2023 IEEE/CVF International Conference on Computer Vision (ICCV) | 979-8-3503-0718-4/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICCV51070.2023.01628

RICO: Regularizing the Unobservable for Indoor Compositional Reconstruction

Zizhang Li[1], Xiaoyang Lyu[2], Yuanyuan Ding[1], Mengmeng Wang[1], Yiyi Liao[1]*, Yong Liu[1]*

[1] Zhejiang University, [2] The University of Hong Kong

Abstract

Recently, neural implicit surfaces have become popular for multi-view reconstruction. To facilitate practical applications like scene editing and manipulation, some works extend the framework with semantic masks input for the object-compositional reconstruction rather than the holistic perspective. Though achieving plausible disentanglement, the performance drops significantly when processing the indoor scenes where objects are usually partially observed. We propose RICO to address this by regularizing the unobservable regions for indoor compositional reconstruction. Our key idea is to first regularize the smoothness of the occluded background, which then in turn guides the foreground object reconstruction in unobservable regions based on the object-background relationship. Particularly, we regularize the geometry smoothness of occluded background patches. With the improved background surface, the signed distance function and the reversedly rendered depth of objects can be optimized to bound them within the background range. Extensive experiments show our method outperforms other methods on synthetic and real-world indoor scenes and prove the effectiveness of proposed regularizations. The code is available at https://github.com/kyleleey/RICO

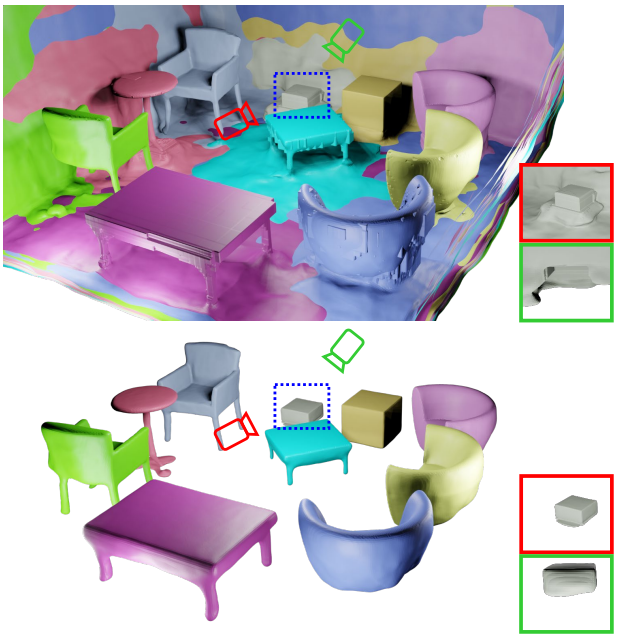

Figure 1. **Comparison** of [37] (Top) and ours (Bottom). Different objects are visualized in different colors. Previous reconstructions are connected with large artifacts. Taking the partially observed cubic object's front and back views (in red and green rectangles respectively) as an example, [37] can only get the visible surface, while ours can complement the complete body.

1. Introduction

Reconstructing 3D geometry from images is a fundamental problem in computer vision and has many downstream applications like VR/AR and game assets creation. With the advance of neural implicit representations [20], recent reconstruction methods [25, 36, 41, 43] can recover accurate geometry from multi-view images. However, existing methods typically regard the whole scene as an entirety and reconstruct everything altogether, thus preventing applications like scene editing. In indoor scenes with plenty of reconfigurable objects, a disentangled object-compositional reconstruction, i.e. decomposing the geometry into different instantiated objects and background, can be more suitable for further applications like moving the sofa in the scene.

In this paper, we aim to recover the room-level indoor scenes with decomposed geometry of individual objects and background. We assume that multi-view posed images and semantic masks that assign different labels to each instantiated object and the background are given as input. Existing object-compositional methods [9, 45, 40] concentrate more on the rendering performance rather than the underlying geometry, and thus can not be directly used for reconstruction. The most recent work ObjSDF [37] learns an object-compositional signed distance function (SDF) field by proposing a transform function between SDF values and semantic logits. Specifically speaking, ObjSDF predicts multiple SDF values at each 3D point for different seman-

*Corresponding authors.

2380-7504/23/$31.00 ©2023 IEEE
DOI 10.1109/ICCV51070.2023.01628

Authorized licensed use limited to: Zhejiang University. Downloaded on September 29,2024 at 07:14:50 UTC from IEEE Xplore. Restrictions apply.

tic classes, and converts them to semantic logits, allowing for separating object SDF values from the background when supervised by semantic masks. Although achieving plausible shape disentanglement, it suffers from a common problem in indoor scenes: objects and background can only be *partially observed due to occlusions*. When the object is partially observed, e.g. a cubic object against the wall, ObjSDF can not properly reconstruct the geometry between them (see Fig. 1 top-right). The reason is that the existing works [45, 37] can only effectively regularize semantic labels and geometry of observed regions, and have little impact on the unobserved regions. When processing the indoor scenes where a large portion of objects are partially observed, the reconstruction results of these objects will be visible surface connected with the unconstrained structures (as shown in Fig. 1, see Section 3.3 for a detailed analysis). Fig. 2 shows that even with reasonable reconstruction when composing all the objects together, each object's result in the unobserved region is far from satisfactory and can hinder further applications like manipulating the object.

We propose **RICO**, which realizes the proper geometry disentanglement for indoor scenes (see Fig. 1 bottom) by explicitly regularizing the unobserved regions. To be more specific, when the object is partially observed, recovering its geometry is an ill-posed problem even with corresponding masks. Thus, introducing prior regularization for unobserved regions is necessary. We exploit two types of prior knowledge for indoor scenes in this work: 1) background smoothness and 2) object-background relations. First, when one ray hits the object surface, the existing method [37] can properly regularize the geometry and appearance on the hitting point, but can not account for the background surface behind this object. This drawback leads to artifacts and holes on the unobserved background surface (see Fig. 2). We propose a patch-based smoothness loss to regularize the SDF values of unobserved background regions. Then, since the background reconstruction is improved, we can leverage another strong prior: *the objects are all within the room*, i.e. using the background surface to regularize the SDF field of objects. We design two regularization terms: an object point-SDF loss for sampled points behind the background surface and a reversed depth loss to regularize the SDF distribution of the entire ray. Both terms aim to bound the object within the background surface's range, thus preventing the aforementioned unmeaning structure, making the object reconstruction a *watertight and plausible* shape instead of an open surface with severe artifacts.

In summary, we propose RICO to realize compositional reconstruction in indoor scenes where a large portion of objects are partially observed. Our main contributions are: i) A patch-based background smoothness regularizer for unobserved background surface geometry. ii) Guided by the improved background surface, we exploit the object-background relation as prior knowledge and design objectives that effectively regularize objects' unobserved regions. iii) Extensive experiments on both real-world and synthetic datasets prove our superior reconstruction performance compared to previous works, especially for the partially observed objects.

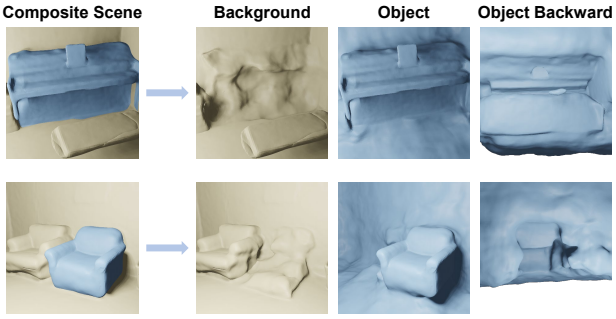

Figure 2. **ObjSDF results.** Interested objects are dyed in blue. Despite of the plausible composition, the disentangled backgrounds have artifacts and sunk holes, and partially observed objects can only get the visible surface (illustrated in 'Object Backward').

2. Related Work

Neural Implicit Representation for Reconstruction: Recently, neural implicit representations have emerged as a popular representation for 3D reconstruction. While early approaches [19, 26] rely on 3D supervision, a few works [24, 42] exploit surface rendering to use multi-view image supervision only, but also suffer from the unstable training. Neural radiance field (NeRF) [20] adopts volume rendering technique to achieve photorealistic scene representation and stable optimization. However, NeRF's formulation can not guarantee accurate underlying geometry. Therefore, [25, 36, 41] combine the geometry representation with iso-surface (e.g. occupancy [19], SDF [36, 41]) and volume density to accurately reconstruct object-level scenes from RGB images. [8] further applies the planar regularization for scene-level reconstruction. To tackle the problem in texture-less regions, [35, 43] utilize results from pretrained normal and depth estimation networks to guide the SDF training and boost the reconstruction performance.

However, despite the promising reconstruction performance, the aforementioned methods all consider the whole scene as an entirety. Our method focuses on decomposing the scene reconstruction into the background and different foreground objects, which can be regarded as compositional scene reconstruction.

Compositional Scene Reconstruction: Decomposing a scene into its different components could benefit downstream applications like scene editing. Many works have been proposed to recover the scene in a compositional man-

Authorized licensed use limited to: Zhejiang University. Downloaded on September 29,2024 at 07:14:50 UTC from IEEE Xplore. Restrictions apply.

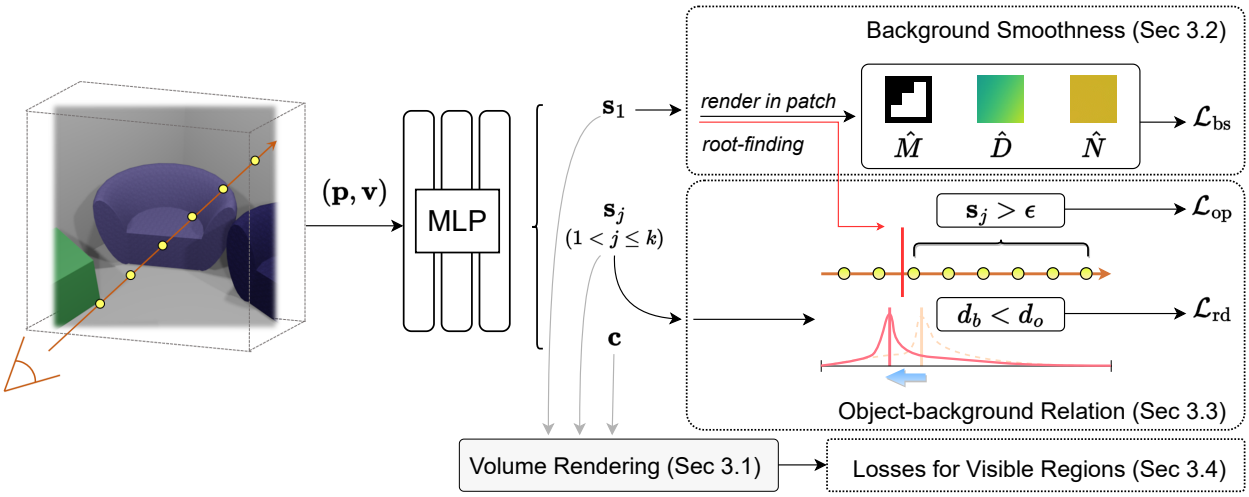

Figure 3. **Overview.** In this work, we propose two different regularizations. We first regularize the geometry smoothness of the unobserved background regions in a sampled patch. Then, we exploit the background surface as the prior to constrain the objects' surface. In detail, a per-point SDF loss and a reversed depth loss are introduced to regularize the manifold of objects' SDF functions. Combined with other reconstruction losses, our method reaches a neat and disentangled compositional reconstruction in indoor scenes.

ner from different perspectives. [22, 44, 10, 21] detect and reconstruct different objects in the given monocular image, and predict the scene's layout at the same time. But most of these methods require large-scale datasets with 3D ground truth for training. [14, 32, 18] optimize a feature field from a large pretrained model [1, 16], which enables deep feature based decomposition and manipulation. [38, 31] exploit the self-supervise paradigm to decompose the scenes into static and dynamic parts. More works [9, 45, 40, 39, 34, 6, 15, 33] focus on recovering the object-compositional scenes given semantic masks with images. However, this line of works concerns the rendering outputs rather than the geometry, i.e. the reconstruction results are sub-optimal.

Recently, ObjSDF [37] proposes a transform function between semantic logits and SDFs of different objects, which enables optimizing SDF fields with image and semantic mask supervision, and decomposing the whole scene with accurate reconstruction. However, in the indoor scenes where many objects are in partial observation, its reconstruction results are far from satisfactory and can not be used for further applications (see Fig. 2). On the contrary, our method introduces geometry prior to unobserved regions, yielding better compositional scene reconstruction.

Prior Regularization in Neural Implicit Reconstruction: In addition to the commonly used RGB image supervision, many different priors have been proposed to benefit the neural implicit representation. For example, [4, 30] use explicit point cloud as the depth prior, [23] employs regularization in unseen views and [11] introduces pretrained model's feature consistency. [12, 28] adopt explicit human model as the structural prior for the human-centric scenes. As for surface reconstruction, [8] proposes a manhattan-world assumption for the planar regions, [43] utilizes the normal and depth prediction from off-the-shelf model [5] as prior to regularize the texture-less regions.

Our method exploits the geometry prior in the unobserved region for compositional scene reconstruction. The proposed regularization can effectively reconstruct the objects and disentangle them from the background, even if the object itself is partially observed.

3. Methodology

Our goal is to recover decomposed geometry surface of the objects and background within a scene from the images and semantic masks inputs. To this end, we first review the SDF-based neural implicit representation and how to use the semantic logits for compositional reconstruction in Section 3.1. Next, we propose two types of regularizations on unobserved regions to address the partial observation problem: patch-based background smoothness (Section 3.2) and object-background relation (Section 3.3). Finally, we introduce the overall optimization procedure in Section 3.4. An overview of our method is provided in Fig 3.

3.1. Background

Volume Rendering of SDF-based Implicit Surface: For implicit reconstruction, the geometry of the scene is represented as the signed distance function (SDF) $s(\mathbf{p})$ of each spatial point $\mathbf{p}$, which is the point's distance to the closest surface. In practice, the SDF function is implemented as a multi-layer perceptron (MLP) network $f(\cdot)$. The appear-

Authorized licensed use limited to: Zhejiang University. Downloaded on September 29,2024 at 07:14:50 UTC from IEEE Xplore. Restrictions apply.

ance of the scene is also defined as an MLP $g(\cdot)$:

$$\begin{aligned} &f : \mathbf{p} \in \mathbb{R}^3 \mapsto (s \in \mathbb{R}, \mathbf{f} \in \mathbb{R}^{256}) \\ &g : (\mathbf{p} \in \mathbb{R}^3, \mathbf{n} \in \mathbb{R}^3, \mathbf{v} \in \mathbb{S}^2, \mathbf{f} \in \mathbb{R}^{256}) \mapsto \mathbf{c} \in \mathbb{R}^3 \end{aligned} \tag{1}$$

where $\mathbf{f}$ is a geometry feature vector, $\mathbf{n}$ is the normal at $\mathbf{p}$, $\mathbf{v}$ is the viewing direction and $\mathbf{c}$ is the view-dependent color. We adopt the unbiased rendering proposed in [36] to render the image. For each camera ray $\mathbf{r} = (\mathbf{o}, \mathbf{v})$ with $\mathbf{o}$ as the ray origin, n points $\{\mathbf{p}(t_i) = \mathbf{o} + t_i\mathbf{v} | i = 0, 1, \ldots, n-1\}$ are sampled, and the pixel color can be approximated as:

$$\hat{\mathbf{C}}(\mathbf{r}) = \sum_{i=0}^{n-1} T_i \alpha_i \mathbf{c}_i. \tag{2}$$

The T_i is the discrete accumulated transmittance derived from $T_i = \prod_{j=0}^{i-1}(1 - \alpha_j)$, and α_i is the discrete density value defined as

$$\alpha_i = \max\left(\frac{\Phi_u(s(\mathbf{p}(t_i))) - \Phi_u(s(\mathbf{p}(t_{i+1})))}{\Phi_u(s(\mathbf{p}(t_i)))}, 0\right), \tag{3}$$

where $\Phi_u(x) = (1 + e^{-ux})^{-1}$ and u is a learnable parameter. By minimizing the difference between predicted and ground-truth pixel colors, we can learn the SDF and appearance function of the desired scene.

Learning SDF with Semantic Logits: In this work, we consider compositional reconstruction of k objects given their masks. Note that we also consider the background as an instantiated object for brevity as in [37] and follow their network structure. In detail, for a scene with k objects, the SDF MLP $f(\cdot)$ now outputs k SDFs at each point, and the j-th ($1 \leq j \leq k$) SDF represents the geometry of j-th object. Without loss of generality, we set $j = 1$ as the background category and others for objects in Fig. 3 and the rest of the paper. The *scene* SDF is the minimum of $\{s_j\}$, which is used for sampling points along the ray and aforementioned volume rendering (Eq. 2). Moreover, each point's k SDFs can be transformed into semantic logits $\mathbf{h}(\mathbf{p})$ as

$$\begin{aligned} h_j(\mathbf{p}) &= \gamma / (1 + \exp(\gamma \cdot s_j(\mathbf{p}))), \\ \mathbf{h}(\mathbf{p}) &= [h_1(\mathbf{p}), h_2(\mathbf{p}), ..., h_k(\mathbf{p})], \end{aligned} \tag{4}$$

where γ is a fixed parameter. Using volume rendering to accumulate the semantic logits of all the points along a ray, we can get the semantic logits $\mathbf{H}(\mathbf{r}) \in \mathbb{R}^k$ of each pixel. During training, the cross-entropy loss applied to $\mathbf{H}(\mathbf{r})$ is backpropagated to the SDF values, allowing for learning the compositional geometry.

3.2. Patch-based Background Smoothness

Although the volume rendering can propagate gradients along the entire ray, the optimization mainly focuses on the surface-hitting point, as its accumulated weight can be much larger than the others. As a result, the geometry of the points behind the first-hit surface can not be optimized correctly. In the indoor scenes, the occluded part of the background surface is invisible in all the images, which can be with holes and random artifacts (see Fig. 2).

Since we cannot tell the exact color, depth or normal of the occluded part, it is intractable to optimize this region wrt. its ground truth. Therefore, we propose to regularize the geometry of the occluded background to be *smooth*, thus preventing some clearly wrong artifacts.

In detail, we regularize the smoothness of rendered depth and normal of background surface within a small patch region. To save the computation budget, we randomly sample a $\mathcal{P} \times \mathcal{P}$ size patch every $\mathcal{T}_\mathcal{P}$ iterations in the given image and sample points along the patch rays using the *background SDF* only. We compute depth map $\hat{D}(\mathbf{r})$ and normal map $\hat{N}(\mathbf{r})$ of the sampled patch following [43], $\mathbf{r}$ denotes the sampled ray in patch. The semantic map of the patch is also computed and transformed into a patch Mask $\hat{M}(\mathbf{r})$:

$$\hat{M}(\mathbf{r}) = \mathbb{1}[\arg\max(\mathbf{H}(\mathbf{r})) \neq 1], \tag{5}$$

which means the mask value is 1 if the rendered class is not the background, so that only the occluded background is regularized. Taking rendered depth as an example, the patch-based background smoothness loss is

$$\begin{aligned} \mathcal{L}(\hat{D}) = \sum_{d=0}^{3} \sum_{m,n=0}^{\mathcal{P}-1-2^d} &\hat{M}(\mathbf{r}_{m,n}) \odot (|\hat{D}(\mathbf{r}_{m,n}) - \\ &\hat{D}(\mathbf{r}_{m,n+2^d})| + |\hat{D}(\mathbf{r}_{m,n}) - \hat{D}(\mathbf{r}_{m+2^d,n})|). \end{aligned} \tag{6}$$

Here the smoothness is applied on different intervals controlled by d. m and n are the pixel indices within the patch and the mask is multiplied at each position with hadamard product $\odot$. The normal smoothness loss $\mathcal{L}(\hat{N})$ can be obtained similarly. We define the overall background smoothness loss $\mathcal{L}_{bs}$ as:

$$\mathcal{L}_{\text{bs}} = \mathcal{L}(\hat{D}) + \mathcal{L}(\hat{N}). \tag{7}$$

Here in contrast to [23] which applies a patch-based regularization to visible regions in other views, we instead regularize the *occluded* regions of the background.

3.3. Object-background Relation as Regularization

With the help of the patch-based background smoothness loss, most artifacts of the background are resolved, yielding a smooth surface. We further leverage this smooth background surface to regularize the SDF fields of other objects, leveraging a key prior knowledge that *all the other objects are confined to the room*, i.e., the background surface.

In the original framework of [37], if an object is partially observed, e.g. against the background, the object's reconstructed surface won't be a "closed" surface (see Fig. 1 and

Authorized licensed use limited to: Zhejiang University. Downloaded on September 29,2024 at 07:14:50 UTC from IEEE Xplore. Restrictions apply.

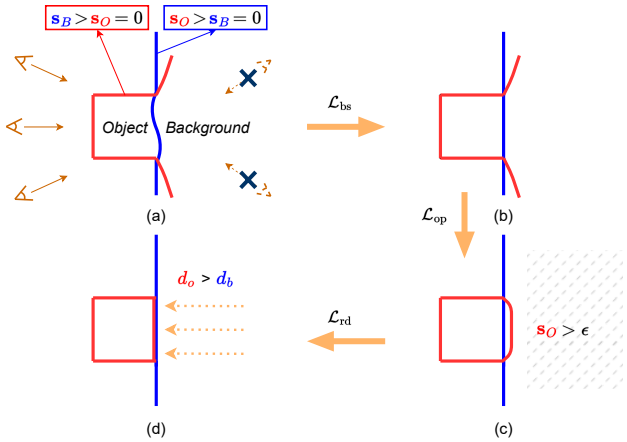

Figure 4. **Toy-case analysis.** A bird-eye view of an object against the background is shown here, the SDF and surface of object is shown in red and background's in blue. (a) In [37] the minimum SDF is used for volume rendering and semantic loss, so on object's visible surface where object SDF is optimized to 0 and the background SDF here is positive, and similar for the visible background surface. Since the scene is partially observed from left, the right part of object surface is open and unobserved background region is unconstraint. (b) With the smoothness prior the background surface can be plausible, (c) and object point SDF loss close the object surface but still have intersections, (d) finally the reversed depth loss optimizes the entire ray thus the object can be within the background.

Fig. 2). We refer to the toy-case analysis shown in Fig. 4 for the reason of the current unsatisfactory reconstruction. To encourage the reconstruction to be *watertight*, we design two types of regularization on the SDF fields of objects.

Object Point-SDF Loss: A straightforward solution is to directly regularize the objects' SDFs on every sampled point behind the background surface. To be more specific, the regular volume rendering only guarantees a single change of the sign (positive SDF to negative) when a ray hits a visible surface. For watertight objects, the ray should hit another occluded surface (negative SDF to positive). With the prior that the object is confined within the room, the occluded object surface should be closer than the background surface, meaning that the object SDFs of points behind the background surface should all be positive (Fig. 4 (c)).

We implement an object point-SDF loss based on the above analysis. For the sampled points along the rays, we first utilize the root-finding algorithm among the background SDF of these points and find the zero-SDF ray depth t'. Then the object point-SDF loss can be formulated as

$$\mathcal{L}_{\text{op}} = \frac{1}{k-1}\sum_{j=2}^{k} \max\left(0, \epsilon - \mathbf{s}_j(\mathbf{p}(t_i))\right) \cdot \mathbb{1}[t_i > t'], \quad (8)$$

which pushes the objects' SDFs at points behind the surface larger than a positive threshold ϵ.

Reversed Depth Loss: Although the $\mathcal{L}_{\text{op}}$ can effectively regularize the SDF fields of each object, in practice we find the reconstructed object surface can still have intersections with the background surface (Fig. 4 (c)). The reason is that the sampled points are discrete and in most cases, the background surface is between two sampled points. Therefore, the sign change of the occluded surface may still occur after hitting the background surface.

Since per-point optimization can not effectively propagate to the distribution of the entire ray. To optimize the entire ray's SDF distribution for better regularization, we compute a *reversed depth* along each ray. With the help of $\mathcal{L}_{\text{op}}$, the sign of the object SDF along one ray now is positive-negative-positive, which enables rendering a depth value backward. We first transform the ray depth $\{t_i | i = 0, 1, \ldots, n-1\}$ into the reversed ray depth named $\{\hat{t}_i | i = 0, 1, \ldots, n-1\}$, where

$$\hat{t}_i = (t_0 + t_{n-1}) - t_{n-1-i}. \quad (9)$$

With the reversed ray depth values, we use the background and object SDF both in reversed order to compute the accumulated weight and get the reversed depth respectively. Remarkably, in order to compute the exact correct depth, the points should be re-sampled along the reverse direction. Here we directly use the sampled points to avoid computation overhead, and empirical results prove its effectiveness. We only compute the reverse depth of one pixel if satisfying two conditions: 1) this pixel's $\hat{M}(\mathbf{r}) = 1$; 2) the SDF value of the rendered object at the furthest point is positive. Note that the second condition is usually satisfied when the object point-SDF loss is applied. By computing the reversed depth d_o of the hitting object (determined by the pixel's rendered semantic) and d_b of the background, we can get the reversed depth loss:

$$\mathcal{L}_{\text{rd}} = \max(0, d_b - d_o), \quad (10)$$

which pushes the object surface within the background as illustrated in Fig. 4 (d).

3.4. Training Objectives Details

Monocular geometric cues are essential for indoor scene reconstruction as proved in [43]. Following [43], we add the depth and normal consistency loss ($\mathcal{L}_{\text{D}}$,$\mathcal{L}_{\text{N}}$) with pseudo ground truth from Omnidata [5] model. In experiments, we also add the monocular cues on [37] to get a stronger baseline for a fair comparison.

The SDF network is also regularized by an Eikonal [7] loss item $\mathcal{L}_{\text{E}}$. We further use the semantic loss $\mathcal{L}_{\text{S}}$ proposed in [37] to learn the compositional geometry. The overall loss function for compositional reconstruction is:

$$\begin{aligned}\mathcal{L} =& \mathcal{L}_{\text{RGB}} + \lambda_{\text{D}}\mathcal{L}_{\text{D}} + \lambda_{\text{N}}\mathcal{L}_{\text{N}} + \lambda_{\text{E}}\mathcal{L}_{\text{E}} + \lambda_{\text{S}}\mathcal{L}_{\text{S}} \\ &+ \lambda_{\text{bs}}\mathcal{L}_{\text{bs}} + \lambda_{\text{op}}\mathcal{L}_{\text{op}} + \lambda_{\text{rd}}\mathcal{L}_{\text{rd}}.\end{aligned} \quad (11)$$

Authorized licensed use limited to: Zhejiang University. Downloaded on September 29,2024 at 07:14:50 UTC from IEEE Xplore. Restrictions apply.

We set $\lambda_{\text{bs}}, \lambda_{\text{op}}, \lambda_{\text{rd}} = 0.1$ in our experiments. For other loss terms, given the weight of RGB reconstruction loss as 1, we set $\lambda_{\text{D}} = 0.1$ and $\lambda_{\text{N}} = \lambda_{\text{E}} = 0.05$ following [43]. For the semantic loss, we follow the implementation in [37] and set $\lambda_{\text{D}} = 0.04$. The detailed calculation of other losses can be found in supplementary.

4. Experiments

We first conduct ablation study on each proposed component. Then we provide quantitative and qualitative results of object-compositional reconstruction on real-world and synthetic scenes. Finally, we show some possible object manipulation with our compositional reconstruction.

Datasets: We consider three types of indoor datasets with multi-view RGB images and masks: 1) ScanNet [3], a real-world dataset widely used in previous works [8, 35, 37, 43]; 2) ToyDesk dataset, a real-world dataset with two scenes and each has four objects placed on the table plane. This dataset has also been widely used in previous compositional works [40, 37] 3) a hand-crafted synthetic dataset with five scenes, each containing $5 \sim 10$ objects. The synthetic dataset is considered such that the ground truth geometry of both, occluded and non-occluded regions are available.

Metrics: For reconstruction performance, we report Chamfer Distance, F-score for evaluation on ScanNet, following [8, 43]. On synthetic scenes we divide metrics for two aspects: objects and background. For objects we report the reconstruction performance compared to the complete ground truth object mesh, and the final results are averaged across all the objects in all the scenes. For background we report the reconstruction metrics and rendered depth errors of the occluded regions only to highlight the effectiveness of our regularization. The results are also averaged across all the scenes. See supplementary for a detailed introduction to datasets and metrics. We also report two 2D metrics, PSNR and mIoU, on the ToyDesk and ScanNet scenes. Here we omit the synthetic dataset considering its relatively simple texture. For the test image set, on ToyDesk dataset we use the official split, and on ScanNet we sample frames from their original camera trajectories. The metrics are averaged across different images of different scenes.

Baselines: We mainly compare with **ObjSDF** [37] in this work, as it is the only method that focuses on the same task, a specific discussion is provided in the supplementary. Since we focus on the indoor scene where monocular cues can benefit a lot [43], we add losses proposed in [43] to our method for better performance. For a fair comparison, we also combine [37] and [43] for a stronger baseline, named **ObjSDF***. Moreover, since ObjSDF*'s object reconstructions inevitably have artifacts, we develop a post-process method to cull the parts outside the background range and name this baseline **ObjSDF*-C**. We provide the details of how to set the range in the supplementary. Note that our method directly generates clean watertight meshes and doesn't need such post-processing. In ScanNet experiments, we also report results for MonoSDF [43] as a reference since we can only evaluate the overall reconstruction (details in Section 4.3).

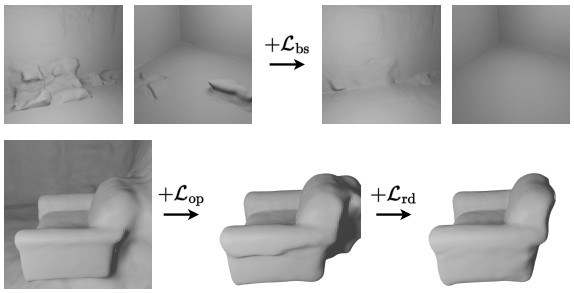

Figure 5. **Effects of Regularizations.** The first row shows two backgrounds, the $\mathcal{L}_{\text{bs}}$ clearly mitigates most of the artifacts. The last row shows that with $\mathcal{L}_{\text{op}}$ we can get the watertight object and $\mathcal{L}_{\text{rd}}$ further constrains the unobserved surface.

Implementation Details: We implement our method in PyTorch [27]. We use the Adam [13] optimizer with a learning rate of 5e-4 for 50k iterations and sample 1024 rays per iteration. The weight initialization scheme for SDF MLP is identical to [41, 36, 37]. The u is initialized as 0.05 and we set γ=20 as proposed in [37]. $\mathcal{P}$, $\mathcal{T}_{\mathcal{P}}$ and ϵ are set as 32, 10 and 0.05 respectively. All the reconstructions are acquired by using marching cube [17] at the resolution of 512.

Moreover, we adopt the geometric initialization [7] for the geometry network, which initializes the reconstruction with a unit sphere and the surface normals are facing inside at the beginning of the optimization. For the ScanNet dataset, we follow the protocol in [43] to crop the input image to 384×384 and adjust the intrinsics accordingly. For the synthetic dataset, we directly render the image at the same resolution. Since we focus on the indoor scenes in this paper, we adopt the common practice to process the "bounded" scene. For each scene, we normalize the camera poses so that all the cameras lie in a unit sphere. The rays intersection, i.e. the furthest sampling location, is computed based on this sphere and we also conduct Marching Cubes [17] within the same area for the final reconstruction.

4.1. Ablation Study

We first quantitatively analyze the effectiveness of the proposed regularizations on synthetic scenes by comparing our full method to four variants in Tab. 1. First, adding the smoothness loss (V2) significantly decreases the depth error of occluded background regions. Next, the $\mathcal{L}_{\text{op}}$ makes the object reconstruction watertight and the object-level metrics are greatly improved (V3). The $\mathcal{L}_{\text{rd}}$ further improves the object reconstruction performance (Full). Results of V4 prove

Authorized licensed use limited to: Zhejiang University. Downloaded on September 29,2024 at 07:14:50 UTC from IEEE Xplore. Restrictions apply.

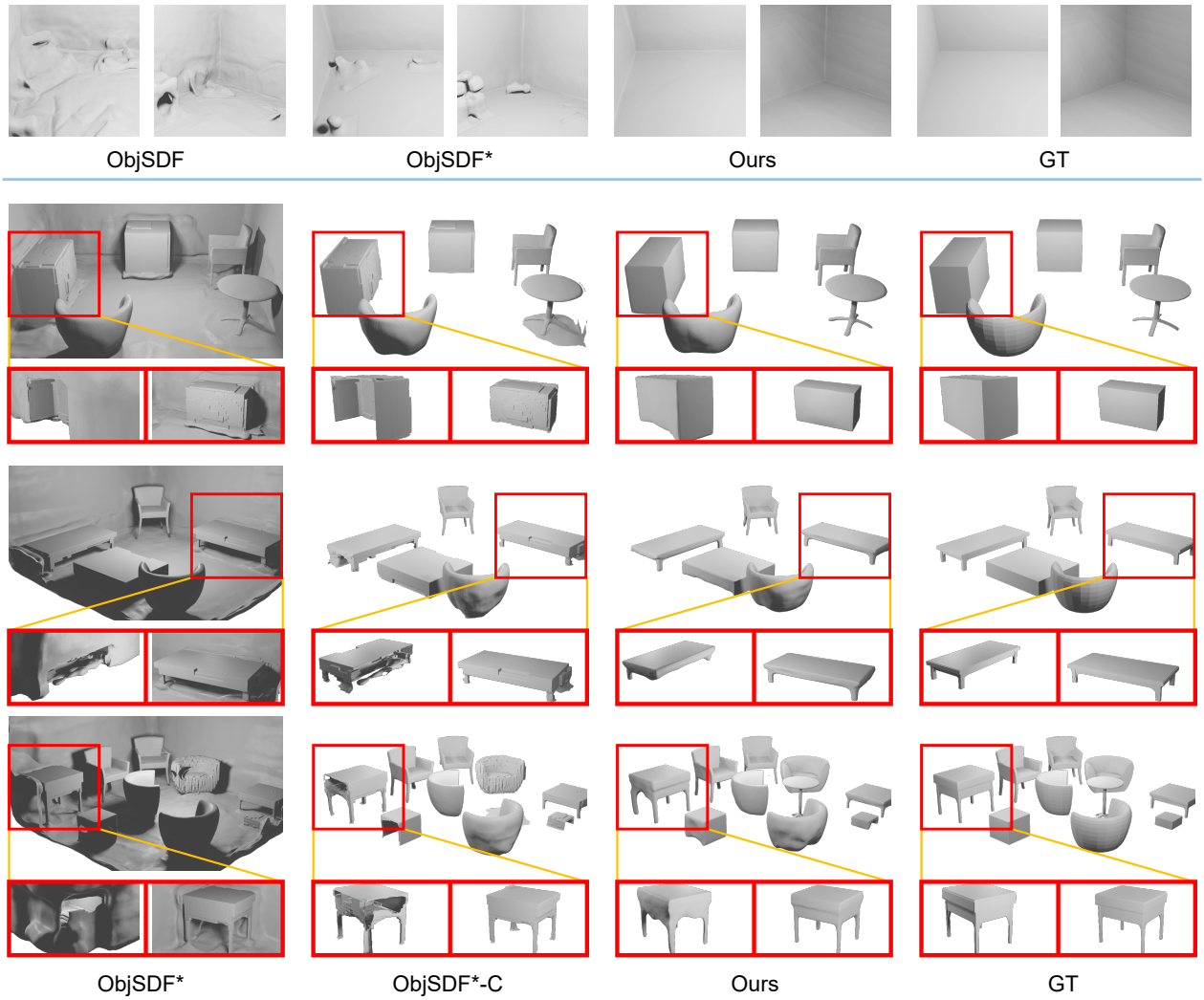

Figure 6. **Qualitative Results on Synthetic Scenes.** Above the blue line we show the comparison of different methods on two background scenes. Below we provide results of two scenes where only the object results are shown. In red rectangles at the bottom of each picture, we show the back (left part) and front (right part) views of an example object. Detailed descriptions in Section 4.2.

				Background			Object	
	$\mathcal{L}_{bs}$	$\mathcal{L}_{op}$	$\mathcal{L}_{rd}$	D.↓	C.↓	F.↑	C.↓	F.↑
V1	-	-	-	0.021	0.045	0.768	1.24	0.065
V2	✓	-	-	0.004	0.009	0.99	1.25	0.066
V3	✓	✓	-	0.003	0.009	0.99	0.045	0.735
V4	-	✓	✓	0.015	0.029	0.888	0.039	0.785
Full	✓	✓	✓	**0.003**	**0.009**	**0.99**	**0.033**	**0.817**

Table 1. **Ablation Study.** The D. denotes the depth error of the occluded background. The C. and F. mean the Chamfer-L_1 and F-score respectively, for the reconstruction of occluded background regions and full complete objects. Details in Section 4.1.

	Background			Objects	
	D.↓	C.↓	F.↑	C.↓	F.↑
ObjSDF	0.033	0.076	0.669	1.040	0.084
ObjSDF*	0.021	0.044	0.772	1.030	0.085
ObjSDF*-C	0.021	0.044	0.772	0.134	0.681
Ours	**0.003**	**0.009**	**0.990**	**0.033**	**0.817**

Table 2. **Synthetic Scenes Reconstruction Quantitative Results.** The D. C. and F. have the same meanings as in Tab. 1.

that without $\mathcal{L}_{bs}$ to improve the background, the object reconstruction quality also drops. Note that $\mathcal{L}_{rd}$ needs to be applied jointly with $\mathcal{L}_{op}$ to guarantee its second condition.

A qualitative case is shown in Fig. 5 for better illustration. In the first row, the occluded part of the background

Authorized licensed use limited to: Zhejiang University. Downloaded on September 29,2024 at 07:14:50 UTC from IEEE Xplore. Restrictions apply.

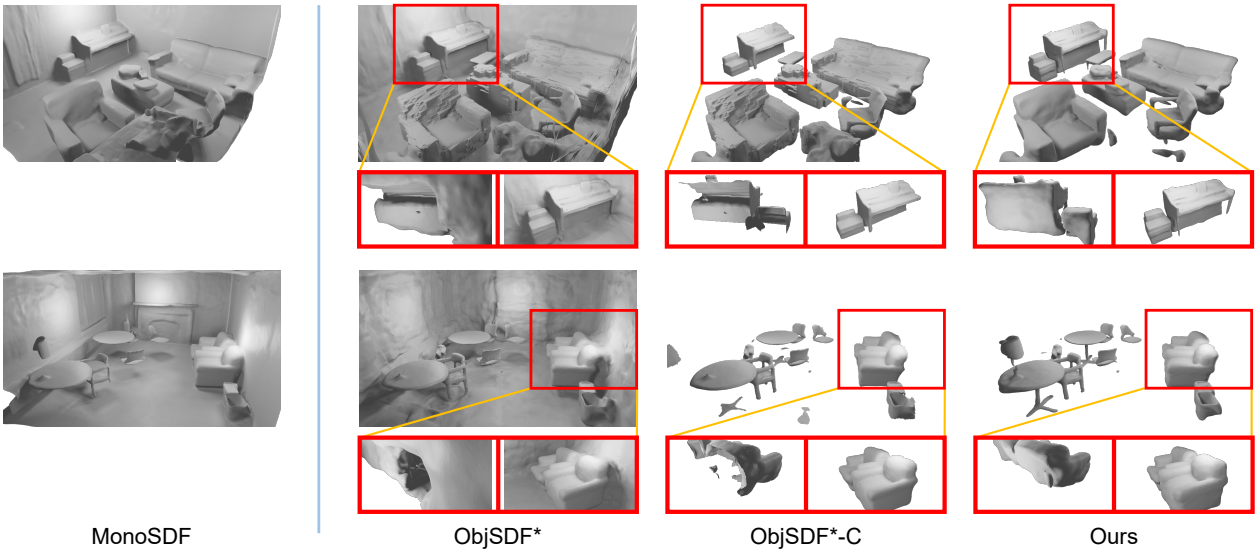

Figure 7. **Qualitative Results on ScanNet [3].** On the left of blue line we show the overall reconstruction from [43] as the reference. On the right we show the comparison between our method and two baselines. Similarly, the back and front views of objects with partial observations are provided in red rectangles. Detailed descriptions in Section 4.3.

can have many artifacts because of occlusion. With the smoothness regularization, the reconstructed backgrounds are much cleaner. Since the $\mathcal{L}_{\text{bs}}$ only applies to the occluded region, the other parts won't be affected. In the second row, the sofa is pushed against the wall and reconstruction from [37] contains only visible open surface. The $\mathcal{L}_{\text{op}}$ first regularizes the SDF field to be a watertight mesh. The $\mathcal{L}_{\text{rd}}$ further mitigates the flaws at the back by regularizing the object to be confined to the background.

4.2. Reconstruction in Synthetic Scenes

Since the synthetic scenes are rendered with objects that have accurate object-level 3D ground truth geometry, we compare our methods with previous object-compositional reconstruction method [37] on the 5 generated scenes and report the quantitative results in Tab. 2.

In Tab. 2 we provide detailed metrics on object-level reconstruction. A qualitative comparison is shown in Fig. 6. The computation procedure of all the metrics is available in the supplementary. With monocular cues, ObjSDF* can achieve better performance than the original ObjSDF. However, their performance on object reconstruction is far from satisfactory, since they can only accurately obtain the visible surface with large parts of irrelevant structures in the indoor scenes as pointed out in the analysis before. Though ObjSDF*-C can eliminate most of the outliers and get improved performance, it can only reconstruct visible regions as an open surface as shown in Fig. 6. On the contrary, our regularizations help to smoothen the background and recover the object as a watertight mesh, which promotes performance by a large margin and leads to broader downstream applications. As shown in Fig. 6, RICO can reconstruct the unobservable (back view as in the caption) regions of the objects where the previous methods fail.

	ObjSDF	ObjSDF*	Ours	MonoSDF
Chamfer-$L_1 \downarrow$	0.170	0.092	**0.088**	0.090
F-Score $\uparrow$	0.357	0.567	**0.624**	0.610

Table 3. **ScanNet Reconstruction Quantitative Results.**

4.3. Reconstruction in Real-world Scenes

We conduct experiments on 7 scenes of ScanNet [3] to show the effectiveness of our method on real-world data. Since ScanNet only provides ground truth for the visible surface, here we follow the protocol in [8] and report the reconstruction performance of the entire scene in Tab. 3. By utilizing the rendering formulation in [36] which explicitly models the angle between surface normal and ray, our method can achieve slightly better performance compared to ObjSDF* and [43] on visible surfaces.

In Tab. 4 we evaluate the 2D rendering and segmentation performance and provide the PSNR and mIoU results on two real-world datasets. As our method constrains mainly the unobservable, the PSNR and mIoU metrics are not with significant difference.

In Fig. 7 we also provide the comparisons over two scenes. Note that the ScanNet images are blurry and noisy, and the masks are sometimes inaccurate (e.g., legs of chairs

Authorized licensed use limited to: Zhejiang University. Downloaded on September 29,2024 at 07:14:50 UTC from IEEE Xplore. Restrictions apply.

		ObjSDF	ObjSDF*	Ours
PSNR↑	ToyDesk	21.10	21.21	**21.33**
	ScanNet	22.78	23.15	**23.30**
mIoU↑	ToyDesk	0.88	**0.89**	**0.89**
	ScanNet	0.89	0.91	**0.92**

Table 4. **PSNR and mIoU Results** of different methods on scenes from ScanNet and ToyDesk datasets.

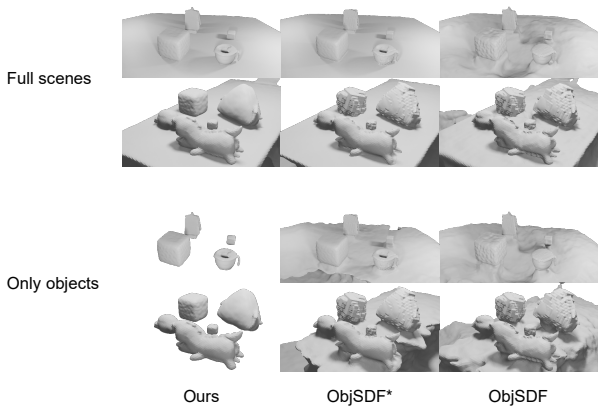

Figure 8. **Qualitative Results on ToyDesk.** Here we show the reconstruction results both with and without the background for better visualization.

missing), leading to less accurate reconstructions. However, it can still be seen that our method can successfully regularize the unobservable regions to get the watertight mesh, while the baselines always result in open surface. For example, the piano in the first row and the sofas in the second row can only be reconstructed on visible surfaces.

And in Fig. 8 we provide the qualitative results of our method and two baselines on two ToyDesk scenes. It can be seen that the baseline methods generate non-watertight surface with undesired artifacts (using table plane to cut the mesh will result in holes on the cut face), while our method can directly get clean results.

4.4. Object Manipulation

As aforementioned, the reconstruction results from ObjSDF* are sub-optimal for downstream applications like object manipulation. On the contrary, since RICO can get a watertight mesh, it can be easily used for such applications. In Fig. 9 we show the volume rendered normal maps and segmentation masks before and after moving an object in the scene. An illustration of how to manipulate one object in our framework and conduct volume rendering accordingly is applicable in the supplementary. As illustrated, after manipulation, the rendered segmentation and normal map are

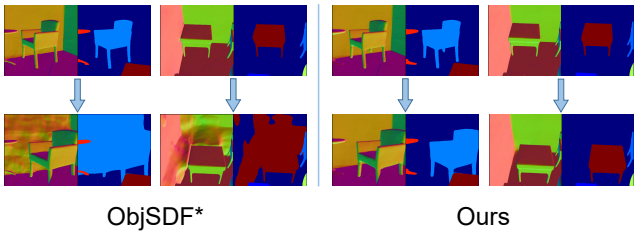

Figure 9. **Object Manipulation Cases.** Here we show the volume rendered normal maps and segmentation masks in two scenes, before (top row) and after (bottom row) the manipulation.

still clean for RICO. In contrast, the results of ObjSDF* are messy because their reconstruction is connected with artifacts that were originally outside of the scenes.

5. Conclusion

We have presented RICO, a novel approach for compositional reconstruction in indoor scenes. Our key motivation is to regularize the unobservable regions for the objects with partial observations in indoor scenes. We exploit the geometry smoothness for the occluded background, and then adopt the improved background as the prior to regularize the objects' geometry. Our experimental results prove that our method achieves compositional reconstruction with fewer artifacts and watertight object geometry, which further facilitates applicable applications like object manipulation.

Future work: Currently, each object's pose is baked with its geometry. An interesting future direction would be represent each object in scene as the combination of its own canonical coordinates and the pose, i.e. the rotated 3D bounding box in the world coordinate. With these disentangled representations, we can cast more regularizations from the shape and box perspectives, and the disentanglement also allows more flexible downstream applications. Another interesting direction would be developing more semantic-based prior knowledge for systemic regularization. Now the geometry motivated regularizations are lack of the understanding of the actual shape of each object. With the general knowledge emerged in recent large vision-language model, like StableDiffusion, it's interesting to distill its knowledge in our compositional reconstruction task. The learnt semantic knowledge can provide a holistic understanding of each object and provide a more comprehensive regularization.

Acknowledgements: We thank Kechun Xu and Huaijin Pi for detailed feedback on the paper. We thank all authors and reviewers for the contributions. This work is in part supported by a Grant from the National Natural Science Foundation of China (No.U21A20484), the NSFC under grant U21B2004, 62202418, and the Zhejiang University Education Foundation Qizhen Scholar Foundation.

Authorized licensed use limited to: Zhejiang University. Downloaded on September 29,2024 at 07:14:50 UTC from IEEE Xplore. Restrictions apply.

References

[1] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proc. of the IEEE International Conf. on Computer Vision (ICCV)*, 2021. 3

[2] Blender Online Community. *Blender – a 3D modelling and rendering package*. Blender Foundation, Stichting Blender Foundation, Amsterdam, 2018. 13

[3] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2017. 6, 8, 15

[4] Kangle Deng, Andrew Liu, Jun-Yan Zhu, and Deva Ramanan. Depth-supervised nerf: Fewer views and faster training for free. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2022. 3

[5] 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 *Proc. of the IEEE International Conf. on Computer Vision (ICCV)*, 2021. 3, 5, 12, 13

[6] Xiao Fu, Shangzhan Zhang, Tianrun Chen, Yichong Lu, Lanyun Zhu, Xiaowei Zhou, Andreas Geiger, and Yiyi Liao. Panoptic nerf: 3d-to-2d label transfer for panoptic urban scene segmentation. In *Proc. of the International Conf. on 3D Vision (3DV)*, 2022. 3

[7] Amos Gropp, Lior Yariv, Niv Haim, Matan Atzmon, and Yaron Lipman. Implicit geometric regularization for learning shapes. *Proc. of the International Conf. on Machine learning (ICML)*, 2020. 5, 6

[8] 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 *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2, 3, 6, 8

[9] Michelle Guo, Alireza Fathi, Jiajun Wu, and Thomas Funkhouser. Object-centric neural scene rendering. *arXiv.org*, 2020. 1, 3

[10] Muhammad Zubair Irshad, Sergey Zakharov, Rares Ambrus, Thomas Kollar, Zsolt Kira, and Adrien Gaidon. Shapo: Implicit representations for multi-object shape, appearance, and pose optimization. In *Proc. of the European Conf. on Computer Vision (ECCV)*, 2022. 3

[11] Ajay Jain, Matthew Tancik, and Pieter Abbeel. Putting nerf on a diet: Semantically consistent few-shot view synthesis. In *Proc. of the IEEE International Conf. on Computer Vision (ICCV)*, 2021. 3

[12] Wei Jiang, Kwang Moo Yi, Golnoosh Samei, Oncel Tuzel, and Anurag Ranjan. Neuman: Neural human radiance field from a single video. *Proc. of the European Conf. on Computer Vision (ECCV)*, 2022. 3

[13] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv.org*, 2014. 6

[14] Sosuke Kobayashi, Eiichi Matsumoto, and Vincent Sitzmann. Decomposing nerf for editing via feature field distillation. *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 3

[15] Abhijit Kundu, Kyle Genova, Xiaoqi Yin, Alireza Fathi, Caroline Pantofaru, Leonidas J Guibas, Andrea Tagliasacchi, Frank Dellaert, and Thomas Funkhouser. Panoptic neural fields: A semantic object-aware neural scene representation. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2022. 3

[16] Boyi Li, Kilian Q Weinberger, Serge Belongie, Vladlen Koltun, and René Ranftl. Language-driven semantic segmentation. *Proc. of the International Conf. on Learning Representations (ICLR)*, 2022. 3

[17] William E Lorensen and Harvey E Cline. Marching cubes: A high resolution 3d surface construction algorithm. *ACM siggraph computer graphics*, 21(4):163–169, 1987. 6

[18] Kirill Mazur, Edgar Sucar, and Andrew J Davison. Feature-realistic neural fusion for real-time, open set scene understanding. *Proc. IEEE International Conf. on Robotics and Automation (ICRA)*, 2023. 3

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

[20] 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. *Proc. of the European Conf. on Computer Vision (ECCV)*, 2020. 1, 2, 12

[21] Yinyu Nie, Angela Dai, Xiaoguang Han, and Matthias Nießner. Learning 3d scene priors with 2d supervision. *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2023. 3

[22] 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 *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2020. 3

[23] Michael Niemeyer, Jonathan T Barron, Ben Mildenhall, Mehdi SM Sajjadi, Andreas Geiger, and Noha Radwan. Regnerf: Regularizing neural radiance fields for view synthesis from sparse inputs. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2022. 3, 4

[24] Michael Niemeyer, Lars Mescheder, Michael Oechsle, and Andreas Geiger. Differentiable volumetric rendering: Learning implicit 3d representations without 3d supervision. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[25] Michael Oechsle, Songyou Peng, and Andreas Geiger. Unisurf: Unifying neural implicit surfaces and radiance fields for multi-view reconstruction. In *Proc. of the IEEE International Conf. on Computer Vision (ICCV)*, 2021. 1, 2

[26] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

Authorized licensed use limited to: Zhejiang University. Downloaded on September 29,2024 at 07:14:50 UTC from IEEE Xplore. Restrictions apply.

[27] 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. *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. 6

[28] Georgios Pavlakos, Ethan Weber, Matthew Tancik, and Angjoo Kanazawa. The one where they reconstructed 3d humans and environments in tv shows. *Proc. of the European Conf. on Computer Vision (ECCV)*, 2022. 3

[29] Maxime Raafat. BlenderNeRF. https://github.com/maximeraafat/BlenderNeRF, 2 2023. 13

[30] Barbara Roessle, Jonathan T Barron, Ben Mildenhall, Pratul P Srinivasan, and Matthias Nießner. Dense depth priors for neural radiance fields from sparse input views. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2022. 3

[31] Prafull Sharma, Ayush Tewari, Yilun Du, Sergey Zakharov, Rares Ambrus, Adrien Gaidon, William T Freeman, Fredo Durand, Joshua B Tenenbaum, and Vincent Sitzmann. Seeing 3d objects in a single image via self-supervised static-dynamic disentanglement. *Proc. of the International Conf. on Learning Representations (ICLR)*, 2023. 3

[32] Vadim Tschernezki, Iro Laina, Diane Larlus, and Andrea Vedaldi. Neural feature fusion fields: 3d distillation of self-supervised 2d image representations. *Proc. of the International Conf. on 3D Vision (3DV)*, 2022. 3

[33] Matthew Wallingford, Aditya Kusupati, Alex Fang, Vivek Ramanujan, Aniruddha Kembhavi, Roozbeh Mottaghi, and Ali Farhadi. Neural radiance field codebooks. *Proc. of the International Conf. on Learning Representations (ICLR)*, 2023. 3

[34] Bing Wang, Lu Chen, and Bo Yang. Dm-nerf: 3d scene geometry decomposition and manipulation from 2d images. *Proc. of the International Conf. on Learning Representations (ICLR)*, 2023. 3

[35] 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 *Proc. of the European Conf. on Computer Vision (ECCV)*, 2022. 2, 6

[36] 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. *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 1, 2, 4, 6, 8, 13, 14

[37] Qianyi Wu, Xian Liu, Yuedong Chen, Kejie Li, Chuanxia Zheng, Jianfei Cai, and Jianmin Zheng. Object-compositional neural implicit surfaces. *Proc. of the European Conf. on Computer Vision (ECCV)*, 2022. 1, 2, 3, 4, 5, 6, 8, 12

[38] Tianhao Wu, Fangcheng Zhong, Andrea Tagliasacchi, Forrester Cole, and Cengiz Oztireli. D2nerf: Self-supervised decoupling of dynamic and static objects from a monocular video. *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 3

[39] Bangbang Yang, Yinda Zhang, Yijin Li, Zhaopeng Cui, Sean Fanello, Hujun Bao, and Guofeng Zhang. Neural rendering in a room: Amodal 3d understanding and free-viewpoint rendering for the closed scene composed of pre-captured objects. *ACM Trans. on Graphics*, 2022. 3

[40] 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 *Proc. of the IEEE International Conf. on Computer Vision (ICCV)*, 2021. 1, 3, 6

[41] Lior Yariv, Jiatao Gu, Yoni Kasten, and Yaron Lipman. Volume rendering of neural implicit surfaces. *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 1, 2, 6, 13, 14

[42] Lior Yariv, Yoni Kasten, Dror Moran, Meirav Galun, Matan Atzmon, Basri Ronen, and Yaron Lipman. Multiview neural surface reconstruction by disentangling geometry and appearance. *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 2

[43] Zehao Yu, Songyou Peng, Michael Niemeyer, Torsten Sattler, and Andreas Geiger. Monosdf: Exploring monocular geometric cues for neural implicit surface reconstruction. *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 1, 2, 3, 4, 5, 6, 8, 12

[44] 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 *Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2021. 3

[45] Shuaifeng Zhi, Tristan Laidlow, Stefan Leutenegger, and Andrew J Davison. In-place scene labelling and understanding with implicit scene representation. In *Proc. of the IEEE International Conf. on Computer Vision (ICCV)*, 2021. 1, 2, 3

Authorized licensed use limited to: Zhejiang University. Downloaded on September 29,2024 at 07:14:50 UTC from IEEE Xplore. Restrictions apply.