.525	0.633	0.756	0.794	0.531	0.809	0.563	0.659
	LIEN†	0.706	0.514	0.591	0.581	0.775	0.619	0.506	0.844	0.481	0.646
	InstPIFu†	0.782	0.646	0.547	0.758	0.753	0.796	0.639	0.951	0.580	0.683
	Ours	**0.825**	**0.689**	**0.693**	**0.776**	**0.866**	**0.835**	**0.645**	**0.960**	**0.599**	**0.778**

Table 3. **Ablation studies on object reconstruction.** We demonstrate the benefits of introducing 2D supervision and employing a properly designed curriculum. The notation $\times 5/10$ indicates increased loss weights.

SDF	C	D	N	*Curr.*	CD ↓	F-Score ↑	NC ↑
✓	×	×	×	×	16.43	64.15	0.806
✓	✓	×	×	×	15.90	65.55	0.813
✓	✓	✓	×	×	14.02	67.31	0.828
✓	✓	✓	✓	×	12.92	67.71	0.841
✓	$\checkmark_{\times 5}$	$\checkmark_{\times 5}$	$\checkmark_{\times 5}$	×	16.19	64.92	0.813
✓	$\checkmark_{\times 10}$	$\checkmark_{\times 10}$	$\checkmark_{\times 10}$	×	19.22	58.26	0.771
✓	✓	✓	✓	$\lambda_0 = 0$	12.88	68.19	0.840
✓	✓	✓	✓	$\lambda_0 = 70$	12.42	68.87	0.845
✓	✓	✓	✓	$\lambda_0 = 150$	**10.45**	**71.36**	**0.854**

Figure 4. **Visual comparisons for ablation study.** (a) SDF only (b) SDF$+C+D+N$ (c) SDF$+C+D+N$ with $\times 10$ loss weights $\checkmark_{\times 10}$ (d) $\lambda_0 = 0$ (e) $\lambda_0 = 150$. Incorporating 2D supervision with our designed curriculum yields the best reconstruction quality.

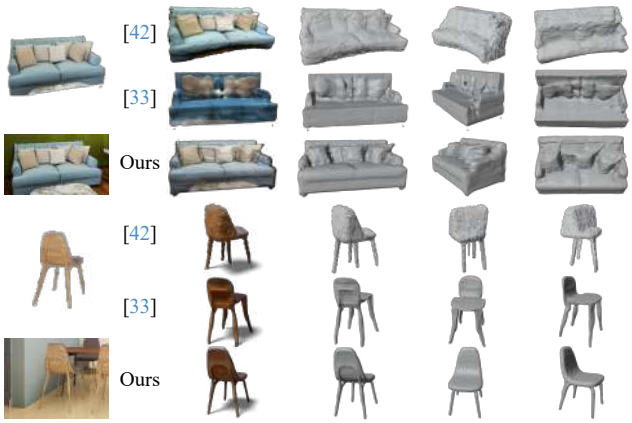

Figure 5. **Comparison with prior-guided models.** Inputs for Zero-1-to-3 [42] and Shap·E [33] only contains foreground objects. Each example is presented with textured mesh and mesh from three views. Our method outperforms prior-guided models in capturing details and 3D shape consistency.

images on particular views, it faces difficulties achieving overall 3D shape consistency. Shap·E captures the rough object shape but lacks detailed modeling of geometry and texture. On the contrary, our model excels at recovering the general 3D shapes while maintaining fine geometrical and

Table 4. **Quantitative comparison with prior-guided models.** Despite the zero-shot generalization ability, methods leveraging 2D or 3D priors fall short in recovering object geometry, especially surface details, compared to our proposed method.

	CD ↓	F-Score ↑	NC ↑
Zero-1-to-3 [42]	39.27	30.07	0.624
Shap·E [33]	29.16	39.86	0.686
Ours	**10.86**	**69.95**	**0.846**

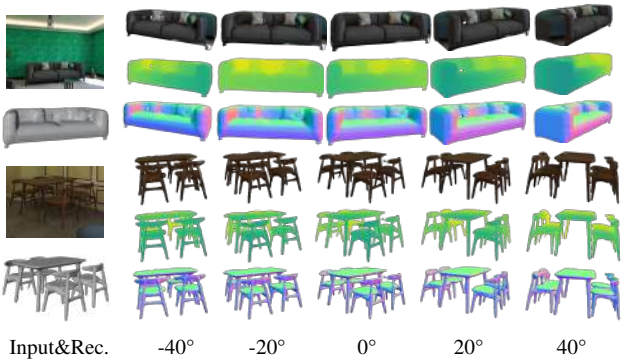

Figure 6. **Novel view rendering from single-view inputs.** Our model can render color, depth, and normal images for both objects (top) and scenes (bottom) from novel views.

textural details. This stresses the significant potential of effectively integrating 2D and 3D priors for future single-view reconstruction models to achieve enhanced results and generalizability. More details can be found in **Sup. Mat.**.

4.2. Rendering capability

Harnessing the advantages of our method, we seamlessly introduce rendering capabilities to a single-view reconstruction model. From the single-view input image, we can render the color, depth, and normal images through volume rendering, even from novel views. The qualitative examples presented in Fig. 6 illustrate that our method excels in producing plausible and consistent rendering results, even when the viewing angles change significantly (*i.e.*, $\pm 40°$).

Novel view synthesis PixelNeRF [81] employs a NeRF representation for novel view synthesis from input images. We compare the class-agnostic model of PixelNeRF, which is pre-trained on ShapeNet [7] and fine-tuned on 3D-FRONT. Qualitative results in Fig. 7 reveal a notable difference between the two approaches: PixelNeRF struggles to render images outside the vicinity of the original viewpoints, whereas our method is capable of generating meaningful renderings from novel viewpoints. This shows the importance of effectively imposing explicit 3D shapes in the scene reconstruction model, particularly when dealing with partial observation in real scenes.

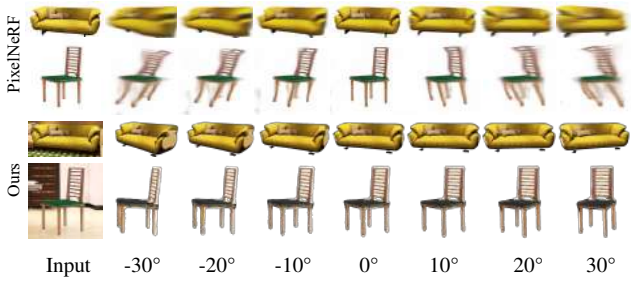

Figure 7. **Visual comparison for novel view rendering.** PixelNeRF [81] (top two rows) struggles to render images outside the vicinity of the input view, whereas our method (bottom two rows) can produce realistic renderings even for views far from the original input. PixelNeRF inputs only contain foreground objects.

Table 5. **Single-view depth and normal estimation.** We evaluate depth using L1 ↓ and normal using L1↓ / Angular°↓ error as metrics. For novel views, we use $\pm 15°$ views to evaluate the accuracy.

	Original View		Novel View	
	Depth	Normal	Depth	Normal
XTC [84]	1.188	12.712 / 14.309	-	-
Omnidata [17]	**0.734**	**10.015 / 11.257**	-	-
Ours	0.992	10.962 / 12.392	**1.179**	**12.094 / 13.710**

Depth and normal estimation Moreover, our model can serve as a proficient single-view depth and normal estimator. To validate this, we compare with zero-shot state-of-the-art methods [17, 84] on 3D-FRONT, following Ranftl *et al*. [62]. Results in Tab. 5 demonstrate that our model performs comparably on the input views. It also shows our model can directly estimate reasonable depth and normal maps on novel views. This is challenging since our model solely relies on single-view inputs, which is in stark contrast from previous work [23, 72, 82] that require multi-view inputs; see **Sup. Mat.** for additional results.

4.3. Applications

Generalizable holistic scene understanding Our method is capable of recovering 3D scene geometry and rendering corresponding color, depth, and normal images by composing object-level implicit representations (see Sec. 3.4 for more details). Fig. 8 showcases qualitative scene reconstruction results on SUNRGB-D [66] by employing existing 3D object detectors [6, 85]. The results demonstrate that *our method can reconstruct detailed object shapes and intricate textures in real images with cross-domain generalization ability*.

Scene editing Finally, we demonstrate our model's potential in representing scenes and enabling 3D scene editing applications. Our method allows for object-level editing, such as object translation, rotation, duplication, and composition of objects from different scenes into a shared 3D space. Qualitative results are shown in Fig. 9. Notably, *our approach can generate both 3D geometry and rendered images for edited scenes*, which differentiates itself from previous work that could only render images of manipulated objects [53, 76, 78], perform color or texture editing [38], or require multi-view posed images as input [41, 83].

Figure 8. **Holistic scene understanding and generalization.** The reconstruction results on SUNRGB-D [66] with existing 3D object detectors demonstrate our model's performance in recovering realistic scenes with generalization ability.

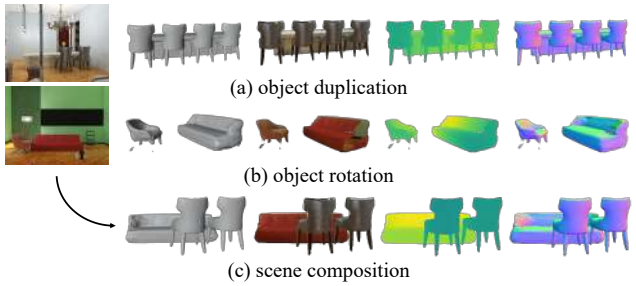

Figure 9. **3D scene editing based on single-view inputs.** Reconstructions and renderings are shown for (a) duplicating the chairs, (b) rotating the sofa and (c) scene composition of (a) and (b).

5. Conclusion

We present a novel framework for single-view scene reconstruction utilizing neural implicit shape and radiance field representations. Our model exhibits a significant advantage in textured 3D object reconstruction compared to state-of-the-art methods, and integrating color, depth, and normal supervision with our designed curriculum is pivotal to achieving improved performance. Furthermore, our model demonstrates impressive rendering capabilities and performs well in single-view depth and normal estimation, showing promise for generalization in holistic scene understanding and facilitating applications like 3D scene editing. Potential limitations include the model's ability to reconstruct objects from novel categories and texture recovery for unseen object parts. Effectively incorporating 2D or 3D priors from large-scale datasets offers a promising avenue for future direction.

Acknowledgment The authors thank colleagues from BIGAI for fruitful discussions and anonymous reviewers for constructive feedback. This work is supported in part by the National Key R&D Program of China (2021ZD0150200).

References

[1] Panos Achlioptas, Olga Diamanti, Ioannis Mitliagkas, and Leonidas Guibas. Learning representations and generative models for 3d point clouds. In *International Conference on Machine Learning (ICML)*, 2018. 3

[2] Panos Achlioptas, Ahmed Abdelreheem, Fei Xia, Mohamed Elhoseiny, and Leonidas Guibas. Referit3d: Neural listeners for fine-grained 3d object identification in real-world scenes. In *European Conference on Computer Vision (ECCV)*, 2020. 2

[3] Daichi Azuma, Taiki Miyanishi, Shuhei Kurita, and Motoaki Kawanabe. Scanqa: 3d question answering for spatial scene understanding. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[4] Paul J Besl and Neil D McKay. Method for registration of 3-d shapes. In *Sensor fusion IV: Control paradigms and data structures*, 1992. 2

[5] Alexey Bokhovkin, Shubham Tulsiani, and Angela Dai. Mesh2tex: Generating mesh textures from image queries. In *International Conference on Computer Vision (ICCV)*, 2023. 2

[6] Garrick Brazil, Abhinav Kumar, Julian Straub, Nikhila Ravi, Justin Johnson, and Georgia Gkioxari. Omni3d: A large benchmark and model for 3d object detection in the wild. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 8

[7] Angel X. Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, Jianxiong Xiao, Li Yi, and Fisher Yu. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015. 7, 3

[8] Dave Zhenyu Chen, Angel X Chang, and Matthias Nießner. Scanrefer: 3d object localization in rgb-d scans using natural language. In *European Conference on Computer Vision (ECCV)*, 2020. 2

[9] Yixin Chen, Siyuan Huang, Tao Yuan, Siyuan Qi, Yixin Zhu, and Song-Chun Zhu. Holistic++ scene understanding: Single-view 3d holistic scene parsing and human pose estimation with human-object interaction and physical commonsense. In *International Conference on Computer Vision (ICCV)*, 2019. 2

[10] Zhiqin Chen and Hao Zhang. Learning implicit fields for generative shape modeling. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 3

[11] Zhenyu Chen, Ali Gholami, Matthias Nießner, and Angel X Chang. Scan2cap: Context-aware dense captioning in rgb-d scans. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2

[12] Ismael Colomina and Pere Molina. Unmanned aerial systems for photogrammetry and remote sensing: A review. *ISPRS Journal of photogrammetry and remote sensing*, 2014. 2

[13] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 2

[14] Saumitro Dasgupta, Kuan Fang, Kevin Chen, and Silvio Savarese. Delay: Robust spatial layout estimation for cluttered indoor scenes. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 2

[15] 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 *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 6

[16] Yilun Du, Zhijian Liu, Hector Basevi, Ales Leonardis, Bill Freeman, Josh Tenenbaum, and Jiajun Wu. Learning to exploit stability for 3d scene parsing. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018. 2

[17] 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 *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2, 8, 1, 4

[18] SM Ali Eslami, Danilo Jimenez Rezende, Frederic Besse, Fabio Viola, Ari S Morcos, Marta Garnelo, Avraham Ruderman, Andrei A Rusu, Ivo Danihelka, Karol Gregor, et al. Neural scene representation and rendering. *Science*, 360 (6394):1204–1210, 2018. 3

[19] Huan Fu, Bowen Cai, Lin Gao, Ling-Xiao Zhang, Jiaming Wang, Cao Li, Qixun Zeng, Chengyue Sun, Rongfei Jia, Binqiang Zhao, et al. 3d-front: 3d furnished rooms with layouts and semantics. In *International Conference on Computer Vision (ICCV)*, 2021. 2, 5, 6, 1, 3

[20] Samir Yitzhak Gadre, Mitchell Wortsman, Gabriel Ilharco, Ludwig Schmidt, and Shuran Song. Cows on pasture: Baselines and benchmarks for language-driven zero-shot object navigation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[21] Georgia Gkioxari, Jitendra Malik, and Justin Johnson. Mesh r-cnn. In *International Conference on Computer Vision (ICCV)*, 2019. 2, 3

[22] Benoit Guillard, Federico Stella, and Pascal Fua. Meshudf: Fast and differentiable meshing of unsigned distance field networks. In *European Conference on Computer Vision (ECCV)*, 2022. 3

[23] Haoyu Guo, Sida Peng, Haotong Lin, Qianqian Wang, Guofeng Zhang, Hujun Bao, and Xiaowei Zhou. Neural 3d scene reconstruction with the manhattan-world assumption. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 8

[24] Feng Han and Song-Chun Zhu. Bottom-up/top-down image parsing by attribute graph grammar. In *International Conference on Computer Vision (ICCV)*, 2005. 2

[25] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *International Conference on Computer Vision (ICCV)*, 2017. 4

[26] Varsha Hedau, Derek Hoiem, and David Forsyth. Recovering the spatial layout of cluttered rooms. In *International Conference on Computer Vision (ICCV)*, 2009. 2

[27] Derek Hoiem, Alexei A Efros, and Martial Hebert. Automatic photo pop-up. In *ACM SIGGRAPH / Eurographics Symposium on Computer Animation (SCA)*, 2005. 2

[28] Chenguang Huang, Oier Mees, Andy Zeng, and Wolfram Burgard. Visual language maps for robot navigation. In *International Conference on Robotics and Automation (ICRA)*, 2023. 2

[29] Siyuan Huang, Siyuan Qi, Yinxue Xiao, Yixin Zhu, Ying Nian Wu, and Song-Chun Zhu. Cooperative holistic scene understanding: Unifying 3d object, layout, and camera pose estimation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018. 2

[30] 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 *European Conference on Computer Vision (ECCV)*, 2018. 2

[31] Siyuan Huang, Yixin Chen, Tao Yuan, Siyuan Qi, Yixin Zhu, and Song-Chun Zhu. Perspectivenet: 3d object detection from a single rgb image via perspective points. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. 2

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

[33] Heewoo Jun and Alex Nichol. Shap-e: Generating conditional 3d implicit functions. *arXiv preprint arXiv:2305.02463*, 2023. 2, 6, 7, 3

[34] Abhishek Kar, Christian Häne, and Jitendra Malik. Learning a multi-view stereo machine. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017. 3

[35] Hiroharu Kato, Deniz Beker, Mihai Morariu, Takahiro Ando, Toru Matsuoka, Wadim Kehl, and Adrien Gaidon. Differentiable rendering: A survey. *arXiv preprint arXiv:2006.12057*, 2020. 3

[36] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2014. 2

[37] Arno Knapitsch, Jaesik Park, Qian-Yi Zhou, and Vladlen Koltun. Tanks and temples: Benchmarking large-scale scene reconstruction. *ACM Transactions on Graphics (TOG)*, 36 (4):1–13, 2017. 5, 2

[38] Verica Lazova, Vladimir Guzov, Kyle Olszewski, Sergey Tulyakov, and Gerard Pons-Moll. Control-nerf: Editable feature volumes for scene rendering and manipulation. In *Proceedings of Winter Conference on Applications of Computer Vision (WACV)*, 2023. 8

[39] David C Lee, Martial Hebert, and Takeo Kanade. Geometric reasoning for single image structure recovery. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009. 2

[40] 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 (ECCV)*, 2022. 2, 3, 4, 5, 6, 1

[41] Lingjie Liu, Jiatao Gu, Kyaw Zaw Lin, Tat-Seng Chua, and Christian Theobalt. Neural sparse voxel fields. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 3, 8

[42] 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 *International Conference on Computer Vision (ICCV)*, 2023. 2, 6, 7, 3

[43] Xiaoxiao Long, Cheng Lin, Lingjie Liu, Yuan Liu, Peng Wang, Christian Theobalt, Taku Komura, and Wenping Wang. Neuraludf: Learning unsigned distance fields for multi-view reconstruction of surfaces with arbitrary topologies. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 3

[44] William E Lorensen and Harvey E Cline. Marching cubes: A high resolution 3d surface construction algorithm. *ACM Transactions on Graphics (TOG)*, 21(4):163–169, 1987. 5

[45] Xiaojian Ma, Silong Yong, Zilong Zheng, Qing Li, Yitao Liang, Song-Chun Zhu, and Siyuan Huang. Sqa3d: Situated question answering in 3d scenes. In *International Conference on Learning Representations (ICLR)*, 2023. 2

[46] Ricardo Martin-Brualla, Noha Radwan, Mehdi SM Sajjadi, Jonathan T Barron, Alexey Dosovitskiy, and Daniel Duckworth. Nerf in the wild: Neural radiance fields for unconstrained photo collections. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 3

[47] Luke Melas-Kyriazi, Iro Laina, Christian Rupprecht, and Andrea Vedaldi. Realfusion: 360deg reconstruction of any object from a single image. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 2

[48] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 3, 5, 2

[49] 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. 3, 1

[50] Niloy J Mitra, Vladimir Kim, Ersin Yumer, Moos Hueting, Nathan Carr, and Pradyumna Reddy. Seethrough: Finding objects in heavily occluded indoor scene images. In *International Conference on 3D Vision (3DV)*, 2018. 2

[51] Richard A Newcombe, Steven J Lovegrove, and Andrew J Davison. Dtam: Dense tracking and mapping in real-time. In *International Conference on Computer Vision (ICCV)*, 2011. 2

[52] 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 *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2, 3, 5, 6, 1

[53] Michael Niemeyer and Andreas Geiger. Giraffe: Representing scenes as compositional generative neural feature fields. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 8

[54] Michael Niemeyer, Lars Mescheder, Michael Oechsle, and Andreas Geiger. Occupancy flow: 4d reconstruction by learning particle dynamics. In *International Conference on Computer Vision (ICCV)*, 2019. 3

[55] Michael Niemeyer, Lars Mescheder, Michael Oechsle, and Andreas Geiger. Differentiable volumetric rendering: Learning implicit 3d representations without 3d supervision. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 3

[56] Michael Oechsle, Songyou Peng, and Andreas Geiger. Unisurf: Unifying neural implicit surfaces and radiance fields for multi-view reconstruction. In *International Conference on Computer Vision (ICCV)*, 2021. 3

[57] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 3, 4

[58] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. 2

[59] Tobias Pfaff, Meire Fortunato, Alvaro Sanchez-Gonzalez, and Peter W Battaglia. Learning mesh-based simulation with graph networks. In *International Conference on Learning Representations (ICLR)*, 2021. 3

[60] Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. In *International Conference on Learning Representations (ICLR)*, 2023. 3

[61] Charles R Qi, Hao Su, Kaichun Mo, and Leonidas J Guibas. Pointnet: Deep learning on point sets for 3d classification and segmentation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 3

[62] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *International Conference on Computer Vision (ICCV)*, 2021. 8

[63] Shunsuke Saito, Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, and Hao Li. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *International Conference on Computer Vision (ICCV)*, 2019. 3

[64] Qiuhong Shen, Xingyi Yang, and Xinchao Wang. Anything-3d: Towards single-view anything reconstruction in the wild. *arXiv preprint arXiv:2304.10261*, 2023. 2

[65] Yawar Siddiqui, Justus Thies, Fangchang Ma, Qi Shan, Matthias Nießner, and Angela Dai. Texturify: Generating textures on 3d shape surfaces. In *European Conference on Computer Vision (ECCV)*, 2022. 2

[66] Shuran Song, Samuel P Lichtenberg, and Jianxiong Xiao. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 2, 8

[67] 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 *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 2, 5, 6, 1, 3

[68] Jiaxiang Tang. Stable-dreamfusion: Text-to-3d with stable-diffusion, 2022. 3

[69] Junshu Tang, Tengfei Wang, Bo Zhang, Ting Zhang, Ran Yi, Lizhuang Ma, and Dong Chen. Make-it-3d: High-fidelity 3d creation from a single image with diffusion prior. In *International Conference on Computer Vision (ICCV)*, 2023. 2

[70] Ayush Tewari, Ohad Fried, Justus Thies, Vincent Sitzmann, Stephen Lombardi, Kalyan Sunkavalli, Ricardo Martin-Brualla, Tomas Simon, Jason Saragih, Matthias Niessner, et al. State of the art on neural rendering. In *Computer Graphics Forum*, 2020. 3

[71] Haochen Wang, Xiaodan Du, Jiahao Li, Raymond A Yeh, and Greg Shakhnarovich. Score jacobian chaining: Lifting pretrained 2d diffusion models for 3d generation. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 3

[72] Jiepeng Wang, Peng Wang, Xiaoxiao Long, Christian Theobalt, Taku Komura, Lingjie Liu, and Wenping Wang. Neuris: Neural reconstruction of indoor scenes using normal priors. In *European Conference on Computer Vision (ECCV)*, 2022. 8

[73] Nanyang Wang, Yinda Zhang, Zhuwen Li, Yanwei Fu, Wei Liu, and Yu-Gang Jiang. Pixel2mesh: Generating 3d mesh models from single rgb images. In *European Conference on Computer Vision (ECCV)*, 2018. 5, 2

[74] Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 3

[75] Jiajun Wu, Chengkai Zhang, Tianfan Xue, Bill Freeman, and Josh Tenenbaum. Learning a probabilistic latent space of object shapes via 3d generative-adversarial modeling. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2016. 3

[76] 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. 8

[77] Qiangeng Xu, Weiyue Wang, Duygu Ceylan, Radomir Mech, and Ulrich Neumann. Disn: Deep implicit surface network for high-quality single-view 3d reconstruction. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. 3

[78] Bangbang Yang, Yinda Zhang, Yinghao Xu, Yijin Li, Han Zhou, Hujun Bao, Guofeng Zhang, and Zhaopeng Cui. Learning object-compositional neural radiance field for editable scene rendering. In *International Conference on Computer Vision (ICCV)*, 2021. 8

[79] Lior Yariv, Yoni Kasten, Dror Moran, Meirav Galun, Matan Atzmon, Basri Ronen, and Yaron Lipman. Multiview neural surface reconstruction by disentangling geometry and appearance. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 4

[80] Lior Yariv, Jiatao Gu, Yoni Kasten, and Yaron Lipman. Volume rendering of neural implicit surfaces. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 3, 2

[81] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelnerf: Neural radiance fields from one or few images. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 4, 7, 8, 1, 3

[82] 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 (NeurIPS)*, 2022. 4, 8, 1

[83] Yu-Jie Yuan, Yang-Tian Sun, Yu-Kun Lai, Yuewen Ma, Rongfei Jia, and Lin Gao. Nerf-editing: geometry editing of neural radiance fields. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 8

[84] Amir R Zamir, Alexander Sax, Nikhil Cheerla, Rohan Suri, Zhangjie Cao, Jitendra Malik, and Leonidas J Guibas. Robust learning through cross-task consistency. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2, 8, 4

[85] 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 *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2, 3, 5, 6, 8, 1

[86] Lichen Zhao, Daigang Cai, Jing Zhang, Lu Sheng, Dong Xu, Rui Zheng, Yinjie Zhao, Lipeng Wang, and Xibo Fan. Towards explainable 3d grounded visual question answering: A new benchmark and strong baseline. *IEEE Transactions on Circuits and Systems for Video Technology*, 2022. 2

Single-view 3D Scene Reconstruction with High-fidelity Shape and Texture

Supplementary Material

In Sec. S1, we present how we prepare the training data on the 3D-FRONT [19] and Pix3D [67] datasets. In Sec. S2, we report more implementation details, including model architecture, learning curriculum, and training strategies. Sec. S3 presents more experimental details, evaluation metrics, results, and failure case discussion. For a more comprehensive view of the qualitative results, we recommend referring to the supplementary video with detailed visualizations and animations.

S1. Training data preparation

S1.1. Datasets and splits

3D-FRONT [19] contains synthetic and professionally-designed indoor scenes populated by high-quality textured 3D models from 3D-FUTURE [19]. Following Liu *et al*. [40], we use about 20K scene images for training and validation and report quantitative evaluation on 2000 images in 8 different categories. Pix3D [67] provides real-world images along with corresponding 3D furniture models that are aligned with the images in 9 object categories. In the splits from Nie *et al*. [52] and [85], there is a significant overlap of objects between the training and testing split. Consequently, we follow Liu *et al*. [40] to employ a split without overlapping objects.

S1.2. Fixing CAD models

One major issue with 3D-FRONT [19] dataset is that the majority of the CAD models are not watertight, which hinders the learning of SDF as the neural implicit surface representations in our framework. To mitigate this issue, we first utilize the automatic remesh method following Liu *et al*. [40] to transform the non-watertight models into watertight ones with a more evenly-distributed topology. Additionally, we have noticed that some models, primarily beds and sofas, lack a back or underside, confusing the model when learning 3D object priors. Consequently, we have manually fixed a total of 1734 models to resolve this issue.

S1.3. SDF supervision

In our framework, we employ supervision from both explicit 3D shapes and volume rendering of color, depth, and normal images. The direct 3D supervision focuses on minimizing the differences between the predicted and actual SDF values of the sampled points. To generate ground-truth SDF values for a given CAD model, we voxelize its 3D bounding box in a normalized coordinate system with a resolution of 64. The SDF value for each voxel grid center is calculated as the distance to the mesh surface. The SDF value is positive if the point is outside the surface and negative if inside. During training, we obtain the ground-truth SDF value for the query points using trilinear interpolation.

S1.4. Monocular cues rendering

To further alleviate the ambiguities in recovering 3D shapes from single-view inputs, we follow Yu *et al*. [82] to exploit monocular depth, normal, and segmentation cues to facilitate the training process. However, since these images are not available in the 3D-FRONT [19] dataset, we render them using the 3D scans of the scene, 3D CAD models of the objects, and the camera's intrinsic and extrinsic parameters provided in the dataset. The Pix3D [67] dataset offers instance segmentation but lacks depth and normal images. Since rendering is impossible, we utilize the estimated depth and normal maps as the pseudo-ground-truth from state-of-the-art estimator [17]. Note that the depth, normal, and segmentation information is solely used during the training stage to guide the model's learning process, and none is required during the inference stage. This ensures that our model remains flexible and applicable to various scenarios.

S2. Technical details

S2.1. Model architecture

For the implicit network, we use an 8-layer MLP with hidden dimension 256. We implement the rendering network with a 2-layer MLP with hidden dimension 256. We use Softplus activation for the implicit network and Sigmoid activation for the rendering network. We use a ResNet34 backbone pre-trained on ImageNet as the image encoder following Yu *et al*. [81]. We use positional encoding γ from NeRF [49] for the spatial coordinates, with $L = 6$ exponentially increasing frequencies:

$$\begin{aligned}\gamma(x) = (&\sin(2^0\omega\mathbf{x}), \cos(2^0\omega \\ &\sin(2^1\omega\mathbf{x}), \cos(2^1\omega\mathbf{x}), \\ &\dots, \\ &\sin(2^{L-1}\omega\mathbf{x}), \cos(2^{L-1}\omega\mathbf{x}))\end{aligned} \tag{S1}$$

S2.2. Learning curriculum

As discussed in the paper, we propose a two-stage learning curriculum to effectively employ supervision from both 3D shapes and volume rendering. In Stage One, the loss weight is set to be 1 for the 3D supervision $\mathcal{L}_{3D}$ and 0 for the rest of the losses. In Stage Two, we linearly increase the loss weights for color, depth, and normal supervision while

maintaining a constant weight for the 3D supervision loss. We utilize the $\mathcal{L}_{3D}$ loss curve when training with 3D supervision only to determine the suitable value for λ_0. Specifically, once the training approaches convergence according to the SDF loss curve, we identify the epoch at this point as the suitable λ_0 value. In our paper, we choose $\lambda_0 = 150$. Slight deviations below or above this threshold have minimal impact. However, significantly reducing the value, such as $\lambda_0 = 0$ or $\lambda_0 = 70$, results in substantial degradation of performance because early injection of 2D supervision may affect the 3D shape learning due to the shape-appearance ambiguity. More discussion can be referred to the ablation experiments in Sec. 4.1.

S2.3. Training strategy

During training, the image, depth, and normal images have the same resolution of 484×648. The implicit network takes 3D points as input in canonical coordinate system to ease the learning of reconstructing indoor objects with implicit representations. The volume rendering is performed along the sampled points in the camera coordinates to calculate the color, depth, and normal values for all the pixels in a minibatch. We sample 64 rays per iteration and apply the error-bounded sampling strategy introduced by Yariv *et al*. [80]. We additionally apply 3D supervision to another 30,000 points uniformly sampled near the ground-truth surfaces (set $\mathcal{X}$). Our model is trained for 400 epochs on 4 NVIDIA-A100 GPUs with a batch size of 96. We implement our method in PyTorch [58] and use the Adam optimizer [36]. The learning rate is initialized as 1e-3 and decays by a factor of 0.2 in the 330^{th} and 370^{th} epochs.

S3. More experiment details and results

S3.1. Evaluation metrics

3D object reconstruction Following Wang *et al*. [73] and Mescheder *et al*. [48], we adopt Chamfer Distance (CD), F-Score and Normal Consistency as the metrics to evaluate 3D object reconstruction. Following prior work [40], we proportionally scale the longest edge of reconstructed objects to $2m$ to calculate CD and F-Score. After mesh alignment with Iterative Closest Point (ICP) [4], we uniformly sample points from our prediction and ground-truth. CD is calculated by summing the squared distances between the nearest neighbor correspondences of two point clouds after mesh alignment. The values of CD are reported in units of 10^{-3}. We calculate precision and recall by checking the percentage of points in prediction or ground-truth that can find the nearest neighbor from the other within a threshold of $2mm$. F-Score [37] is the harmonic mean of precision and recall on in prediction and ground-truth. Finally, to measure how well the methods can capture higher-order information, the normal consistency score is computed as the mean absolute dot product of the normals in one mesh and the normals at the corresponding nearest neighbors in the other mesh after alignment.

Depth and normal estimation We adopt the L1 error for depth estimation and utilize both L1 and Angular errors for normal estimation following Eftekhar *et al*. [17]. Since the baseline methods [17, 84] estimate relative depth values rather than absolute ones, we first align the estimated depth values with the ground-truth values to the range of $[0, 1]$ using the approach outlined in Eftekhar *et al*. [17]. After the alignment, we compute the L1 error to quantify the discrepancy. For normal estimation, we normalize both the estimated and ground-truth values to unit vectors and then compute the L1 error and the angle error for evaluation.

S3.2. Indoor object reconstruction

S3.2.1 Experiments on 3D-FRONT and Pix3D

In this subsection, we provide more details about our reproduced results and more qualitative results. Tabs. S1 and S3 list the quantitative outcomes on 3D-FRONT [19] and Pix3D [67] datasets documented by Liu *et al*. [40] in the original paper. The "NC" (normal consistency) columns in these tables are empty, as Liu *et al*. [40] did not measure normal consistency in their paper. Tabs. S2 and S4 present our reproduced results, which are comparable with the results in Tabs. S1 and S3. We additionally provide more qualitative results, Fig. S3 on 3D-FRONT and Fig. S4 on Pix3D. As can be seen from these figures, our model can learn finer and smoother surfaces with high-fidelity textures.

S3.2.2 Failure cases

In this section, we present and diagnose some representative failure examples. Fig. S1(a) illustrates that occlusion between objects can lead to distortions in both shape and appearance recovery, particularly in the occluded areas. Prior work [33, 42] necessitates unoccluded object images as in-

Table S1. **Object reconstruction on the 3D-FRONT [19] dataset.** Our model achieves the best performance on mean CD and F-Score, as well as the best NC on all object categories, outperforming MGN [52], LIEN [85], and InstPIFu [40].

Category		bed	chair	sofa	table	desk	nightstand	cabinet	bookshelf	mean
CD ↓	MGN	15.48	11.67	8.72	20.90	17.59	17.11	13.13	10.21	14.07
	LIEN	16.81	41.40	9.51	35.65	26.63	16.78	7.44	11.70	28.52
	InstPIFu	18.17	14.06	7.66	23.25	33.33	**11.73**	**6.04**	8.03	14.46
	Ours	**4.96**	**10.52**	**4.53**	**16.12**	**25.86**	17.90	6.79	**3.89**	**10.45**
F-Score ↑	MGN	46.81	57.49	64.61	49.80	46.82	47.91	54.18	54.55	55.64
	LIEN	44.28	31.61	61.40	43.22	37.04	50.76	69.21	55.33	45.63
	InstPIFu	47.85	59.08	67.60	56.43	**48.49**	57.14	**73.32**	66.13	61.32
	Ours	**76.34**	**69.17**	**80.06**	**67.29**	47.12	**58.48**	70.45	**85.93**	**71.36**
NC ↑	MGN	-	-	-	-	-	-	-	-	-
	LIEN	-	-	-	-	-	-	-	-	-
	InstPIFu	-	-	-	-	-	-	-	-	-
	Ours	**0.896**	**0.833**	**0.894**	**0.838**	**0.764**	**0.897**	**0.856**	**0.862**	**0.854**

put to avoid this problem. One of the future directions is to involve the accurate reconstruction of object shapes with textures under heavy occlusions. The bookshelf depicted in Fig. S1(b) exemplifies the challenges our framework encounters when attempting to reconstruct intricate geometry. We hypothesize that this can be attributed to the limited representation power of SDF as the implicit surface representations for thin surfaces. Specifically, it requires the model to recognize and reconstruct abrupt changes in the signed distance field within centimeters, transitioning from positive (outside) to negative (inside) and then back to positive again. To address such issues, we suggest future endeavors in integrating unsigned distance field [22, 43] with volume rendering to capture such intricate geometry and non-watertight meshes.

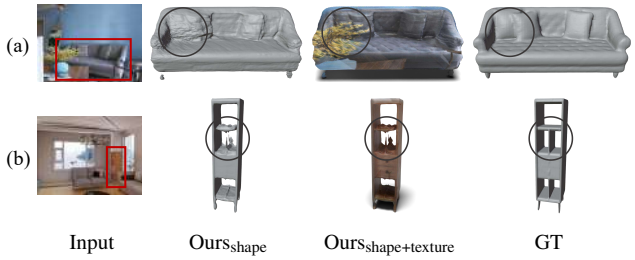

Figure S1. **Qualitative examples for failure cases.** Two representatives with (a) occlusions and (b) intricate geometry.

S3.2.3 Comparison with prior-guided models

We randomly select 100 samples from the test split in 3D-FRONT to perform a comparative assessment of the reconstruction performance between our framework and prior-guided models, *i.e.*, Zero-1-to-3 [42] and Shap·E [33]. In the original paper, Zero-1-to-3 [42] utilized SJC [71] for the 3D reconstruction task, whereas we follow Tang *et al.* [68] to employ DreamFusion [60] to achieve enhanced results. In Shap·E [33], the text prompts are required for reconstruction using the conditional generative model. We utilize the ground-truth object category as the designated text prompt to maintain fairness. It is noteworthy that both Zero-1-to-3 [42] and Shap·E [33] necessitate background-free input images for the target objects. Consequently, we utilize the ground truth mask to segment the object as input. For our model, we use the original image as input.

Table S2. **Reproduced results on the 3D-FRONT [19] dataset.** †: Results reproduced from the official repository.

Category		bed	chair	sofa	table	desk	nightstand	cabinet	bookshelf	mean
CD ↓	MGN†	5.95	14.31	6.01	24.21	38.07	20.27	12.49	11.42	14.97
	LIEN†	4.58	11.85	7.21	30.40	42.52	20.84	14.81	18.13	16.33
	InstPIFu†	7.68	13.79	6.78	21.56	31.32	**13.14**	**5.94**	13.79	13.63
	Ours	**4.96**	**10.52**	**4.53**	**16.12**	**25.86**	17.90	6.79	**3.89**	**10.45**
F-Score ↑	MGN†	65.47	52.83	68.98	54.04	42.9	45.01	52.01	55.26	57.19
	LIEN†	72.16	62.23	68.25	48.18	32.07	42.87	49.62	43.12	58.37
	InstPIFu†	62.28	66.31	69.65	58.73	40.49	57.52	**76.59**	71.48	65.34
	Ours	**76.34**	**69.17**	**80.06**	**67.29**	**47.12**	**58.48**	70.45	**85.93**	**71.36**
NC ↑	MGN†	0.829	0.758	0.819	0.785	0.711	0.833	0.802	0.719	0.787
	LIEN†	0.822	0.793	0.803	0.755	0.701	0.814	0.801	0.747	0.786
	InstPIFu†	0.799	0.782	0.846	0.804	0.708	0.844	0.841	0.790	0.810
	Ours	**0.896**	**0.833**	**0.894**	**0.838**	**0.764**	**0.897**	**0.856**	**0.862**	**0.854**

Table S3. **Object Reconstruction on the Pix3D [67] dataset.** On the non-overlapped split [40], our model outperforms the state-of-the-art methods by significant margins.

Category		bed	bookcase	chair	desk	sofa	table	tool	wardrobe	misc	mean
CD ↓	MGN	22.91	33.61	56.47	33.95	9.27	81.19	94.70	10.43	137.50	44.32
	LIEN	11.18	29.61	40.01	65.36	10.54	146.13	29.63	4.88	144.06	51.31
	InstPIFu	10.90	7.55	32.44	**22.09**	8.13	45.82	10.29	**1.29**	47.31	24.65
	Ours	**6.31**	**7.21**	**26.23**	28.63	**5.68**	**43.87**	**8.29**	2.07	**35.03**	**21.79**
F-Score ↑	MGN	34.69	28.42	35.67	65.36	51.15	17.05	57.16	52.04	10.41	36.20
	LIEN	37.13	15.51	25.70	26.01	49.71	21.16	5.85	59.46	11.04	31.45
	InstPIFu	54.99	62.26	35.30	**47.30**	56.54	37.51	64.24	**94.62**	27.03	45.62
	Ours	**68.78**	**66.69**	**55.18**	42.49	**71.22**	**51.93**	**65.38**	91.84	**46.92**	**59.71**
NC ↑	MGN	-	-	-	-	-	-	-	-	-	-
	LIEN	-	-	-	-	-	-	-	-	-	-
	InstPIFu	-	-	-	-	-	-	-	-	-	-
	Ours	**0.825**	**0.689**	**0.693**	**0.776**	**0.866**	**0.835**	**0.645**	**0.960**	**0.599**	**0.778**

Table S4. **Reproduced results on the Pix3D [67] dataset.** †: Results reproduced from the official repository.

Category		bed	bookcase	chair	desk	sofa	table	tool	wardrobe	misc	mean
CD ↓	MGN†	11.73	17.50	36.74	29.63	7.02	47.85	25.63	6.61	62.32	27.9
	LIEN†	16.31	26.05	33.06	39.84	7.34	76.97	27.84	4.66	126.88	33.66
	InstPIFu†	9.82	7.73	31.55	**23.18**	7.28	45.89	9.64	**1.79**	43.4	24.35
	Ours	**6.31**	**7.21**	**26.23**	28.63	**5.68**	**43.87**	**8.29**	2.07	**35.03**	**21.79**
F-Score ↑	MGN†	51.51	40.85	49.42	44.06	59.48	45.02	45.52	64.76	26.23	50.55
	LIEN†	50.20	36.26	48.41	32.55	64.98	29.97	38.72	76.03	18.29	47.87
	InstPIFu†	57.02	61.45	38.06	**45.98**	62.76	37.26	63.50	**93.94**	40.14	48.09
	Ours	**68.78**	**66.69**	**55.18**	42.49	**71.22**	**51.93**	**65.38**	91.84	**46.92**	**59.71**
NC ↑	MGN†	0.737	0.592	0.525	0.633	0.756	0.794	0.531	0.809	0.563	0.659
	LIEN†	0.706	0.514	0.591	0.581	0.775	0.619	0.506	0.844	0.481	0.646
	InstPIFu†	0.782	0.646	0.547	0.758	0.753	0.796	0.639	0.951	0.580	0.683
	Ours	**0.825**	**0.689**	**0.693**	**0.776**	**0.866**	**0.835**	**0.645**	**0.960**	**0.599**	**0.778**

S3.3. Rendering capability

S3.3.1 Novel view synthesis

To measure the capability for novel view synthesis, we compare our model with PixelNeRF [81] on the category-agnostic model, which is first trained on the ShapeNet [7] and then fine-tuned using the 3D-FRONT [19]. Results in Fig. 7 show that PixelNeRF struggles to render images outside the vicinity of the original viewpoints where our model is capable of generating meaningful renderings from novel views. The main reason behind this is that our method employs 3D supervision, which helps the model learn better 3D object priors. PixelNeRF fails to acquire a meaningful 3D prior from the training data, especially when each image exists independently in 3D-FRONT [19], which is in stark contrast to the ShapeNet [7] where images are presented in a sequence of related perspectives. Fig. S5 shows that our model not only generates high-quality new perspective images but also produces reasonable depth and normal maps by volume rendering.

Table S5. **Novel views depth and normal estimation.** We evaluate depth using L1 ↓ and normal using L1↓ / Angular°↓ error.

Angle	Method	Depth(L1↓)	Normal(L1↓ / Angular°↓)
0°	XTC [84]	1.188	12.712 / 14.309
	Omnidata [17]	0.734	10.015 / 11.257
	Ours	0.992	10.962 / 12.392
5°	Ours	1.0313	11.2132 / 12.631
10°		1.0911	11.5866 / 13.1078
15°		1.1796	12.0943 / 13.7095
20°		1.2624	12.6315 / 14.2108
30°		1.3934	13.4501 / 15.1012
40°		1.4694	15.0034 / 16.6108

S3.3.2 Depth and normal estimation

Our framework can serve as a proficient single-view depth and normal estimator, and to compare with zero-shot state-of-the-art methods [17, 84], we randomly select 200 samples from the test split in 3D-FRONT. To assess our model's capability to generate depth and normal maps from novel views, we rotated the camera vertically left and right by 5°, 10°, 15°, 20°, 30°, and 40°. It is worth noting that only our method possesses the capability to estimate depth and normal from novel views. Tab. S5 and Fig. S2 show that our model can consistently produce satisfactory outcomes, even when the viewing angle changes significantly.

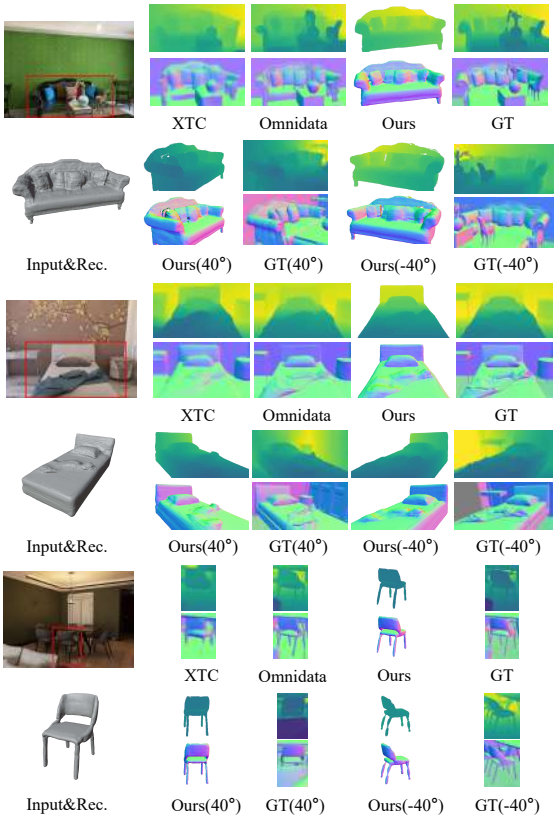

Figure S2. **Qualitative results for depth and normal estimation.** Our method produces results comparable with [17, 84] on the input view, and can estimate depth and normal for novel views.

Figure S3. **More qualitative results from 3D-FRONT [40]**.

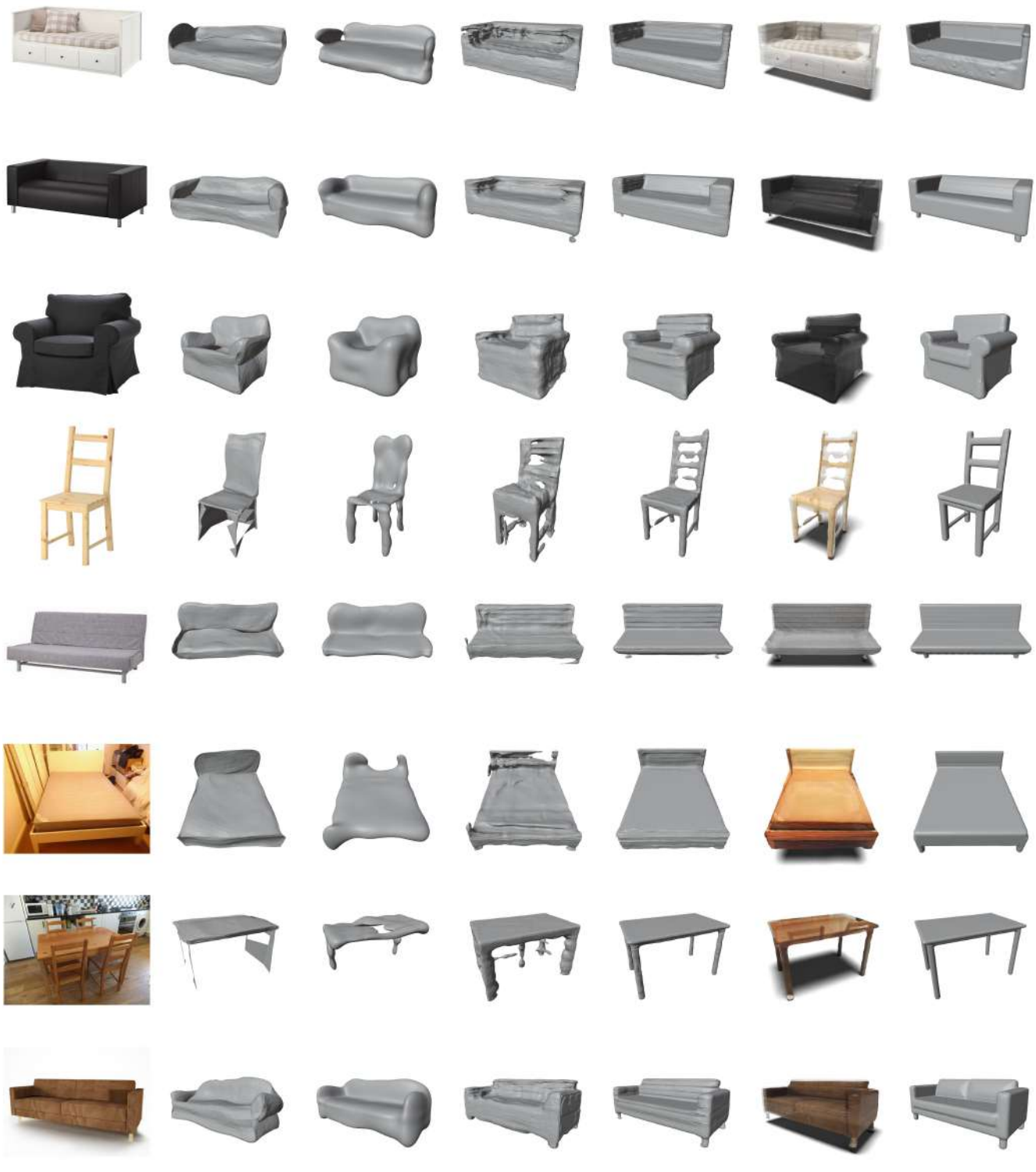

Figure S4. **More qualitative results from Pix3D [67].**

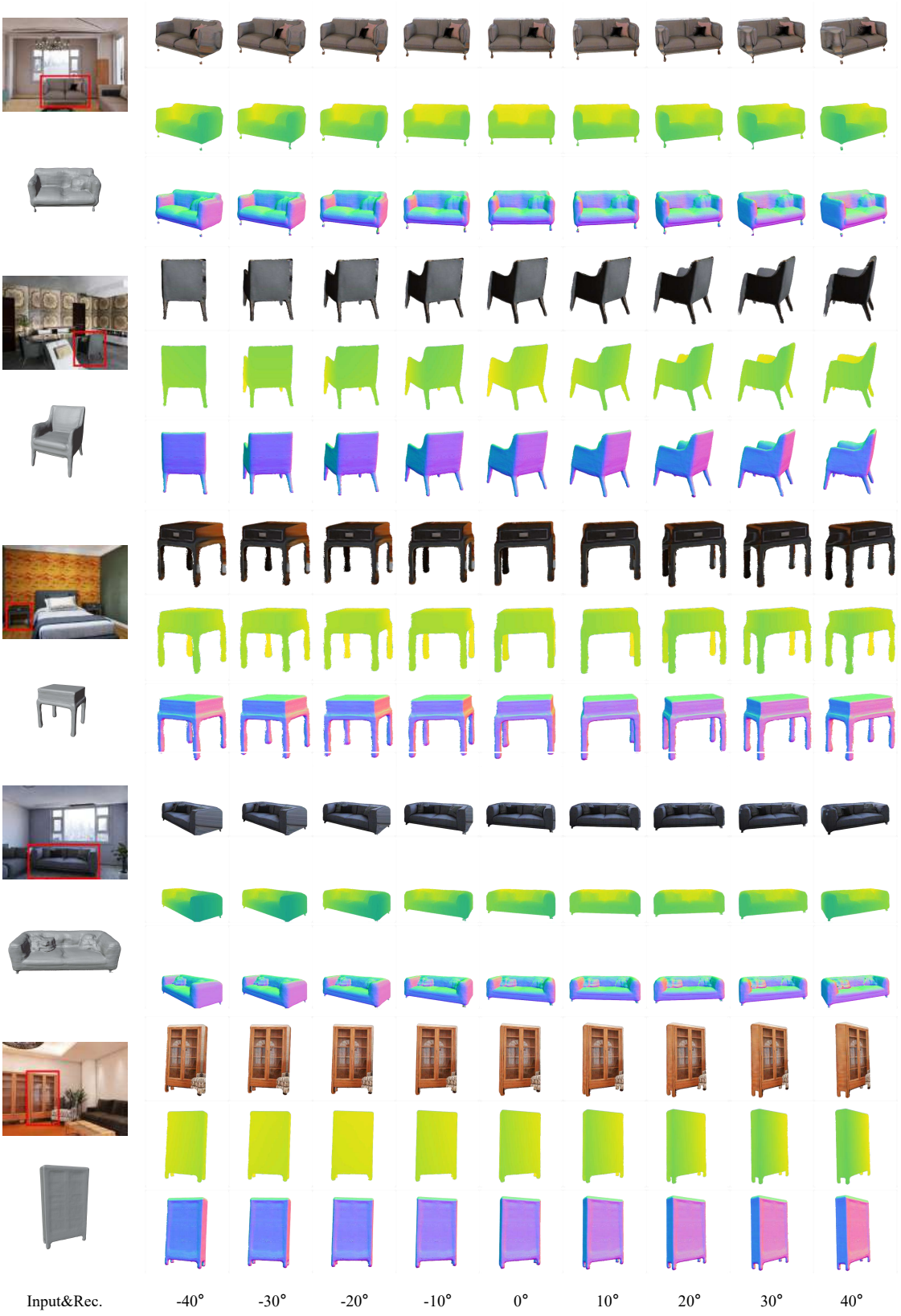

Figure S5. **More qualitative results for novel views synthesis.**