ObjectSDF++: Improved Object-Compositional Neural Implicit Surfaces

Qianyi Wu[1]   Kaisiyuan Wang[2]   Kejie Li[3]   Jianmin Zheng[4]   Jianfei Cai[1]

[1]Monash University   [2]University of Sydney   [3]University of Oxford   [4]Nanyang Technological University

{qianyi.wu, jianfei.cai}@monash.edu kaisiyuan.wang@sydney.edu.au

kejie.li@outlook.com asjmzheng@ntu.edu.sg

Abstract

In recent years, neural implicit surface reconstruction has emerged as a popular paradigm for multi-view 3D reconstruction. Unlike traditional multi-view stereo approaches, the neural implicit surface-based methods leverage neural networks to represent 3D scenes as signed distance functions (SDFs). However, they tend to disregard the reconstruction of individual objects within the scene, which limits their performance and practical applications. To address this issue, previous work ObjectSDF introduced a nice framework of object-composition neural implicit surfaces, which utilizes 2D instance masks to supervise individual object SDFs. In this paper, we propose a new framework called ObjectSDF++ to overcome the limitations of ObjectSDF. First, in contrast to ObjectSDF whose performance is primarily restricted by its converted semantic field, the core component of our model is an occlusion-aware object opacity rendering formulation that directly volume-renders object opacity to be supervised with instance masks. Second, we design a novel regularization term for object distinction, which can effectively mitigate the issue that ObjectSDF may result in unexpected reconstruction in invisible regions due to the lack of constraint to prevent collisions. Our extensive experiments demonstrate that our novel framework not only produces superior object reconstruction results but also significantly improves the quality of scene reconstruction. Code and more resources can be found in https://qianyiwu.github.io/objectsdf++.

1. Introduction

Recently, the neural implicit surface representation [55, 54, 45] has revived the classical area of multi-view 3D reconstruction that constructs a 3D scene from a set of multi-view images. Unlike traditional multi-view stereo approaches, neural implicit surface reconstruction leverages neural networks to implicitly represent 3D scenes as signed distance functions (SDFs). Due to its large modeling capability and resolution-free representation and also the wide applications of 3D reconstruction (for instance, in virtual and augmented reality and ditigal twins), this novel paradigm has brought multi-view 3D reconstruction to a new level and garnered the attention of the research community.

The success of the neural implicit surface representation should be attributed to the seminal work of Neural Radiance Field (NeRF) [23]. NeRF revolutionarily links the neural implicit 3D radiance field with 2D images via *volume rendering*, which simulates how light rays travel from the 3D scene to the 2D image plane by sampling points along each ray and accumulating their radiance. The neural implicit representation is designed as a Multi-Layer Perceptron (MLP), which takes 3D coordinates of a point as input and outputs RGB color and scene density. Since NeRF is mainly designed for view synthesis, its performance on 3D reconstruction is limited. This triggers the emergence of the neural implicit surface representation and reconstruction, which predicts an SDF value instead of a radiance field density and has proven its ability in 3D reconstruction [55, 26, 54, 45, 58, 7, 27].

One major limitation of most existing neural implicit surface reconstruction methods is that they basically focus on the representation of one entire scene or a single object, without paying much attention to individual objects within the scene. There have been a few attempts [53, 48, 17] that address this issue by introducing the object-compositional representation with additional guidance from 2D instance masks. In particular, ObjectSDF [48] proposes a simple and effective object-compositional neural implicit representation, where the MLP outputs individual object SDFs and the scene SDF is simply the minimum over all object SDFs. Moreover, ObjectSDF introduces a mapping function to convert each object SDF to a 3D semantic logit, which is volume-rendered to be compared with the given 2D semantic mask.

However, ObjectSDF also has a few limitations. First, it requires the additional mapping function with a tuning parameter for constructing the vectorized semantic field from

object SDFs, which introduces additional mapping noise during model optimization. Second, it mainly emphasizes the semantic field of the front object by applying the cross entropy loss to supervise the rendered semantic label with the ground-truth semantic label, lacking the supervision for other occluded object SDFs. Third, there is no constraint between any two nearby objects to prevent collisions and overlaps. Moreover, the model convergence speed of ObjectSDF is very slow due to its heavy MLP structure.

Therefore, in this paper, we propose a new framework called *ObjectSDF++* to address the limitations of ObjectSDF. While our method is built upon ObjectSDF, it is not merely a straightforward extension of ObjectSDF. The core of our model is an occlusion-aware object opacity rendering formulation (see Fig. 1), which removes the entire semantic field of ObjectSDF and directly volume-renders object opacity to predict individual object opacity masks. The key insight behind this is that a ground truth instance mask not only provides information about which object is in the front but also indicates that *this object should absorb all the light emanating along the ray in volume rendering*. Our new representation gives a stronger constraint compared with the solution in ObjectSDF that only considers the rendered semantic logit close to the ground truth label. Our experimental results show that such a design substantially improves the reconstruction quality of both scenes and objects.

To alleviate the problem of overgrown object shapes, we further design a physical-based regularization that penalizes an object intruding into other objects. The key observation is that object SDFs should be mutually exclusive. Specifically, we notice that if one point is located inside an object with a negative SDF value of s, it suggests its distance to any other objects should be greater than or equal to $-s$. Such a simple and intuitive regularization term helps the model separate different objects and alleviate the object collision problem. In addition, inspired by previous works [57, 6, 47, 33, 4] that use geometry cues to improve neural rendering and reconstruction, we also enforce monocular geometry cues from a pre-trained model to facilitate high-quality object-compositional surface reconstruction and faster model convergence. To further speed up the training, we adopt a multi-resolution feature grid as position embedding together with a tiny MLP structure, following [24, 57].

Overall, our main contributions are as follows.

- We propose a new occlusion-aware object opacity rendering scheme to improve both scene and object reconstruction in the object-compositional neural implicit surface representation.

- To alleviate object collisions in the reconstruction, we introduce an object distinction regularization term,

which is simple and effective.

- We evaluate our method on different datasets and demonstrate the effectiveness of our proposed components, not only on object-level reconstruction but also on scene-scale surface reconstruction.

2. Related Work

Neural Implicit Representation. Coordinate-based MLPs, also termed as neural implicit representation have become a powerful representation form in approximating different signals like 3D geometry. Occupancy Networks [22] and DeepSDF[29] are two pioneers which first introduce the concept of implicitly encoding objects or scenes using a neural network, despite that these studies need 3D ground-truth models. Scene Representation Networks (SRN) [35] and Neural Radiance Field (NeRF) [23] reveal that geometry and appearance can be simultaneously learned from multiple RGB photos utilizing multi-view consistency. Further predictions using this implicit representation include semantic labels [59, 42], deformation fields [31, 30], and high-fidelity specular reflections [41].

Among the existing neural implicit representation works, NeRF's learning-by-volume-rendering paradigm has received tremendous attentions. It sets the basis for many subsequent efforts, including ours. On the other hand, several works [55, 26, 20, 54, 45, 7] show that rendering neural implicit surfaces—where gradients are concentrated around surface regions—can produce a high-quality 3D reconstruction, compared with rendering a neural radiance field. Particularly, the recent efforts [54, 45] combine volume rendering and neural implicit surfaces, achieving very accurate surface reconstruction. Despite the good reconstruction performance, all these methods mainly focus on the representation of one entire scene or a single object. In contrast, our framework jointly represents and reconstructs both scene and its objects in an object-compositional manner.

Object-Compositional Neural Implicit Representation. Decomposing a holistic neural implicit representation into several parts or individual object representations can benefit many downstream applications such as scene editing, generation and understanding. Several attempts have been made. The existing methods can be categorized as category-specific or category-agnostic methods.

The category-specific methods learn the object-centric representation from many training data in the same category [9, 14]. For example, Ost *et al.* [28] utilize the scene graph information in rendering a dynamic street scene. Yu *et al.* [56] and Guo *et al.* [13] concentrate on extracting object-centric implicit representations by exploiting a large number of synthetic data. Niemeyer and Geiger [25] introduce GIRAFFE that conditions on latent codes to obtain object-centric NeRFs and thus represents a scene as a com-
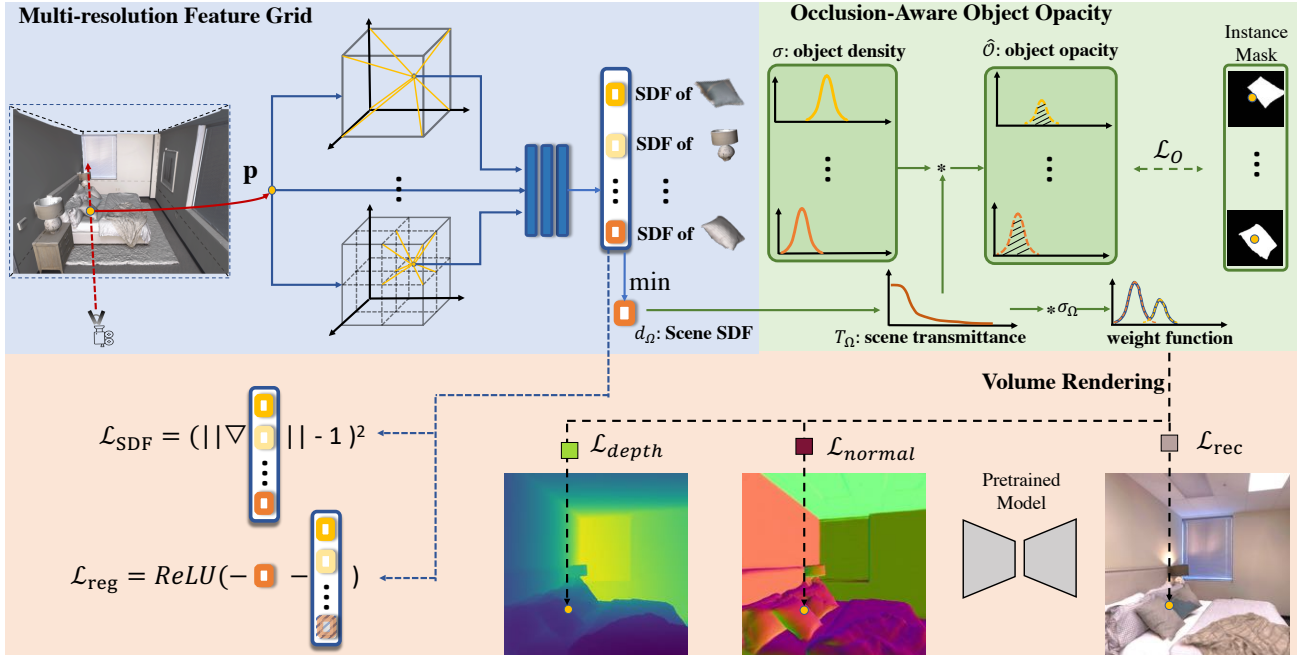
Figure 1. **Overview of our proposed framework.** Our model is able to reconstruct high-fidelity object geometry by predicting the SDF of each object inside the scene. The key to its success is the proposed occlusion-aware object opacity rendering scheme that assures the frontal object to absorb all photons during volume rendering, which helps in object surface reconstruction. A novel object distinction regularization loss is also designed to decompose objects better. To facilitate the training convergence, we use the multi-resolution feature grid and monocular geometry cues from the pre-trained model [8] to guide the optimization of the network.

positional generative neural feature field. Xu [52] proposes an object-centric 3D scene generative NeRF framework DiscoScene by leveraging the 3D object location prior to achieving object-compositional generation. Although these methods can utilize the fruitful information in the dataset, they have difficulty in generalizing to unseen objects.

The category-agnostic approaches learn an object-compositional neural implicit representation without involving a large dataset for learning object priors. They rely on additional information to figure out objects in a scene. For example, D2Nerf [50] proposes to learn the dynamic object inside the scene from dynamic video data. Different from a dynamic video [36, 40] which can provide a natural decomposition between dynamic objects and the static background, learning an object-compositional representation from a static scene requires extra guidance during training. The existing methods mainly adopt pixel-level guidances such as segmentation masks or neural features [39, 10, 21, 43, 18, 34, 19, 51, 46, 3]. For instance, Sosuke and Eiichi [16] used a well-trained model like CLIP [32] or DINO [2] to learn a neural feature field to support further editing and manipulation by feature matching.

Compared with the neural features, segmentation masks can provide more accurate object information and are not difficult to obtain. In particular, ObjectNeRF [53] uses instance masks to supervise the learning of each object

NeRF. To better capture the high-fidelity instance geometry, ObjectSDF [48] introduces a strong connection between each object SDF and its semantic label to obtain object-compositional neural implicit surfaces with the guidance of RGB images and their associated instance masks. However, it still suffers from several issues such as inaccurate object and scene reconstruction, slow convergence and training speed. A very recent concurrent work, vMAP [17], aims to solve the object-compositional SLAM from an RGB-D video with instance masks. With the help of a depth sensor, vMAP can sample the points around the depth for fast and accurate surface reconstruction. In contrast, we focus on the more challenging case that has no accurate metric depth but only RGB multiview images and instance masks, same as the setting of ObjectSDF. Our method is built upon ObjectSDF but resolves its limitations.

3. Methodology

Given a set of N posed RGB images $\mathcal{I} = \{I_1, \ldots, I_N\}$ with the associated instance-segmentation mask $\mathcal{S} = \{S_1, \ldots, S_N\}$, our goal is to reconstruct the scene geometry in an object-compositional neural implicit representation which describes not only the entire scene geometry but also each individual instance's geometry. We propose a new framework called *ObjectSDF++* for this purpose, which includes several modules as shown in Fig. 1. The technical details of the core modules are explained in this section.

3.1. Preliminary

Volume Rendering of Scene Density. The key to learning the neural implicit representation from multi-view images is volume rendering [15]. Considering a ray $\mathbf{r}(v) = \mathbf{o} + v\mathbf{d}$ emanated from a camera position $\mathbf{o}$ in the direction of $\mathbf{d}$, the volume rendering is essentially about the integrated light radiance along the ray. There are two crucial factors in the volume rendering: scene density $\sigma(v)$ and scene radiance $\mathbf{c}(v)$ on each 3D point along a ray. An important quantity used in the volume rendering computation is the scene transmittance function $T(v) = \exp(-\int_{v_n}^{v} \sigma(\mathbf{r}(u))du)$, measuring the energy loss during the ray traversing. Thus, the scene opacity O can be defined as the complement probability of T, *i.e.*, $O(v) = 1 - T(v)$. In this way, a target pixel color of volume rendering can be calculated as the integral of the probability density function of scene opacity $\frac{dO(v)}{dv} = T(v)\sigma(\mathbf{r}(v))$ [38] and radiance $\mathbf{c}(\mathbf{r}(v))$ along the ray from near bound v_n to far bound v_f:

$$\hat{C}(\mathbf{r}) = \int_{v_n}^{v_f} T(v)\sigma(\mathbf{r}(v))\mathbf{c}(\mathbf{r}(v))dv. \quad (1)$$

This integral is approximated using numerical quadrature.

SDF-based Neural Implicit Surface Representation. SDF representation is often used to directly characterize the geometry at the surface. Specifically, given a scene $\Omega \subset \mathbb{R}^3$, $\mathcal{M} = \partial\Omega$ is the boundary surface. The SDF d_Ω is defined as the signed distance from point $\mathbf{p}$ to the boundary $\mathcal{M}$:

$$d_\Omega(\mathbf{p}) = (-1)^{\mathbb{1}_\Omega(\mathbf{p})} \min_{\mathbf{y} \in \mathcal{M}} ||\mathbf{p} - \mathbf{y}||_2, \quad (2)$$

where $\mathbb{1}_\Omega(\mathbf{p})$ is the indicator function denoting whether $\mathbf{p}$ belongs to the scene Ω or not. If the point is outside the scene, $\mathbb{1}_\Omega(\mathbf{p})$ returns 0; otherwise returns 1. The standard l_2-norm is commonly used to compute the distance.

To learn the SDF from multiview images without the need for ground truth SDF values, pioneer works [45, 54] combine SDF with neural implicit function and volume rendering to get supervision from RGB images. The common idea is to replace the NeRF volume density output $\sigma(\mathbf{p})$ with a transformation function of the SDF value $d_\Omega(\mathbf{p})$. Specifically, in [54], the density $\sigma(\mathbf{p})$ is converted from a specific tractable transformation:

$$\sigma(\mathbf{p}) = \begin{cases} \frac{1}{2\beta} \exp\left(\frac{-d_\Omega(\mathbf{p})}{\beta}\right) & \text{if } d_\Omega(\mathbf{p}) \geq 0 \\ \frac{1}{\beta} - \frac{1}{2\beta} \exp\left(\frac{d_\Omega(\mathbf{p})}{\beta}\right) & \text{if } d_\Omega(\mathbf{p}) < 0 \end{cases} \quad (3)$$

where β is a learnable parameter to decide the sharpness of the surface density. This design implicitly assumes that the object's surface is solid. The neural implicit surface is supervised by RGB reconstruction loss and an Ekional regularization loss [11] for the SDF field.
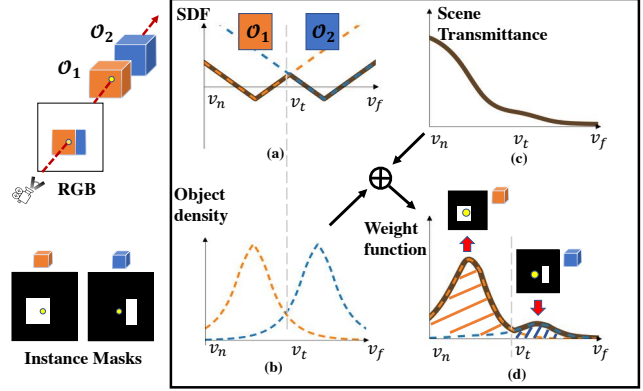

Figure 2. **Occlusion-Aware Object Opacity Rendering**. We show a toy example of occlusion-aware object opacity rendering. The opacity is essentially the integration of object density and scene transmittance. For different objects, the integration interval will differ according to its range. We propose an approximated form of object opacity rendering, which can be supervised by the pixel value in the ground truth instance mask.

Modeling an object-compositional scene geometry. To represent each individual instance geometry inside the scene, the previous method ObjectSDF [48] predicts the SDF of each object. Suppose that one scene Ω is a composition of K different objects $\{\mathcal{O}_i \in \mathbb{R}^3 | i = 1, \ldots, K\}$, we have $\Omega = \bigcup_{i=1}^{k} \mathcal{O}_i$. Then, the scene SDF can be calculated as the minimum of all object SDFs:

$$d_\Omega(\mathbf{p}) = \min_{i=1\ldots k} d_{\mathcal{O}_i}(\mathbf{p}). \quad (4)$$

With the scene SDF, volume rendering can be applied to render corresponding RGB images, normal maps, *etc*.

3.2. Occlusion-Aware Object Opacity Rendering

In order to learn the individual object geometry from the supervision of instance segmentation masks, the state-of-the-art approach ObjectSDF [48] adopts a Sigmoid-type transition function to convert each object SDF to a 3D semantic logit, which only changes rapidly around the object boundary. With the help of volume rendering over the transformed semantic field, the model will push the surface of the corresponding object to the front to meet the guidance from the 2D instance [48].

However, such a design has a few limitations. First, it mainly focuses on the semantic field of the front object by employing the cross entropy loss to oversee the predicted semantic label with the semantic logit of reality, which fails to provide guidance for the SDFs of obstructed objects. Second, the softmax operation across different rendered semantic labels often has energy leaking, which cannot exactly match the one-hot ground-truth label. Third, it requires an

additional mapping function to convert object SDFs to semantic logits with a tuning hyperparameter.

To tackle these limitations, we propose to remove the need to build a vectorized semantic field and directly volume-render the opacity of each object. Specifically, consider that when a frontal object occludes another object, the frontal one should block *all* the photons emanating along the inverse ray. In terms of volume rendering, the object opacity of the frontal object should be 1 while that of the occluded object should be 0. For example, as shown in Fig 2, the ray casting from left to right crosses the two color cubes $\mathcal{O}_1$ and $\mathcal{O}_2$. If we inspect the selected pixel in the instance mask, we are not able to see the blue cube from a ray casting through the selected pixel (bottom-left of Fig. 2). It means that when the photons travel through the inverse ray casting from the camera, the transmittance should become zero after they cross the *first* object. As all the objects decompose the entire scene, they naturally form a Voronoi diagram [1] based on individual object SDF. The light energy absorbed by the first object can be approximated by the integration of the weight function from near v_n to the SDF turning point v_t, (dash line in Fig 2(a), where the minimum SDF field changes from $\mathcal{O}_1$ to $\mathcal{O}_2$. Specifically, the object opacity for the first object $\mathcal{O}_1$ can be expressed as:

$$O_{\mathcal{O}_1}(\mathbf{r}) = \int_{v_n}^{v_t} T_\Omega(v)\sigma_\Omega(\mathbf{r}(v))dv$$
$$= \int_{v_n}^{v_t} T_\Omega(v)\sigma_{\mathcal{O}_1}(\mathbf{r}(v))dv \quad (5)$$

where T_Ω, σ_Ω are the scene transmittance and scene density function. Because in $[v_n, v_t]$, the scene SDF is dominated by the SDF of $\mathcal{O}_1$, resulting in $\sigma_\Omega = \sigma_{\mathcal{O}_1}$.

This rendering equation can also be used to render the second object opacity by changing the integration interval, *i.e,* $O_{\mathcal{O}_2}(\mathbf{r}) = \int_{v_t}^{v_f} T_\Omega(v)\sigma_{\mathcal{O}_2}(\mathbf{r}(v))dv$. Thanks to the consideration of scene transmittance in the integration, such an object opacity rendering is occlusion-aware. However, it is hard to find the scene SDF turning points during training. On the other hand, consider the following relationship

$$O_{\mathcal{O}_2}(\mathbf{r}) = \int_{v_t}^{v_f} T_\Omega(v)\sigma_{\mathcal{O}_2}(\mathbf{r}(v))dv$$
$$\leq \int_{v_n}^{v_f} T_\Omega(v)\sigma_{\mathcal{O}_2}(\mathbf{r}(v))dv, \quad (6)$$

where the non-negative of transmittance and density guarantee this inequality. This motivates us to approximate the first integration in (6) by the second one since $\sigma_{\mathcal{O}_2}$ is expected to be close to 0 in $[v_n, v_t]$, and similarly extend the integration in (5) to $[v_n, v_f]$ since T_Ω is expected to be close to 0 in $[v_t, v_f]$. In this way, we reach a general form of approximated occlusion-aware object opacity rendering:

$$\hat{O}_{\mathcal{O}_i}(\mathbf{r}) = \int_{v_n}^{v_f} T_\Omega(v)\sigma_{\mathcal{O}_i}(\mathbf{r}(v))dv, i = 1, \dots, K. \quad (7)$$

Then, we adopt each instance segmentation mask as the opacity supervision to guide the learning of each object surface. The loss function is defined as:

$$\mathcal{L}_O = \mathbb{E}_{\mathbf{r}\in\mathcal{R}}[\frac{1}{K}\sum_{i=1\dots K}\|\hat{O}_{\mathcal{O}_i}(\mathbf{r}) - O_{\mathcal{O}_i}(\mathbf{r})\|] \quad (8)$$

where $\mathcal{R}$ denotes the set of rays in a minibatch and $O_{\mathcal{O}_i}(\mathbf{r})$ is the GT object opacity from the instance mask. Further discussion and detailed information can be found in the supplementary.

3.3. Object Distinction Regularization Term

Although the occlusion-aware object opacity rendering equation describes the proper light behavior during the volume rendering, there is still a lack of enough constraint to regulate individual object geometries when the transmittance approaches zero. In fact, the 2D guidance of multi-view images and instance segmentation, and the geometry guidance like depth or normal map, can only guide the visible regions of a neural radiance field. The information in invisible regions could be arbitrary since they have no influence on the loss metric between the rendered results and the 2D guidance. It could make an object overgrow in an invisible region just like an underwater iceberg.

To alleviate this issue, we enforce that each object inside the scene should be distinct and there should be no overlap between any two objects. In particular, suppose there are K objects in a scene, for any input point coordinates $\mathbf{p}$, our model will predict the K SDFs for the K object surfaces. If the point is located inside the object $\mathcal{O}_1$ with a negative SDF value $d_{\mathcal{O}_1}$, one important property for the remaining objects $\{d_{\mathcal{O}_2}, d_{\mathcal{O}_3}, \dots, d_{\mathcal{O}_K}\}$ is that their SDFs need to be positive and should not less than $-d_{\mathcal{O}_1}$. In other words, it also means for the rest of the objects, their SDFs should be large than the scene SDF $-d_\Omega$ (*i.e,* $d_{\mathcal{O}_i} \geq -d_\Omega, \forall d_{\mathcal{O}_i} \neq d_\Omega$). Therefore, for the SDF vector prediction of one point $\mathbf{p}$, we propose to minimize the following object distinction regularization term:

$$\mathcal{L}_{reg} = \mathbb{E}_{\mathbf{p}}[\sum_{d_{\mathcal{O}_i}(\mathbf{p})\neq d_\Omega(\mathbf{p})} \text{ReLU}(-d_{\mathcal{O}_i}(\mathbf{p}) - d_\Omega(\mathbf{p}))] \quad (9)$$

We also find that such a requirement also holds for any point located outside of all objects. In this case, $d_\Omega > 0$ and for each object SDF $d_{\mathcal{O}_i} \geq d_\Omega > -d_\Omega$. Hence, we random sample points in the entire 3D space to apply this regularization term.

3.4. Model Training

The slow convergence of implicit neural surfaces is a common issue in most of the existing methods. The inherent reason is the heavy structure of MLP as a network backbone. To accelerate the convergence speed, inspired by the
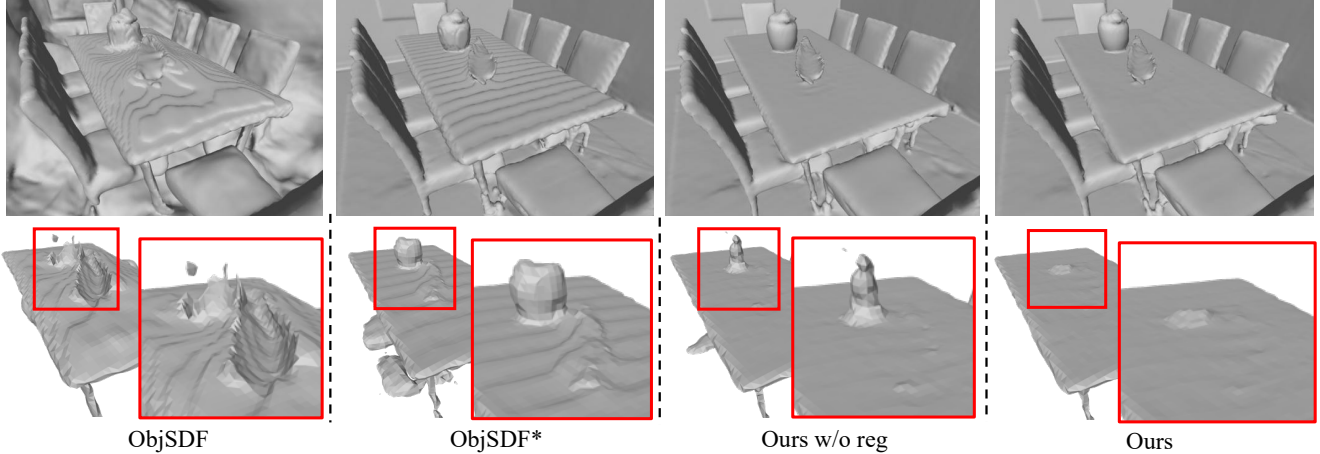
| ObjSDF | ObjSDF* | Ours w/o reg | Ours |

Figure 3. **Qualitative improvements of our proposed components.** The first row is the scene reconstruction results and the second row is the reconstruction of a single object (the dining table). Our model aims to improve both scene and object reconstruction in an object-compositional way. We can see that the occlusion-aware object opacity can significantly improve the surface reconstruction and the object distinction regularization term can further help to reduce the artifacts in the invisible region.

Method	Model Components			Scene Reconstruction		Object Reconstruction	
	Object Guidance	Mono Cue	Regularizer	Chamfer-L1↓	F-score ↑	Chamfer-L1↓	F-score ↑
ObjSDF [48]	Semantic			22.8±14.8	25.74 ±5.70	7.05±0.40	59.91±6.67
ObjSDF*	Semantic	✓		4.14±1.53	78.34±15.34	4.65±0.11	74.06±3.12
Ours w/o reg	Occlusion Opacity	✓		3.60± 1.13	85.59± 8.12	3.78± 0.09	79.51± 2.92
Ours	Occlusion Opacity	✓	✓	3.58±1.08	85.69 ±6.55	3.74± 0.09	80.10 ± 2.63

Table 1. **The quantitative average results from 8 Replica scenes evaluated on scene and object reconstruction**. We compare our results with vanilla ObjectSDF [48] and its improved version, ObjSDF* (adding monocular geometry cue).

current efficient neural implicit represent work [24, 49, 57], we take a multi-resolution feature grid (multi-res grid) as position embedding together with a tiny MLP to boost the model convergence and the inference speed.

To optimize the network of our object-compositional neural implicit surface, we first minimize the image reconstruction loss between the predicted color $\hat{C}(\mathbf{r})$ and the ground truth color $C(\mathbf{r})$:

$$\mathcal{L}_{rec} = \sum_{\mathbf{r} \in \mathcal{R}} \|\hat{C}(\mathbf{r}) - C(\mathbf{r})\|_1 \qquad (10)$$

where $\mathcal{R}$ denotes the set of sampling rays in a minibatch. Following [11, 54, 48], we add an Eiknoal term on the sampled points to regularize object SDFs and scene SDF:

$$\mathcal{L}_{SDF} = \sum_{i=1}^{k} \mathbb{E}_{d_{\mathcal{O}_i}}(\|\nabla d_{\mathcal{O}_i}(\mathbf{p})\| - 1)^2 + \mathbb{E}_{d_{\Omega}}(\|\nabla d_{\Omega}(\mathbf{p})\| - 1)^2.$$
$$(11)$$

Moreover, recent approaches [57, 44, 12] have proven that surface reconstruction could be significantly improved by incorporating geometry guidance such as depth and surface normal. Thus, we adopt the design from [57] to enforce the similarity between the rendered depth/normal map and the estimated monocular depth/normal from a pretrained model Omnidata [8], which is termed as the monocular cue loss

$\mathcal{L}_{mono} = \lambda_{depth}\mathcal{L}_{depth} + \lambda_{normal}\mathcal{L}_{normal}$. Finally, the overall training loss of ObjectSDF++:

$$\mathcal{L} = \mathcal{L}_{rec} + \mathcal{L}_O + \mathcal{L}_{mono} + \lambda_1\mathcal{L}_{SDF} + \lambda_2\mathcal{L}_{reg}. \quad (12)$$

4. Experiment

Our proposed ObjectSDF++ aims to create a neural implicit representation for a high-fidelity object-compositional surface reconstruction that can be beneficial for both object and scene reconstruction. To assess the effectiveness of our approach, we conducted experiments on two datasets: one simulated and one real-world dataset. To evaluate both scene and object reconstruction, we compare ours with the SOTA methods for object-compositional surface reconstruction. We further compare our approach with many SOTA methods for neural surface reconstruction on large-scale real-world scenes to examine the benefit of object-compositional surface reconstruction.

Implementation details. We implement our method in Pytorch and apply the Adam optimizer with a learning rate of $5e - 4$ for the tiny MLP part (2 layers with 256 channels for the SDF prediction and 4 layers with 256 channels for the color prediction) and $1e - 2$ for the multi-res feature grid. The hyperparameters $\lambda_{depth}, \lambda_{normal}, \lambda_1, \lambda_2$ in (12) are set as $0.1, 0.05, 0.1, 0.5$ respectively. For each mini-
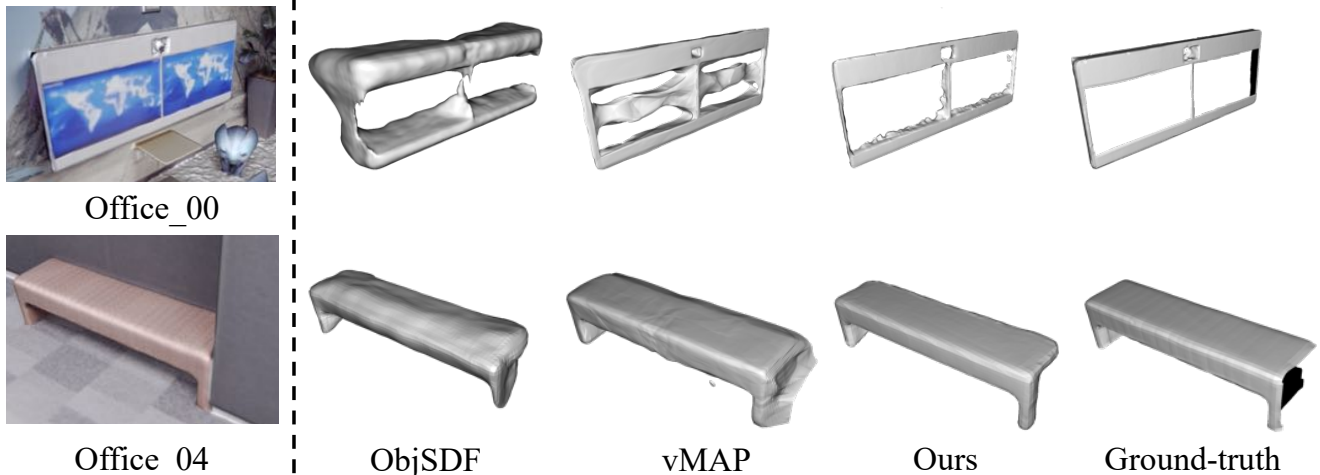
Figure 4. **Qualitative comparisons of object reconstructions on Replica by our method and two SOTA approaches [48, 17]**. Our model can produce fewer artifacts in the invisible region due to the guidance of the regularization term and occlusion-aware object opacity rendering. All meshes of vMAP are provided by the original authors.

batch during training, we sample 1024 rays and apply the error-bounded sampling algorithm [54] to decide the sampling points along a ray. Our method can be trained in a single 12 GB GTX 2080Ti GPU. The initial value of β in (3) is 0.1. More details can be found in the supplementary.

Dataset and metrics. We conduct our experiments on two datasets: a) Replica [37] and b) ScanNet [5]. Replica is a synthesized dataset with accurate camera poses and clear object masks. Following the setting in [60, 57, 17], we use 8 scenes from Replica for experiments. For quantitative evaluation, we measure Chamfer Distance together with F-score with a threshold of 5cm on Replica[1]. ScanNet is a real-world dataset captured by an RGB-D camera. It contains 2D instance segmentation masks with RGB images. Following the setting [12, 57, 60], we select 4 scenes from ScanNet for experiments and report the accuracy, completeness, Chamfer Distance, precision, recall and F-score.

Baseline Methods. We mainly compare our method with the following representative works in the realm of object-compositional neural implicit surface reconstruction.

ObjectSDF [48]. ObjectSDF is the first to use the neural implicit representation for object-compositional surface reconstruction. Its key idea is to predict individual object SDFs and convert each of them to a semantic label with a pre-defined function. The converted 3D semantic labels will accumulate along the ray by volume rendering and be supervised with the 2D semantic map.

ObjectSDF*. Considering our method was built upon ObjectSDF with several improved designs. To better verify the effectiveness, we also create another baseline called **ObjectSDF***, where we simply add the monocular guidance and multi-res feature grid into the vanilla ObjectSDF to improve the reconstruction and the model convergence.

---

[1]Details about metric definition are provided in the supplementary.

vMAP [17]. vMAP is the latest concurrent work, which uses RGB-D images for an object-level dense SLAM system. With accurate depth information, it can sample points near the surface to handle the occlusion intuitively. We also compare our object reconstruction results with it.

4.1. Object and Scene Reconstruction on Replica

We begin with the experiments on the Replica dataset, aiming at evaluating the efficacy of our proposed components via a fair comparison of ObjSDF, ObjSDF*, and our method for both scene and object reconstructions in RGB image settings. Table 1 shows that integrating monocular cues with vanilla ObjectSDF significantly enhances the overall scene reconstruction performance. However, as illustrated in Figure 3, the reconstructed scene is not entirely smooth and clean. More critically, in the object reconstruction result (*i.e.*, the dining table) in Figure 3, artifacts appear in the pot area. Our conjecture is that the design of volume-rendering the semantic field in ObjectSDF only ensures the frontal object is in the correct position, but does not guarantee that it absorbs all photons. As a result, the geometry of other objects situated behind may cause unexpected artifacts due to light energy leak.

In contrast, our occlusion-aware object opacity design, which differs from the semantic field rendering in ObjectSDF, results in a considerable improvement, especially in terms of visual quality. As shown in Figure 3, the artifacts on the dining table geometry are significantly reduced in the result of 'Ours w/o reg', which also increases the accuracy of the scene geometry by 7.25% in the F-score of the scene reconstruction. However, occlusion-aware object opacity rendering alone is inadequate for eliminating artifacts in the invisible region. Our proposed object distinction regularizer plays an essential role in enhancing the geome-

	Accuracy ↓	Completeness ↓	Chamfer-L1 ↓	Precision ↑	Recall ↑	F-Score ↑
UNISURF [26]	0.554	0.164	0.359	0.212	0.362	0.267
Neus [45]	0.179	0.208	0.194	0.313	0.275	0.291
VolSDF [54]	0.414	0.120	0.267	0.321	0.394	0.346
Manhattan-SDF [12]	0.072	0.068	0.070	0.621	0.586	0.602
NeuRIS [44]	0.050	0.049	0.050	0.717	0.669	0.692
MonoSDF (Multi-Res Grids) [57]	0.072	0.057	0.064	0.660	0.601	0.626
Ours (Multi-Res Grids)	0.047	0.045	0.046	0.749	0.707	0.726
MonoSDF (MLP) [57]	0.035	0.048	0.042	0.799	0.681	0.733
Ours (MLP)	0.039	0.045	0.042	0.781	0.706	0.740

Table 2. **The quantitative results of the scene reconstruction on ScanNet**. We compare our method against various recent neural implicit surface reconstruction methods. The top-2 results are highlighted in red and yellow, respectively.
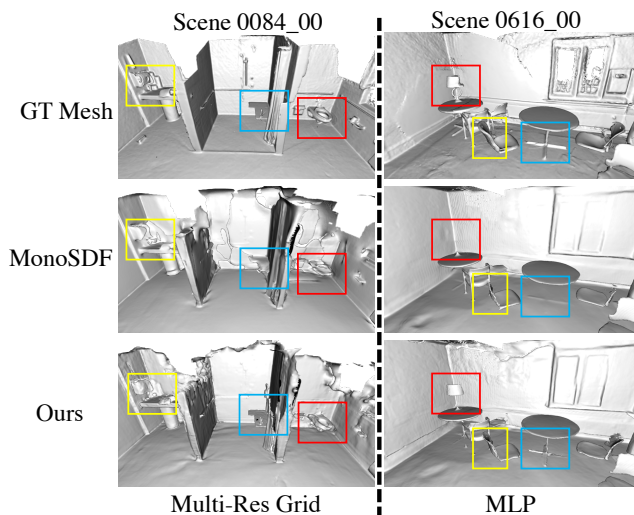

Figure 5. **Qualitative comparisons on ScanNet**. We compare with the recent MonoSDF [57] with two types of network architectures (multi-res grid-based and MLP-based) on two ScanNet examples. With the help of object-compositional modeling, our method produces better results with more details (*e.g.*, the desk lamp, toilet, and table). All MonoSDF meshes are provided by the original authors.

try of invisible regions, leading to further quantitative and qualitative improvements. Note that these two components are not computationally intensive and can be easily incorporated into other frameworks.

Figure 4 gives more qualitative comparisons of the object reconstruction results on Replica, where we also include the latest RGB-D based object-compositional NeRF method vMAP [17]. Although vMAP is able to comprehend object surface geometry by utilizing accurate depth information, it still remains vulnerable to imprecise geometry modeling when dealing with occlusion scenarios (e.g., the distortions in the TV screen region) and geometry within concealed areas (e.g., the chair placed on the wall), as illustrated in Figure 4. In contrast, our proposed method that solely relies on RGB inputs and predicted monocular cues

attains impressive geometry reconstruction results by employing our novel occlusion-aware object opacity rendering and object-distinct regularizer.

4.2. Real-World Large-Scale Scene Reconstruction

Table 2 gives a comprehensive quantitative comparison with various recent neural implicit surface reconstruction methods which only focus on scene reconstruction without considering the compositing objects. Even only considering scene reconstruction, our results are either the best or the second best. It is worth noting that MonoSDF reports two different results based on two distinct network structures: multi-resolution (multi-res) grid and multi-layer perceptron (MLP). For a fair comparison, we also implement our model with these two structures. As can be seen from Table 2, when using the light network architecture (i.e., the multi-res grid structure), compared to MonoSDF, our model improves F-score by 10% and Chamfer distance by 28%. When employing MLP, our quantitative results are comparable to those of the corresponding MonoSDF, while our visual results contain more detailed reconstruction (e.g., the desk lamp in Fig. 5). Moreover, we achieve the best F-score among all methods. Note that, unlike MonoSDF, the performance difference of our method with the two different network structures is negligible, which further affirms the robustness of our proposed components.

Fig. 5 gives a qualitative comparison between our method and MonoSDF on ScanNet. No matter which one of the two network architectures is being used, our model produces cleaner surface reconstruction and more details than MonoSDF. This suggests that the overall scene reconstruction quality gets improved when considering the composition of individual objects in the scene.

5. Conclusion

We have presented *ObjectSDF++*, a new framework for object-composition neural implicit surface reconstruction. Our key observation is that the frontal object should absorb all energy of light during volume rendering. Leveraging this

insight, we propose an occlusion-aware object opacity rendering scheme during training. Moreover, we design an object distinction regularization term to penalize the intersection of different objects in invisible regions. These two techniques significantly improve the quality of both scene-level and object-level surface reconstruction. Our framework can be extended in the future to support additional types of objects, such as transparent or semi-transparent objects, and be applied to various 3D scene editing applications.

References

[1] Franz Aurenhammer, Rolf Klein, and Der-Tsai Lee. *Voronoi diagrams and Delaunay triangulations*. World Scientific Publishing Company, 2013. 5

[2] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, 2021. 3

[3] Xiaokang Chen, Jiaxiang Tang, Diwen Wan, Jingbo Wang, and Gang Zeng. Interactive segment anything nerf with feature imitation. *arXiv preprint arXiv:2305.16233*, 2023. 3

[4] Yuedong Chen, Haofei Xu, Qianyi Wu, Chuanxia Zheng, Tat-Jen Cham, and Jianfei Cai. Explicit correspondence matching for generalizable neural radiance fields. *arXiv preprint arXiv:2304.12294*, 2023. 2

[5] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. ScanNet: Richly-annotated 3d reconstructions of indoor scenes. In *CVPR*, 2017. 7

[6] Kangle Deng, Andrew Liu, Jun-Yan Zhu, and Deva Ramanan. Depth-supervised nerf: Fewer views and faster training for free. In *CVPR*, 2022. 2

[7] Yu Deng, Jiaolong Yang, Jianfeng Xiang, and Xin Tong. Gram: Generative radiance manifolds for 3d-aware image generation. In *CVPR*, 2022. 1, 2

[8] 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 *ICCV*, 2021. 3, 6

[9] Dave Epstein, Taesung Park, Richard Zhang, Eli Shechtman, and Alexei A. Efros. Blobgan: Spatially disentangled scene representations. *ECCV*, 2022. 2

[10] Zhiwen Fan, Peihao Wang, Yifan Jiang, Xinyu Gong, Dejia Xu, and Zhangyang Wang. Nerf-sos: Any-view self-supervised object segmentation on complex scenes. In *ICLR*, 2023. 3

[11] Amos Gropp, Lior Yariv, Niv Haim, Matan Atzmon, and Yaron Lipman. Implicit geometric regularization for learning shapes. In *ICML*, 2020. 4, 6

[12] 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 *CVPR*, 2022. 6, 7, 8

[13] Michelle Guo, Alireza Fathi, Jiajun Wu, and Thomas Funkhouser. Object-centric neural scene rendering. *arXiv preprint arXiv:2012.08503*, 2020. 2

[14] 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 *ECCV*, 2022. 2

[15] James T Kajiya and Brian P Von Herzen. Ray tracing volume densities. *ACM SIGGRAPH*, 18(3):165–174, 1984. 4

[16] Sosuke Kobayashi, Eiichi Matsumoto, and Vincent Sitzmann. Decomposing nerf for editing via feature field distillation. In *NeurIPS*, 2022. 3

[17] Xin Kong, Shikun Liu, Marwan Taher, and Andrew J Davison. vmap: Vectorised object mapping for neural field slam. In *CVPR*, 2023. 1, 3, 7, 8

[18] 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 *CVPR*, 2022. 3

[19] Zizhang Li, Xiaoyang Lyu, Yuanyuan Ding, Mengmeng Wang, Yiyi Liao, and Yong Liu. Rico: Regularizing the unobservable for indoor compositional reconstruction. In *ICCV*, 2023. 3

[20] Fujun Luan, Shuang Zhao, Kavita Bala, and Zhao Dong. Unified shape and svbrdf recovery using differentiable monte carlo rendering. In *CGF*, volume 40, pages 101–113, 2021. 2

[21] Kirill Mazur, Edgar Sucar, and Andrew J Davison. Feature-realistic neural fusion for real-time, open set scene understanding. In *ICRA*, 2023. 3

[22] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *CVPR*, 2019. 2

[23] 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. In *ECCV*, 2020. 1, 2

[24] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM TOG.*, 41(4):102:1–102:15, July 2022. 2, 6

[25] Michael Niemeyer and Andreas Geiger. Giraffe: Representing scenes as compositional generative neural feature fields. In *CVPR*, 2021. 2

[26] Michael Oechsle, Songyou Peng, and Andreas Geiger. Unisurf: Unifying neural implicit surfaces and radiance fields for multi-view reconstruction. In *ICCV*, 2021. 1, 2, 8

[27] Roy Or-El, Xuan Luo, Mengyi Shan, Eli Shechtman, Jeong Joon Park, and Ira Kemelmacher-Shlizerman. StyleSDF: High-Resolution 3D-Consistent Image and Geometry Generation. In *CVPR*, 2022. 1

[28] Julian Ost, Fahim Mannan, Nils Thuerey, Julian Knodt, and Felix Heide. Neural scene graphs for dynamic scenes. In *CVPR*, 2021. 2

[29] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *CVPR*, 2019. 2

[30] Keunhong Park, Utkarsh Sinha, Peter Hedman, Jonathan T. Barron, Sofien Bouaziz, Dan B Goldman, Ricardo Martin-Brualla, and Steven M. Seitz. Hypernerf: A higher-dimensional representation for topologically varying neural radiance fields. *ACM TOG.*, 40(6), 2021. 2

[31] Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. D-nerf: Neural radiance fields for dynamic scenes. In *CVPR*, 2021. 2

[32] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 3

[33] 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 *CVPR*, 2022. 2

[34] Yawar Siddiqui, Lorenzo Porzi, Samuel Rota Buló, Norman Müller, Matthias Nießner, Angela Dai, and Peter Kontschieder. Panoptic lifting for 3d scene understanding with neural fields. In *CVPR*, 2023. 3

[35] Vincent Sitzmann, Michael Zollhöfer, and Gordon Wetzstein. Scene representation networks: Continuous 3d-structure-aware neural scene representations. In *NeurIPS*, 2019. 2

[36] Liangchen Song, Anpei Chen, Zhong Li, Zhang Chen, Lele Chen, Junsong Yuan, Yi Xu, and Andreas Geiger. Nerfplayer: A streamable dynamic scene representation with decomposed neural radiance fields. *IEEE TVCG*, 2023. 3

[37] Julian Straub, Thomas Whelan, Lingni Ma, Yufan Chen, Erik Wijmans, Simon Green, Jakob J. Engel, Raul Mur-Artal, Carl Ren, Shobhit Verma, Anton Clarkson, Mingfei Yan, Brian Budge, Yajie Yan, Xiaqing Pan, June Yon, Yuyang Zou, Kimberly Leon, Nigel Carter, Jesus Briales, Tyler Gillingham, Elias Mueggler, Luis Pesqueira, Manolis Savva, Dhruv Batra, Hauke M. Strasdat, Renzo De Nardi, Michael Goesele, Steven Lovegrove, and Richard Newcombe. The Replica dataset: A digital replica of indoor spaces. *arXiv preprint arXiv:1906.05797*, 2019. 7

[38] Andrea Tagliasacchi and Ben Mildenhall. Volume rendering digest (for nerf), 2022. 4

[39] Vadim Tschernezki, Iro Laina, Diane Larlus, and Andrea Vedaldi. Neural feature fusion fields: 3D distillation of self-supervised 2D image representations. In *3DV*, 2022. 3

[40] Vadim Tschernezki, Diane Larlus, and Andrea Vedaldi. NeuralDiff: Segmenting 3D objects that move in egocentric videos. In *3DV*, 2021. 3

[41] Dor Verbin, Peter Hedman, Ben Mildenhall, Todd Zickler, Jonathan T. Barron, and Pratul P. Srinivasan. Ref-NeRF: Structured view-dependent appearance for neural radiance fields. In *CVPR*, 2022. 2

[42] Suhani Vora, Noha Radwan, Klaus Greff, Henning Meyer, Kyle Genova, Mehdi SM Sajjadi, Etienne Pot, Andrea Tagliasacchi, and Daniel Duckworth. Nesf: Neural semantic fields for generalizable semantic segmentation of 3d scenes. *TMLR*, 2022. 2

[43] Bing Wang, Lu Chen, and Bo Yang. Dm-nerf: 3d scene geometry decomposition and manipulation from 2d images. In *ICLR*, 2023. 3

[44] 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 *ECCV*, 2022. 6, 8

[45] 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 *NeurIPS*, 2021. 1, 2, 4, 8

[46] Yuxin Wang, Wayne Wu, and Dan Xu. Learning unified decompositional and compositional nerf for editable novel view synthesis. In *ICCV*, 2023. 3

[47] Yi Wei, Shaohui Liu, Yongming Rao, Wang Zhao, Jiwen Lu, and Jie Zhou. Nerfingmvs: Guided optimization of neural radiance fields for indoor multi-view stereo. In *ICCV*, 2021. 2

[48] Qianyi Wu, Xian Liu, Yuedong Chen, Kejie Li, Chuanxia Zheng, Jianfei Cai, and Jianmin Zheng. Object-compositional neural implicit surfaces. In *ECCV*, 2022. 1, 3, 4, 6, 7

[49] Tong Wu, Jiaqi Wang, Xingang Pan, Xudong Xu, Christian Theobalt, Ziwei Liu, and Dahua Lin. Voxurf: Voxel-based efficient and accurate neural surface reconstruction. In *ICLR*, 2022. 6

[50] Tianhao Wu, Fangcheng Zhong, Andrea Tagliasacchi, Forrester Cole, and Cengiz Oztireli. D2nerf: Self-supervised decoupling of dynamic and static objects from a monocular video. In *NeurIPS*, 2022. 3

[51] Yuanbo Xiangli, Linning Xu, Xingang Pan, Nanxuan Zhao, Bo Dai, and Dahua Lin. Assetfield: Assets mining and reconfiguration in ground feature plane representation. In *ICCV*, 2023. 3

[52] Yinghao Xu, Menglei Chai, Zifan Shi, Sida Peng, Ivan Skorokhodov, Aliaksandr Siarohin, Ceyuan Yang, Yujun Shen, Hsin-Ying Lee, Bolei Zhou, et al. Discoscene: Spatially disentangled generative radiance fields for controllable 3d-aware scene synthesis. In *CVPR*, 2023. 3

[53] 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 *ICCV*, October 2021. 1, 3

[54] Lior Yariv, Jiatao Gu, Yoni Kasten, and Yaron Lipman. Volume rendering of neural implicit surfaces. In *NeurIPS*, 2021. 1, 2, 4, 6, 7, 8

[55] 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 *NeurIPS*, 2020. 1, 2

[56] Hong-Xing Yu, Leonidas Guibas, and Jiajun Wu. Unsupervised discovery of object radiance fields. In *ICLR*, 2022. 2

[57] Zehao Yu, Songyou Peng, Michael Niemeyer, Torsten Sattler, and Andreas Geiger. Monosdf: Exploring monocular

geometric cues for neural implicit surface reconstruction. In *NeurIPS*, 2022. 2, 6, 7, 8

[58] Kai Zhang, Fujun Luan, Zhengqi Li, and Noah Snavely. Iron: Inverse rendering by optimizing neural sdfs and materials from photometric images. In *CVPR*, 2022. 1

[59] Shuaifeng Zhi, Tristan Laidlow, Stefan Leutenegger, and Andrew Davison. In-place scene labelling and understanding with implicit scene representation. In *ICCV*, 2021. 2

[60] Zihan Zhu, Songyou Peng, Viktor Larsson, Weiwei Xu, Hujun Bao, Zhaopeng Cui, Martin R. Oswald, and Marc Pollefeys. Nice-slam: Neural implicit scalable encoding for slam. In *CVPR*, 2022. 7