A. Network Architecture

We adopt the similar network structure as in [37] and add the appearance code utilized in [43] for all the baseline models and ours. Fig. 10 shows our detailed network architecture.

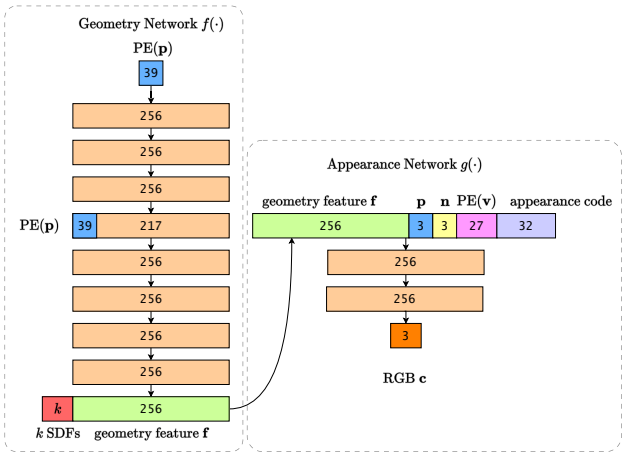

Figure 10. **Network Architecture.** The geometry network corresponds to the SDF function f and the appearance network denotes the appearance g in the main paper.

The geometry network is a 8-layer MLP with hidden dimension 256, and one skip connection at the fourth layer. The input of this network is the point coordinate mapped by a fixed positional encoding [20]. The output consists of a 256 dimensional geometry feature and k SDF values for each object. These SDF values can be transformed to semantic logits using the function proposed in [37].

For the input of the appearance network, we adopt the design in [43]. We optimize a per-frame appearance code during the training and use this per-frame code to model the varying light and blurry condition in each image for better reconstruction. The appearance code is concatenated with the viewing direction (also mapped by positional encoding), the normal of the scene SDF, the geometry feature and the point coordinate. The appearance network consists of two layers with hidden dimension 256, and outputs a 3-channel RGB color.

We use Softplus activation for the geometry network and use ReLU activation for the appearance network, the RGB color is obtained after passing the network's output through the Sigmoid activation.

B. Loss Functions Details

We elaborate all the losses and the weight choices used in the optimization in this section.

RGB Reconstruction Loss: To learn the surface from images input, we need to minimize the difference between ground-truth pixel color and the rendered color. We follow the previous works [43, 37] here for the RGB reconstruction loss:

$$\mathcal{L}_{\text{RGB}} = \sum_{\mathbf{r}} ||\hat{\mathbf{C}}(\mathbf{r}) - \mathbf{C}(\mathbf{r})||_1. \tag{12}$$

Here $\hat{\mathbf{C}}(\mathbf{r})$ is the rendered color from volume rendering and $\mathbf{C}(\mathbf{r})$ denotes the ground truth.

Depth Consistency Loss: Monocular depth and normal cues [43] can greatly benefit indoor scene reconstruction. For the depth consistency, we minimize the difference between rendered depth $\hat{D}(\mathbf{r})$ and the depth estimation $\bar{D}(\mathbf{r})$ from the Omnidata [5] model:

$$\mathcal{L}_{\text{D}} = \sum_{\mathbf{r}} ||(w\hat{D}(\mathbf{r}) + q) - \bar{D}(\mathbf{r})||^2, \tag{13}$$

where w and q are the scale and shift values to match the different scales. We solve w and q with a least-squares criterion, which has the closed-form solution. Please refer to the supplementary of [43] for a detailed computation process.

Normal Consistency Loss: Similar to the depth consistency loss, we also use the normal cues $\bar{N}$ from Omnidata model to supervise the rendered normal. Specifically, the normal consistency loss consists of L1 and the angular losses:

$$\mathcal{L}_{\text{N}} = \sum_{\mathbf{r}} ||\hat{N}(\mathbf{r}) - \bar{N}(\mathbf{r})||_1 + ||1 - \hat{N}(\mathbf{r})^{\text{T}}\bar{N}(\mathbf{r})||_1. \tag{14}$$

Here the volume-rendered normal and normal estimation will be transformed into the same coordinate system by the camera pose.

Semantic Loss: We minimize the semantic loss between volume-rendered semantic logits of each pixel and the ground-truth pixel semantic class. Here the semantic objective is implemented as a cross-entropy loss:

$$\mathcal{L}_{\text{S}} = \sum_{\mathbf{r}} \sum_{j=1}^{k} -\hat{h}_j(\mathbf{r}) \log h_j(\mathbf{r}). \tag{15}$$

The $\hat{h}_j(\mathbf{r})$ is the ground-truth semantic probability for j-th object, which is 1 or 0.

Eikonal Loss: Following common practice, we also add an Eikonal term on the sampled points to regularize SDF values in 3D space:

$$\mathcal{L}_{\text{E}} = \sum_{i}^{n} (||\nabla \min_{1 \le j \le k} s_j(\mathbf{p}_i)||_2 - 1) \tag{16}$$

Here the eikonal loss is applied to the gradient of the scene SDF, which is the minimum of all the SDFs.

C. Evaluation Metrics

To evaluate the reconstruction performance, we use the Chamfer Distance and F-score with a threshold of 5cm in this paper. In detail, Chamfer Distance comes from *Accuracy* and *Completeness*, and F-score is derived from *Precision* and *Recall*. For point clouds P and P^* sampled from the predicted and the ground-truth mesh, we show the detailed computation procedure here:

$$\begin{aligned} \texttt{Accuracy} &= \underset{\mathbf{p}\in P}{\texttt{mean}} \left(\min_{\mathbf{p}^*\in P*} ||\mathbf{p}-\mathbf{p}^*||_1 \right), \\ \texttt{Completeness} &= \underset{\mathbf{p}^*\in P^*}{\texttt{mean}} \left(\min_{\mathbf{p}\in P} ||\mathbf{p}-\mathbf{p}^*||_1 \right), \\ \texttt{Chamfer-}L_1 &= \frac{\texttt{Accuracy}+\texttt{Completeness}}{2}. \end{aligned} \tag{17}$$

$$\begin{aligned} \texttt{Precision} &= \underset{\mathbf{p}\in P}{\texttt{mean}} \left(\min_{\mathbf{p}^*\in P*} ||\mathbf{p}-\mathbf{p}^*||_1 < 0.05 \right), \\ \texttt{Recall} &= \underset{\mathbf{p}^*\in P^*}{\texttt{mean}} \left(\min_{\mathbf{p}\in P} ||\mathbf{p}-\mathbf{p}^*||_1 < 0.05 \right), \\ \texttt{F-score} &= \frac{2\times\texttt{Precision}\times\texttt{Recall}}{\texttt{Precision}+\texttt{Recall}}. \end{aligned} \tag{18}$$

D. Synthetic Dataset Construction

In order to quantitatively evaluate the object-level reconstruction performance in the object-compositional indoor scenes, we create a synthetic dataset with object ground-truth geometry. In this part we elaborate on how to construct the Synthetic Dataset used in this paper. Despite that the dataset is not a major contribution of this paper, we will release it for future comparisons.

We use Blender [2] and an add-on BlenderNeRF [29] to construct the scenes (assign different object locations, lighting conditions and camera trajectories) and render the RGB images together with the camera poses. The Blender's camera coordinate system is different from the coordinate system in ScanNet, which requires an extra 180° rotation along with the x-axis on the recorded extrinsic matrix.

To render semantic masks, we switch each object's surface texture to a certain value and render again with the identical camera trajectories. We create 5 scenes and three of them contain 5 objects while other two contain 10 objects (background not included). For each scene we render 200 images and use the Omnidata [5] model to obtain the corresponding monocular depth and normal cues.

E. Additional Ablation Experiments

E.1. Parameter Ablation Study

In this part we provide the ablation experiment for the ϵ in proposed object point-SDF loss $\mathcal{L}_{\text{op}}$. Particularly, ϵ is a non-negative number as a threshold, that the objects' SDFs outside of the background should be larger than this value. We provide an ablation study on different ϵ values on synthetic scenes in Tab. 5.

	$\epsilon=0$	$\epsilon=0.05$	$\epsilon=0.1$	$\epsilon=0.2$
Chamfer-$L_1\downarrow$	0.187	**0.033**	0.034	0.036
F-Score $\uparrow$	0.755	**0.817**	0.812	0.793

Table 5. **Ablation Study on ϵ.** Metrics are evaluated and averaged on all the objects of all the synthetic scenes.

Intuitively, ϵ should be larger than 0 because the points behind the background are outside of each object's surface, i.e. the object SDFs of these points should be positive. Empirically we find setting $\epsilon = 0.05$ yields slightly better performance than 0.1 and 0.2. When setting $\epsilon = 0$, the object reconstruction performance drops significantly. We found the reason is that, when ϵ is 0, the SDFs of the sampled points can not all be effectively optimized to positive, yielding some negative SDF vaules, which results in the flaws in the empty space.

E.2. Backbone Ablation Study

When utilizing SDF as the surface geometry representation, there are typically two choices to combine SDF and volume rendering as proposed in[41, 36]. In the main paper we adopt the scheme proposed in NeuS [36] for RICO. We provide the comparison of reconstruction results on ScanNet (evaluated on whole scene) and Synthetic scenes (evaluated on each object), and report the results in Tab. 6. Since [36] explicitly models the angle difference of ray direction and the surface normal, it can have slightly better performance by better reconstructing the visible surface.

	ScanNet		Synthetic Object	
	Chamfer-$L_1\downarrow$	F-score $\uparrow$	Chamfer-$L_1\downarrow$	F-score $\uparrow$
RICO-VolSDF	0.090	0.592	0.042	0.751
RICO	0.088	0.624	0.033	0.817

Table 6. **Ablation Study on Backbone**. We show the reconstruction comparison of our methods using the volume rendering scheme proposed in [41] (RICO-VolSDF) and [36] (RICO, which represents the method proposed in main paper).

F. Construct the ObjSDF*-C Baseline

As stated in the main paper, we construct an improved baseline over ObjSDF*, named ObjSDF*-C, to provide better visualization and quantitative results. The main procedure is to use the reconstructed background surface to eliminate the parts of object reconstruction that are outside of the background range. The construction procedure is just a post-process method on the object meshes and do not

change the original nature of the ObjSDF* that only the visible surfaces are reconstructed.

For Synthetic scenes, since we set the background as a cubic room with a range of $[-2m, 2m]$ in three dimensions, i.e. the background is an axis-aligned box, we can directly use this range to segment the object reconstruction meshes. For ScanNet scenes, we use the ground-truth scene meshes to get a coarse range and manually finetune the range of each scene (ObjSDF*-C on ScanNet is only used for visualization, not for quantitative evaluation), then segment the object meshes based on the finetuned range.

G. Object Manipulation Implementation

To manipulate the reconstructed objects, a straightforward way is to directly manipulate the meshes. In the main paper we show the volume rendered normal maps and semantic masks before and after manipulation. In Fig. 11 we show how to implement the volume rendering in current framework. The core is to query the SDF value of manipulated object at the destined point, and combine it with other objects' SDF values. Notably, the color is decided by not only the coordinate but also the geometry feature (illustrated in Fig. 10). However, now the original point for other object SDFs and the manipulated point for the desired object SDF will result in two geometry feature vectors. In contrast to use minimum value to get the scene SDF from all SDFs, it's hard to decide how to fuse these two geometry features together in current framework. Now we only show the volume rendering results that are decided by SDF values, i.e. geometry, like the normal map and semantic mask.

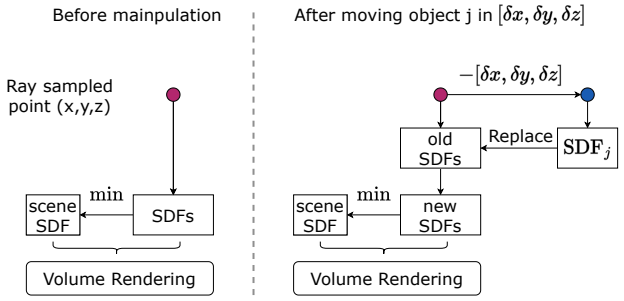

Figure 11. **Object Manipulation Implementation.** In this figure we show when moving the object j with $[\delta x, \delta y, \delta z]$ in each direction, how to implement volume rendering in the current network.

H. More Discussions

Smooth surface: It can be seen from the figures in the paper that surface from RICO is smoother in a way. The smooth effect actually comes from different rendering schemes proposed in VolSDF [41] and NeuS [36]. Empirically we notice that comparing to RICO-VolSDF, RICO-NeuS (default) has better quantitative results but can also be somehow oversmooth. It is an interesting future direction to investigate the effects of different reconstruction backbones in learning the compositional geometry.

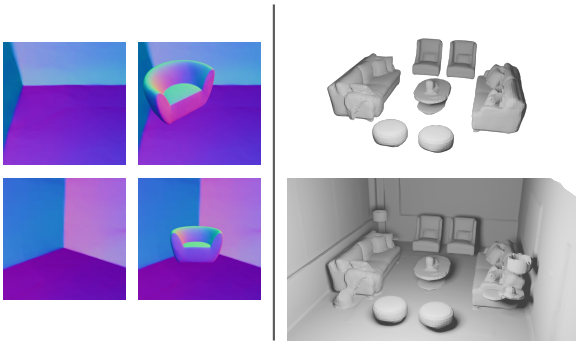

Figure 12. **Visualizations of:** (left) normal map of occluded corner in background, (right) reconstruction results on Replica.

Background smoothness regularization: Fig. 6 of our paper presents sharp background corners with occlusions. Here we provide a better visualization in Fig. 12 (left). There are mainly two scenarios for smoothness loss on occluded sharp geometry: 1) The sharp geometry is observable in some images. The reconstruction loss in these views will be dominant in optimization because the smoothness loss is of small weight and computed only once in several iterations, thus the sharp geometry can be reconstructed correctly; 2) The sharp geometry is completed occluded in all the views as shown in Fig. 12. By regularizing the depth and normal, we observe that the visible regions of wall and ground are smoothly extended, yielding a corner that is not perfectly perpendicular but without artifacts.

I. Qualitative Results on Individual Scenes

In Fig. 13 and Fig. 14, we provide the RICO's object-compositional reconstruction on ScanNet scenes and Synthetic scenes (with the object ground-truth) respectively. Here we also provide a visualization of RICO on one of the Replica scenes in Fig. 12 (right).

J. Limitations

In this work, we assume the indoor scene as convex room, that the ray shot inside of the room penetrates the background surface once. However when processing the more complex indoor scenes where one ray can go through multiple rooms, our object regularizations may require extra conditions to decide which points to be applied to. Additionally, the object-scene relation prior regularized the completeness from only the geometry perspective. The framework can be extended to utilizing more complex category-level prior for better reconstruction.

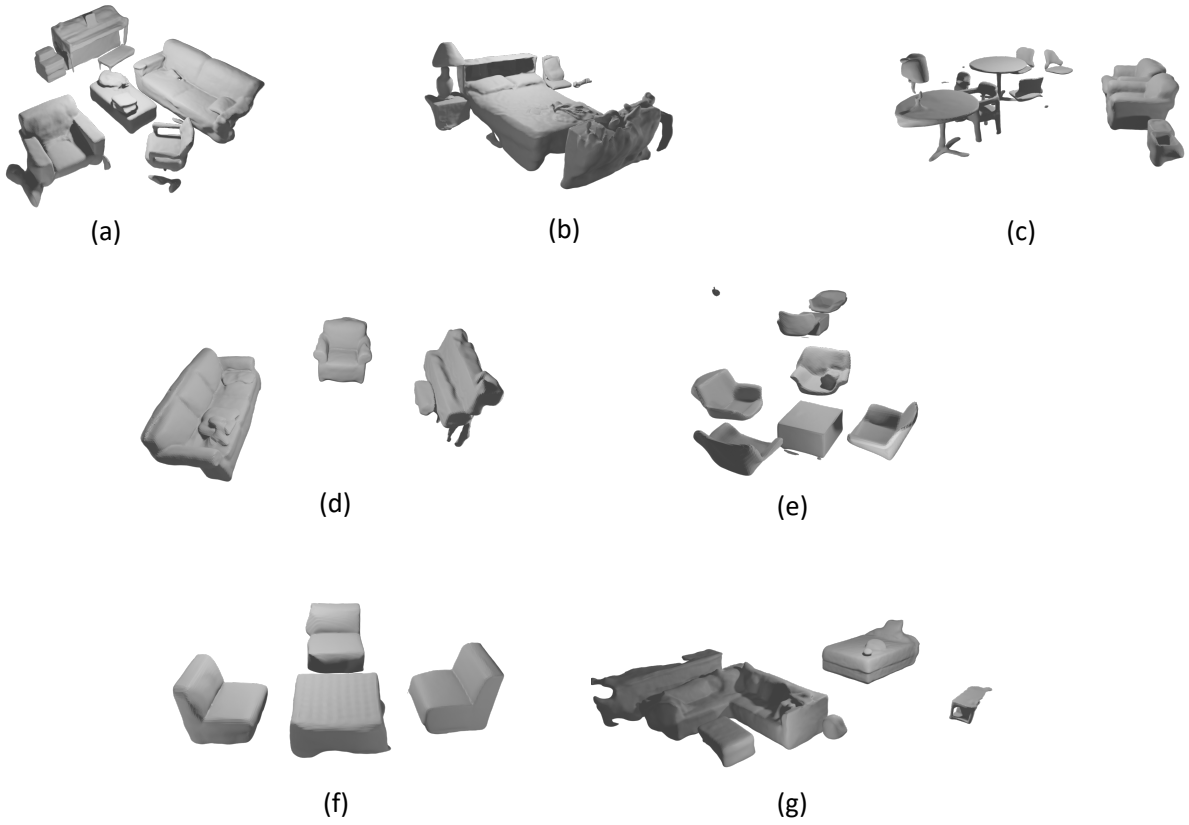

Figure 13. **Qualitative Visulization on ScanNet.** We show the object-compositional reconstruction results from RICO on seven ScanNet [3] scenes.

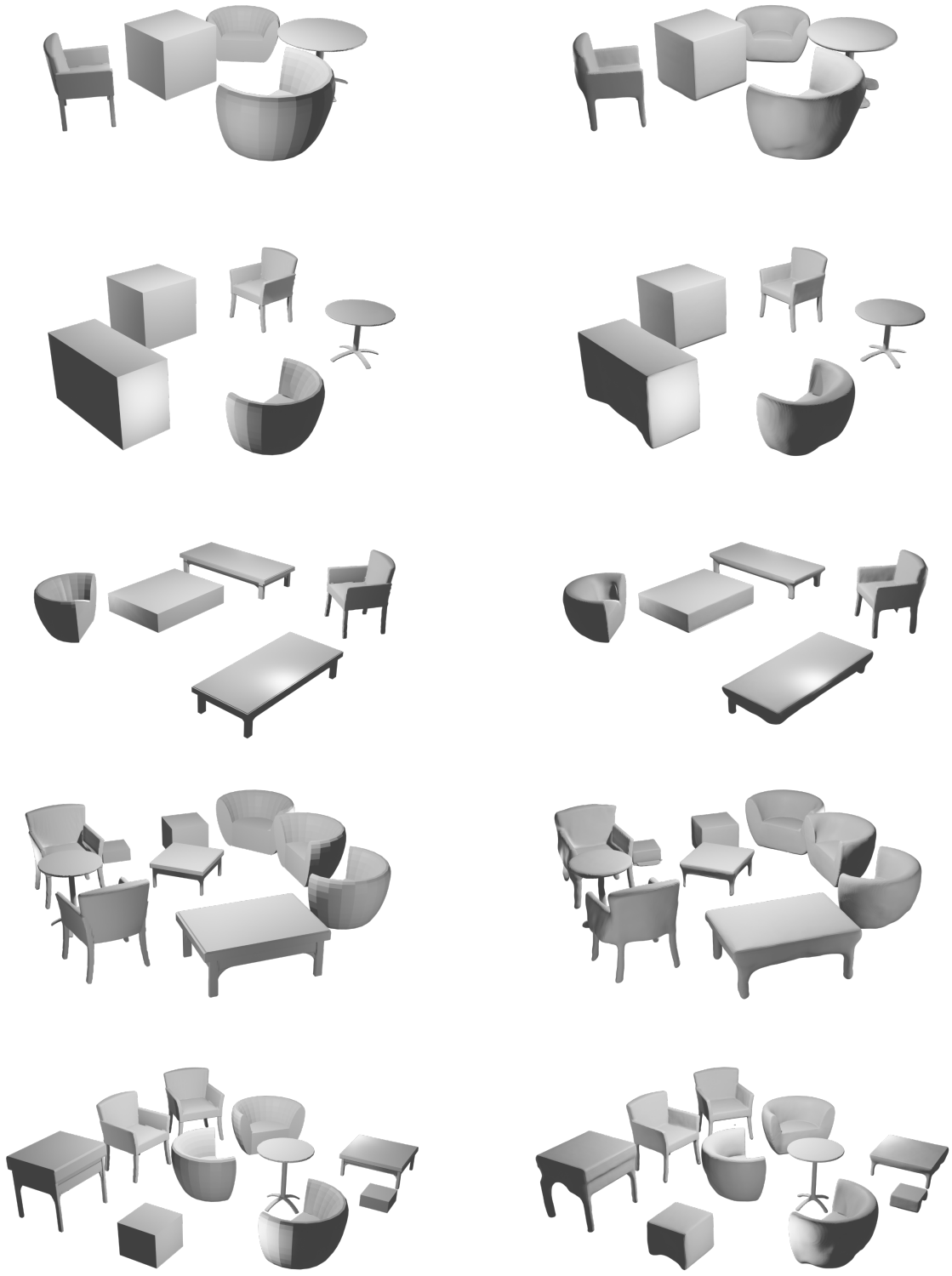

Figure 14. **Qualitative Visualization on Synthetic Scenes.** In the left column we show the ground-truth object geometry of the five synthetic scenes, in the right column we provide the qualitative object-compositional reconstruction results of our proposed RICO.