This ICCV paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Learning A Room with the Occ-SDF Hybrid: Signed Distance Function Mingled with Occupancy Aids Scene Representation

Xiaoyang Lyu[1], Peng Dai[1], Zizhang Li[2], Dongyu Yan[3], Yi Lin[4], Yifan Peng[1], Xiaojuan Qi[1],

[1]The University of Hong Kong , [2] Zhejiang University, [3] Harbin Institute of Technology, [4] DJI

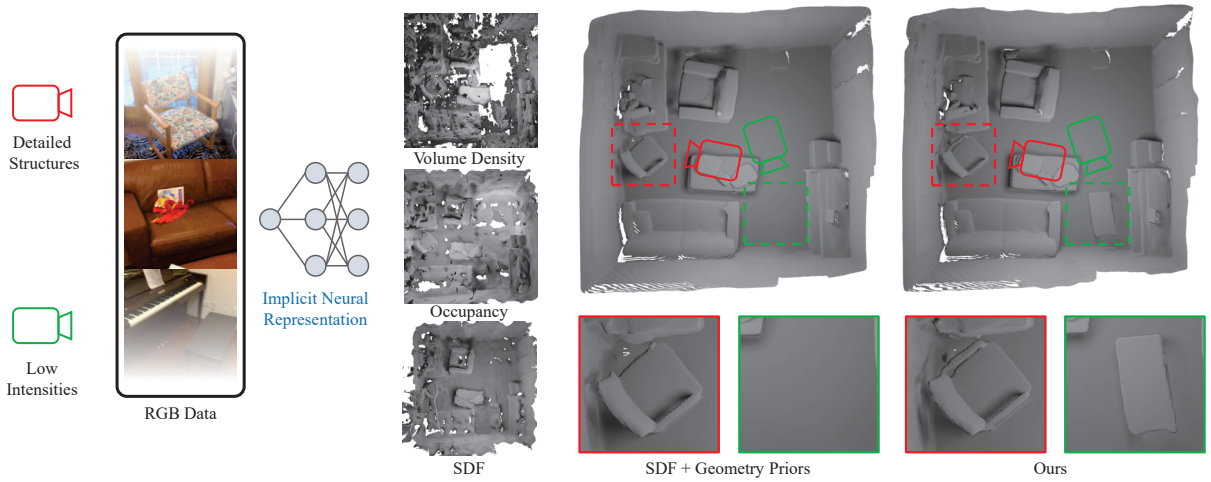


Figure 1. **Overview and reconstruction results of the Occ-SDF hybrid neural scene representation.** Aided with the feature rendering scheme (Sec. 4) and the hybrid representation (Sec. 5), our method yields results with more detailed structures in room-level scenes compared to the state-of-the-art, particularly for those low intensities and detailed structures.

Abstract

Implicit neural rendering, using signed distance function (SDF) representation with geometric priors like depth or surface normal, has made impressive strides in the surface reconstruction of large-scale scenes. However, applying this method to reconstruct a room-level scene from images may miss structures in low-intensity areas and/or small, thin objects. We have conducted experiments on three datasets to identify limitations of the original color rendering loss and priors-embedded SDF scene representation.

Our findings show that the color rendering loss creates an optimization bias against low-intensity areas, resulting in gradient vanishing and leaving these areas unoptimized. To address this issue, we propose a feature-based color rendering loss that utilizes non-zero feature values to bring back optimization signals. Additionally, the SDF representation can be influenced by objects along a ray path, disrupting the monotonic change of SDF values when a single object is present. Accordingly, we explore using the occupancy representation, which encodes each point separately and is unaffected by objects along a querying ray. Our experimental results demonstrate that the joint forces of the feature-based rendering loss and Occ-SDF hybrid representation scheme can provide high-quality reconstruction results, especially in challenging room-level scenarios. The code is available at https://github.com/shawLyu/Occ-SDF-Hybrid

1. Introduction

Reconstructing a 3D scene from a series of multi-view images is a crucial problem in the realm of computer vision. This process has widespread applications in various fields such as animation, gaming, and virtual/augmented reality (VR/AR). The recent trend is to represent a 3D scene as an implicit function parameterized by a neural network [13, 19, 26, 21], whose optimization is supervised by ex-

plicit 3D data like point cloud or real SDF value. Recent advancements in neural radiance field (NeRF) [14] further enable learning an implicit 3D representation from purely sparse posed images [35, 15].

However, when it comes to producing high-quality novel-view synthesis, these methods frequently utilize volume density [14] to represent the 3D geometry. Unfortunately, this approach does not adequately constrain the 3D geometry in the presence of ambiguities [18], ultimately leading to poor surface reconstructions (as depicted in Fig. 1: Volume Density).

Accordingly, research efforts have been made to exploit geometry-friendly representations, including signed distance function (SDF) [32, 34, 19] or occupancy [18], whose zero-level set can be extracted to become the concerned 3D surface. Albeit improving quality, they consider the reconstruction only of a single object, thereby, the performance degrades dramatically when applied to scene-level surface reconstruction, *i.e.*, representing a room (Fig. 1: SDF). An attribute is that reconstructing texture-less areas often suffers from ambiguous visual cues with only RGB loss as the regularization. To address this problem, recent research has attempted to incorporate semantic [8] or geometric priors (depth/normal [36, 31] constraints) to further regularize scene-level reconstruction. With SDF-based representation and geometric priors [36], the reconstruction quality has been greatly improved (Fig. 1: SDF + Geometry Priors), especially concerning large flat areas and objects. However, it still cannot faithfully reconstruct the 3D scene with missing structures in low-intensity dark areas and small/thin objects (Fig. 1: SDF + Geometry Priors).

The above observation motivates us to dive into bridging the remaining missing blocks of existing neural surface representation methods. Notably, we focus on the SDF-based representation as it achieves state-of-the-art performance and has been widely adopted. Our analysis suggests that both the RGB color rendering formulation and SDF representation have clear limitations preventing existing solutions from fully unleashing the potential of implicit neural surface representation for large-scale room-level scenes.

First, the color itself can show a significant impact on the optimization of geometric representation relying on the original RGB-based rendering formula [14], namely color bias. In particular, dark pixels with small intensity values will make the partial derivation of the loss with respect to the corresponding SDF value become zero, corrupting the optimization and resulting in missing structures in dark areas (see for example in Fig. 1: Low Intensities). Accordingly, herein instead of directly calculating the weighted color, we first compute weighted features and then use a learnable multi-layer perceptron (MLP) to decode the final rendering color. In such a way, we would still be able to effectively optimize the corresponding geometry representation as long as the feature vector contains non-zero values.

Second, the vanilla SDF-based neural rendering only considers a single ray directly passing through the object surface from the empty space and ignores objects along the ray [32, 34]. This configuration violates scene-level geometry where the existence of multiple objects clearly affects the distributions of SDF (Fig. 5(a)). Meanwhile, the optimization of thin structures and small objects, which naturally has small sampling probability, will be greatly degraded by this violation even with correct geometry prior and the structure will be erased to minimize the global geometry loss (Fig. 1: Detailed Structures).

In addition, although occupancy-based representations are likely to generate unwanted structure and cannot warrant a smooth surface reconstruction (Fig. 1), they are often sufficiently robust to objects along the ray and free from object interference in scene-level data. Therefore, during optimization, we propose to describe the room-level scene using occupancy in conjunction with signed distance functions (SDFs) to compensate for each other's defects.

The technical contributions are as follows:

- We explore an improved feature rendering scheme to overcome the problem of vanishing gradients in neural implicit reconstruction brought by the vanilla color space rendering formula.
- We carefully investigate insights and limitations in existing SDF and occupancy representations, and accordingly propose a hybrid representation mingling SDF with occupancy, dubbed **Occ-SDF Hybrid**, to resolve surfaces with thin structures and small objects.
- We conduct a large body of qualitative and quantitative experiments against state-of-the-art, indicating that our Occ-SDF hybrid formula can yield a higher-fidelity room-level scene representation, particularly with successfully resolving small and dark objects.

2. Related Work

Multi-view Stereo Conventional algorithms [1, 25, 4] always split the reconstruction into two steps. First, the feature-matching method [22, 17, 24] is applied to estimate the depth of each frame. Then, the resulting depth maps [12] are used to reconstruct the final scene. Notably, the reconstruction may suffer from poor performance in texture-less areas. Learning-based approaches are mainly divided into two categories. Typically, neural networks are embedded into the traditional reconstruction pipeline to replace specific modules, like feature matching [23, 37, 30], depth estimation [33], or depth fusion [5]. These methods often suffer from depth inconsistency problems due to the separately estimated depth maps. Alternatively, neural networks are designed to directly regress input images to trun-

cated signed distance functions (TSDFs) [16, 28], but the reconstruction results often lack enough fine details.

Neural Scene Representation Recently, coordinate-based neural representations can faithfully model a 3D scene with only posed images. Approaches with an implicit differentiable renderer [35] only use volume density as scene representation which can not extract 3D scenes directly. To address this issue, occupancy-based representation [18] and SDF-based representation [32, 34] are proposed to facilitate 3D reconstruction. Notably, these methods already achieve great performance for object-level scenes but exhibit poorly for room-level scenes, especially in texture-less areas. For the room-level scenes, several existing approaches [9, 39] have demonstrated the ability to employ learned shape priors derived from extensive data to reconstruct scenes from incomplete or noisy point clouds. However, these methods face limitations when it comes to reconstructing scenes solely from image data.

Priors for Indoor Scene Reconstruction Existing methods have attempted to introduce priors to resolve higher-fidelity surfaces in texture-less areas. Manhattan-SDF [8] follows semantic-NeRF[38] to estimate the volume density and semantic label at the same time, and then uses Manhattan-World assumption to regularize the geometry in floor and wall regions. NeuRIS [31] and MonoSDF [36] directly exploit the depth and normal predicted from an off-the-shelf neural network to regularize the geometry of each point, but in many cases, thin structures would disappear. NeuRIS [31] proposes a dynamic scheme to eliminate the wrong supervision signal from inaccurate estimated results, however, based on our investigation that fine structures are still lost even with the correct geometry for supervision. In all, we seek to explore a feature rendering scheme and a hybrid representation to overcome the above problems.

3. Overview and Preliminary

Our goal is to examine the *limitations* of existing implicit neural surface representations and explore practical *solutions* for accurately reconstructing large-scale, room-level 3D geometry with fine details from a set of calibrated images. First, we find that the well-adopted color-based rendering formula in [36, 31] will induce optimization bias against low-intensity areas, leaving these areas under-optimized and resulting in missing reconstructions (Sec. 4.1). Accordingly, we propose a simple yet effective feature-based rendering formula to address the problem (Sec. 4.2). Second, our analysis shows that the SDF-based neural surface representation violates scene-level geometry supervised signal and thus prevents the model from obtaining accurate reconstructions, making the model tends to sacrifice small and thin structures (Sec. 5.2). Motivated by this, we propose a hybrid representation mingling occupancy and SDF for accurate reconstruction (Sec. 5.3).

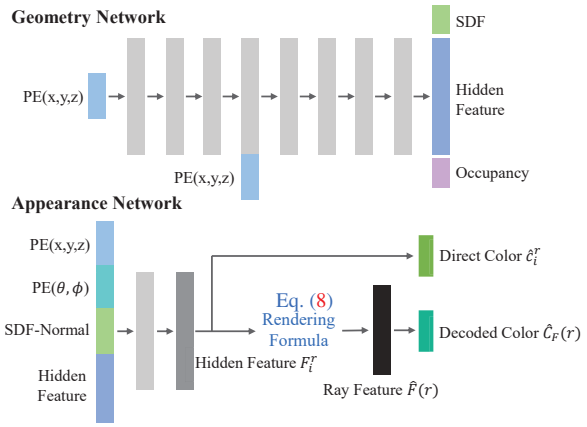


Figure 2. **Network architecture.** The geometry network takes 3D position (x, y, z) after positional encoding(PE) as input and output both SDF and occupancy value. The appearance network takes view direction (θ, ϕ) as input and outputs two types of color, the direct color is used in Eq. (4) to directly obtain the pixel color and the decoded color is calculated via the rendering formula (Eq. (8)).

We here describe the mathematical preliminary on the state-of-the-art surface representation, namely SDF-based Neural Scene Representation [34], for 3D reconstruction.

For implicit neural reconstruction, we can represent the scene as a signed distance function (SDF) field, which is a continuous function f that calculates the distance between each point and its closest surface

$$\begin{aligned} 1_\Omega(\boldsymbol{p}) &= \begin{cases} 1 & \text{if } \boldsymbol{p} \in \Omega \\ 0 & \text{if } \boldsymbol{p} \notin \Omega \end{cases}, \\ f(\boldsymbol{p}) &= (-1)^{1_\Omega(\boldsymbol{p})} \min_{\boldsymbol{y} \in \mathcal{M}} \|\boldsymbol{p} - \boldsymbol{y}\| , \end{aligned} \tag{1}$$

where $1_\Omega(\boldsymbol{p})$ is an indicator function to represent whether the space at position p is occupied, $\mathcal{M} = \partial\Omega$ is the boundary surface of occupied space and $||\cdot||$ is the standard Euclidean 2-norm. Following the VolSDF [34], we use an MLP to represent the function f and convert the SDF value to Laplace density with the following function

$$\sigma_i(\boldsymbol{p}_i) = \alpha \Psi_\beta \left(-f(\boldsymbol{p}_i)\right) , \tag{2}$$

where $\alpha, \beta > 0$ are learnable parameters, and Ψ_β is the cumulative distribution function (CDF) with zero mean and the β scale is defined as

$$\Psi_\beta(s) = \begin{cases} \frac{1}{2} \exp\left(\frac{s}{\beta}\right) & , \text{ if } s \leq 0 \\ 1 - \frac{1}{2} \exp\left(-\frac{s}{\beta}\right) & , \text{ if } s > 0 \end{cases} . \tag{3}$$

Color Rendering Formula According to the rendering formula [14], the color for the current ray r is rendered by

$$\hat{C}(r) = \sum_{i=1}^{M} T_i^r \alpha_i \hat{c}_i^r , \tag{4}$$

where T_i^r and α_i represent the transmittance and alpha

value (a.k.a opacity), respectively, of sampled point. And M represent the number of the sampled point along the ray r. They can be computed by

$$T_i^r = \prod_{j=1}^{i-1}(1-\alpha_i),\ \alpha_i = 1-\exp(-\sigma_i^r \delta_i^r)\ , \tag{5}$$

where δ_r^i is the distance between adjacent sample points. Finally, given the rendered color $\hat{C}(r)$, the SDF field will be optimized from sparsely sampled images by minimizing the color-based rendering loss as

$$\mathcal{L}_{\text{rgb}} = \sum_{r\in R} ||\hat{C}(r) - C(r)||_1\ , \tag{6}$$

where $C(r)$ is the ground-truth color associated with the sampled ray r.

4. Feature Rendering Formula

4.1. Problem of Color-based Rendering

Given the SDF-based scene representation and the color-based rendering loss $\mathcal{L}_{\text{rgb}}$ in Sec. 3, we analyze the derivative of $\mathcal{L}_{\text{rgb}}$ to the opacity α_i of a point p_i. Note that for a single point p_i, as α_i^r are the same regardless of the rays, we thus omit its dependency on ray r and use α_i for simplicity. For a point $\boldsymbol{p}_i$, the derivative of the color loss function to its opacity α_i is

$$\frac{\partial \mathcal{L}_{rgb}}{\partial \alpha_i} = \pm\left(\prod_{j=1}^{i-1}(1-\alpha_j)c_i - \sum_{k=i+1}^{N} c_k\alpha_k \prod_{j=1,j\neq i}^{k-1}(1-\alpha_j)\right), \tag{7}$$

which indicates that when we optimize the SDF value of $\boldsymbol{p}_i$, the gradient is determined by the color of the current point $\boldsymbol{p}_i$ and points behind it (c_k and $k \in \{i+1, ..., N\}$), and opacity of all points on the entire ray except for the current point (α_j and $j \in \{1, 2, ...N\}$ & $j \neq i$). Notably, when processing a dark region, saying that the c_i approaches zero, the first term of Eq. (7) will be close to zero. Similarly, if points behind $\boldsymbol{p}_i$ have low opacity (α_k and $k \in \{i+1, ..., N\}$), the gradient with respect to the SDF value will be small, causing the vanishing problem in dark regions. More generally, the gradient of SDF values can be affected by the color itself, resulting in a biased optimization process that tends to favor high color intensities.

The above analysis is also supported by our experiments below. As shown in Fig. 3, we sample rays in the dark regions and the light regions separately and accordingly record the trend of gradient norms in these two regions during the optimization. In the beginning, the gradient norms from these two regions are similar; and the gradient norm in the dark region (red solid line) decreases as the number of training epochs increases, while the gradient norm in the light region remains stable (blue solid line), indicating the dark areas contribute much less to the optimization process. As the optimization process proceeds, if these points are predicted as dark colors ($c_i^r \to 0$), it would lead to the gradient reduction effects as analyzed in Eq. (7). The shadowed regions keep the same trend as the mean values, further affirming our earlier analysis. Note that the gradient of points in these areas will not equal zero due to the influence of other loss functions, like depth consistency loss.

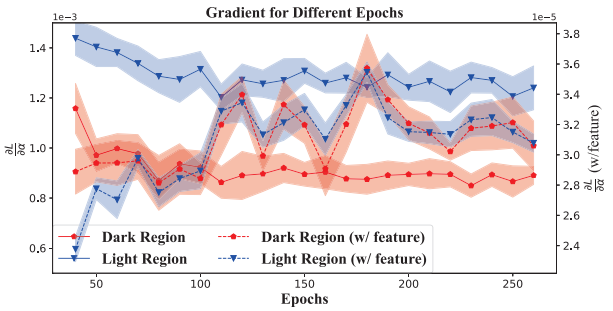


Figure 3. **The trend for the gradient.** The mean and variance of gradient norm (shadowed curves for variance) corresponding to light and dark regions during optimization.

4.2. Feature-based Rendering

To resolve the aforementioned problem, we propose feature-based color rendering loss. As shown in Fig. 2, the *Appearance* network outputs two predictions for each point i along a ray r: one is the color vector $\hat{c}_i^r$, and the other is the hidden feature F_i^r. For direct color $\hat{c}_i^r$, we utilize Eq. (4) to obtain the target pixel color $\hat{C}_c(r)$. And the hidden feature F_r^i is used to render the ray feature $\hat{F}(r)$ by

$$\hat{F}(r) = \sum_{i=1}^{M} T_i^r \alpha_i F_i^r\ . \tag{8}$$

The ray feature $\hat{F}_r$ is further decoded by a decoder $\mathcal{D}$ to yield the decoded target pixel color,

$$\hat{C}_F(r) = \mathcal{D}(\hat{F}(r))\ , \tag{9}$$

where the decoder $\mathcal{D}$ is a single-layer perceptron with 256 nodes. Finally, the decoded color $\hat{C}_F(r)$ from the rendered feature is used to construct the feature-based color rendering loss. As such, the optimization of these dark regions would not be affected by the color itself. As long as there are non-zero values in the rendered feature, there will be non-zero gradients with respect to the volume density of the concerned point. As shown in Fig. 3 (dashed lines), the gradient norm is not influenced by the intensity of colors.

5. Hybrid Representation Scheme

5.1. Incorporating Geometry Prior Matters

It is clear that for room-level scene reconstruction, geometry priors are essential. Existing methods [31, 36] render depth $\hat{D}(r)$ and normal $\hat{N}(r)$ of the surface intersecting

the current ray as

$$\hat{D}(r) = \sum_{i=1}^{M} T_i^r \alpha_i^r t_i^r \quad \text{and} \quad \hat{N}(r) = \sum_{i=1}^{M} T_i^r \alpha_i^r \hat{n}_i^r \ , \quad (10)$$

where $\hat{T}_r^i$ and $\hat{\alpha}_r^i$ have the same meaning as Eq. (4), t_i^r is the distance the ray passing and n_i^r is the normal of point p_i. Next, these methods use depth and normal maps estimated from pre-trained models, such as Omnidata [10], to directly supervise the rendered depth $\hat{D}(r)$ and normal $\hat{N}(r)$ using Eq. (11) and Eq. (12), respectively. Overall, the depth loss function is defined as

$$\mathcal{L}_{\text{depth}} = \sum_{r \in \mathcal{R}} ||(w\hat{D}(r) + q) - \bar{D}(r)||^2 \ , \quad (11)$$

where w and q are the scale and shift computed by the least-squares method [6] to solve scale-ambiguity problem for monocular depth prediction methods. And the normal loss function is

$$\mathcal{L}_{\text{normal}} = \sum_{r \in R} ||\hat{N}(r) - \bar{N}(r)||_1 + ||1 - \hat{N}(r)^T \bar{N}(r)||_1 \ , \quad (12)$$

where $\bar{N}(r)$ is the predicted monocular normal transformed to the same coordinate system with angular.

As shown in Fig. 1, we note that these geometry priors benefit the reconstruction of better surfaces in textureless and sparse-viewed areas. However, thin structures and small objects, such as the yellow flower in Fig. 4, cannot be faithfully reconstructed with geometry priors. Recently, NeuRIS [31] put forward a hypothesis that this phenomenon arises from the inaccurate geometry supervisory signal (*i.e.* depth and surface normal). However, according to our experiment on Replica synthetic dataset, this problem still exists even though we use the perfect ground-truth depth, normal, and RGB to provide supervisory signals (see Fig. 4).

5.2. Problem of SDF Formula with Geometry Prior

To dive into the SDF representation and explore its limitations for surface reconstruction with geometric priors, we create a simplified scenario and simulate object occlusions as shown in Fig. 5(a). The ground-truth SDF intersecting with a horizontal plane is shown in Fig. 5(a), and the SDF distribution along a ray r intersecting with the blue cube at point p_s is shown in Fig. 5(c). There are many local minima and maxima due to the existence of multiple objects, which differs from the single-object scenario following a monotonic function (blue line in Fig. 5(c)). To examine depth priors for surface reconstruction, we employ the approach proposed in MonoSDF [36] to calculate the depth of p_s following Eq. (10) subject to the ground-truth SDF.

However, even with the ground-truth SDF, the estimated depth value (1.59, the red vertical plot in Fig. 5(e)) still deviates from the true depth value (2.94, the green vertical plot in Fig. 5(e)) when multiple objects exist. According to Eq. (10), the estimated normal value would suffer from the same problem. This implies that existing methods incorporating geometry priors [36, 31] to guide the learning of the SDF representation may not necessarily encourage the model to learn the true SDF for scene-level surface reconstruction. In turn, because small objects or thin structures usually have low sampling probability during training, the minimization of $\mathcal{L}_{\text{depth}}$ will encourage the model to predict SDF ignoring small objects along the ray r such that the estimated SDF will produce depth values closer to the depth supervision (see Fig. 5(e)) and minimize the overall loss function, attempting to mimic the single object scenario (right part in Fig. 5). In sum, the supervision from geometric priors tends to sacrifice the reconstruction of small objects to preserve large surface reconstruction, which aligns also with our observation presented in Fig. 4.

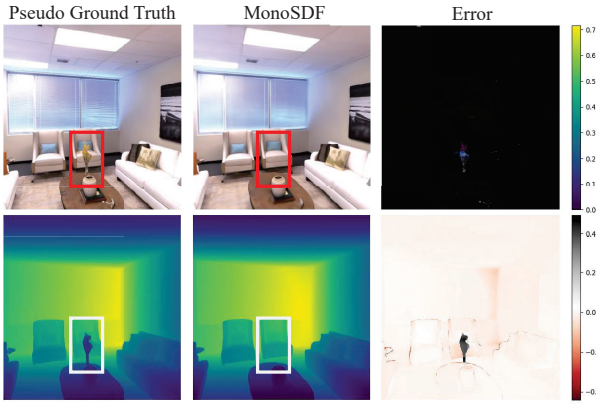


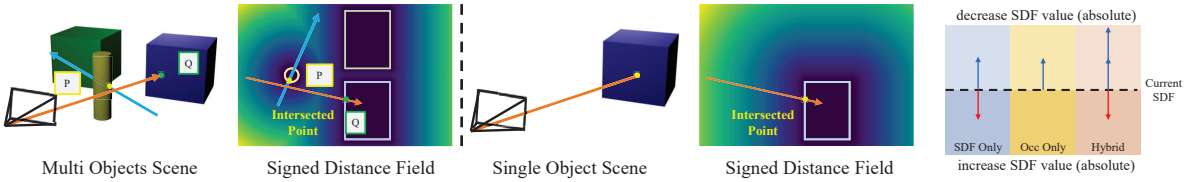
Figure 4. **Illustration of failure cases of state-of-the-arts.** Even though applying the perfect pseudo ground-truth geometry to supervise the model, existing room-level reconstruction methods like [36] can still fail to resolve accurate 3D structures.

5.3. Hybrid Occupancy-SDF Representation

The problem above is essentially caused by the SDF representation, which describes the geometry of a scene as a whole and thus suffers from the interference of other objects, especially small objects, and thin structures are prone to be removed to preserve large structures.

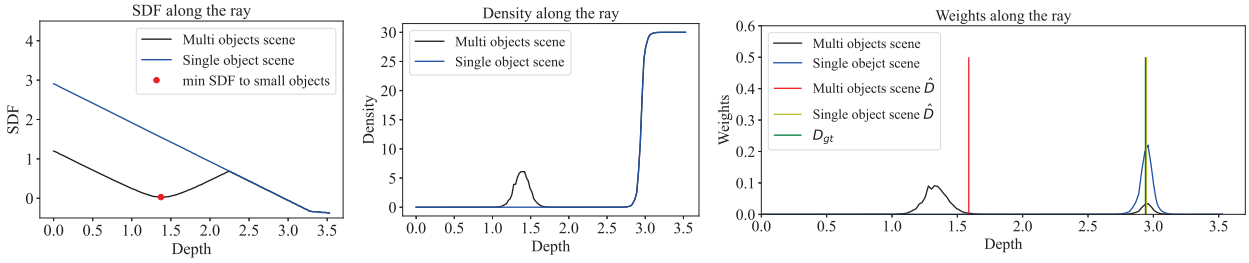
Unlike the SDF representation, occupancy represents each point separately and thus is free from the interference of objects in this challenging scenario. However, occupancy representation only focuses on the intersecting point ignoring the constraint of neighborhood points. Thus, the reconstruction results represented by occupancy will have many floaters and useless structures, as shown in Fig. 1, which can be eliminated in SDF by Eikonal loss [7]. This inspires us to investigate a hybrid of occupancy and SDF as a representation for neural surface reconstruction as shown in Fig. 2. The geometry network θ outputs both SDF and occupancy.

Specifically, the occupancy represents surfaces as the de-



(a) Schematic of the toy 3D structure for the multi-object scene (left) and single object scene (object).

(b) SDF Value at Point P



(c) SDF $f(\boldsymbol{p}_i)$

(d) Laplace density $\delta_i(\boldsymbol{p}_i)$

(e) Weights $T_r^i \alpha_r^i$

Figure 5. **Toy experiments for the room-level scene.** The left part of 5(a), where we stand in front of the yellow cylinder to observe the entire scene, is a widespread scenario for room-level scale scenes. Unlike the single object scenario, where the distribution of SDF value is a monotonic decreasing function from the observed position to the object, the room-level scenario has complex distributions with multi peaks/valleys along the single ray (5(c)). Following the Laplace density function [34], the density distributions of different situations are shown in 5(d), where room-level scenes have a secondary peak near small objects but the single object scene only has one peak. It is because of the existence of this peak, the weights in the room-level scene (5(e)) exhibit a multi-model distribution, while for the single object case a uni-modal distribution. As such, we note that the rendering depth $\hat{D}$ deviates from the ground truth in the room-level scene but is close to the ground truth depth object-level scene. 5(b) means the effect of supervised signal in three different representations.

cision boundary of a binary occupancy classifier, parameterized by a neural network θ

$$o_\theta(\mathbf{p}) : \mathbb{R}^3 \rightarrow [0, 1] \ , \tag{13}$$

where $\mathbf{p}$ is a 3D point. The occupancy representation assumes that objects are solid, thus we can rewrite the neural rendering formula [14] to

$$\hat{C}(r) = \sum_{i=1}^{M} o(\mathbf{p}_i) \prod_{j<i} (1 - o(\mathbf{x}_j)) c(\mathbf{p}_i, d) \ , \tag{14}$$

which replaces the opacity α to a discrete occupancy indicator variable $o \in 0, 1$, where $o = 0$ indicates the free space while $o = 1$ the occupied space. Thus, this representation will not be affected by the objects along the ray, and the rendering depth will be consistent with ground-truth depth in ground-truth occupancy space. Note that the occupancy-based representation is introduced to facilitate optimization. During inference, the SDF is used for reconstruction.

To understand why and how the hybrid representation helps optimize the SDF field for accurate reconstruction, we conduct the following empirical analysis, using the scenario shown in Fig. 5(a). Here, the orange ray hits a surface point Q of the large blue cube and the blue ray hits the surface point P on the small cylinder. During optimization, the depth/normal loss for point Q along the orange ray will encourage the model to predict a large SDF value (absolute) of point P (Fig. 5(b)) which violates the reconstruction of the small cylinder where a small SDF value is desired. In contrast, point Q has no effects on point P with occupancy representation. The hybrid representation joins the forces of SDF and occupancy representations, aiming to use occupancy representation to help overcome the issues of SDF representation in optimization. Although the depth/normal loss from the SDF presentation for point Q still has a negative impact on the optimization of point P. The additional occupancy representation will force the network to predict a large occupancy value for a point P and thus will indirectly regularize the network to predict a small SDF value (see Fig. 5(b): the blue up arrow in "Hybrid"). We admit that this hybrid representation can only alleviate this problem, and our study is more empirical. Fundamental issues arise from insufficient neural scene representation, which requires further research efforts. We explore further why this combination would bring notable benefits to the supplementary with an example.

6. Experiments

Optimization. In the training stage, we minimize the loss

$$\begin{aligned} \mathcal{L} = & \mathcal{L}_{\text{rgb}}^{sdf} + \lambda_1 \mathcal{L}_{\text{rgb}}^{sdf_F} + \lambda_2 \mathcal{L}_{\text{eik}} + \lambda_3 \mathcal{L}_{\text{depth}}^{\text{occ}} \\ & + \lambda_4 \mathcal{L}_{\text{depth}}^{\text{sdf}} + \lambda_5 \mathcal{L}_{\text{normal}}^{occ} + \lambda_6 \mathcal{L}_{\text{normal}}^{sdf}, \end{aligned} \tag{15}$$

where $\mathcal{L}_{rgb}$ means color-based rendering loss following Eq. (6) but only to SDF representation. The rendering color computed by the feature rendering formula (Eq. (8)

and Eq. (9)) is denoted to $\mathcal{L}_{rgb}^{sdf_F}$. Notably, $\mathcal{L}_{eik}$ means the eikonal loss [7], $\mathcal{L}_{depth}$ means the depth rendering loss following Eq. (11), and $\mathcal{L}_{normal}$ means the normal rendering loss following Eq. (12). We apply them for both representations, where the superscript ^{occ} indicates the loss computed by occupancy-based representation, while the ^{sdf} by SDF-based representation. The network is optimized by the Adam optimizer with a learning rate of $5e^{-4}$. We set weights $\lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5, \lambda_6$ to $1, 0.05, 0.5, 0.1, 0.1, 0.05$, respectively. The network architecture and sampling strategy are detailed in the supplement.

Datasets. We use three datasets to assess the performance of our algorithm. **ScanNet** [2] is a real-world dataset that provides 1,513 scenes captured with Kinect V1 RGB-D camera. The BundleFusion [3] is applied to provide high-quality camera poses and surface reconstructions. For each scene, we uniformly sample roughly 500 frames to train our network. **Tanks and Temples** [11] is a real-world, large-scale scene dataset. We use four indoor scenes from their advanced split and run on the official server. **Replica** [27] is a synthetic dataset that provides 18 scenes, with each providing dense geometry, HDR textures, and semantic annotations. We select 8 scenes and use the Habitat simulator [29] to render RGB images following MonoSDF [36] splits. Notably, we conduct ablation studies on this dataset.

Compared Methods. (1) **UNISURF** [18] is an occupancy-based method that unifies surface rendering and volume rendering for neural scene reconstruction. We implement the **UNISURF*** with normal and depth priors for a fair comparison. (2) **MonoSDF** [36] is an SDF-based method that adds depth and normal constraints on VolSDF [34]. (3) **Manhattan-SDF** [8] is an SDF-based method that adds a semantic branch and uses the Manhattan constraint to regularize the geometry in floor and wall regions. (4) **COLMAP** [25] is a classical multi-view stereo method with Poisson surface reconstruction. (5) **NeuRIS** [31] is an SDF-based method that introduces pseudo normal prior to the NeUS [32] architecture. Meanwhile, it leverages multi-view consistency to eliminate the wrong supervision signal from inaccurate estimated results. (6) **NICER-SLAM** [39] is an SDF-based dense SLAM system that uses locally implicit map representation and can simultaneously optimize for camera poses and a hierarchical neural implicit map representation. (7) **LIG** [9] uses the local implicit grid representation to reconstruct the large-scale scene from partial or noise point clouds. (8) **Convolutional Occupancy network(Conv-Occ)** [21] is a locally implicit representation that integrates local information to get better reconstruction results from noisy point cloud.

Notably, local implicit representation [9, 21] can only reconstruct the scene from point clouds, thus we re-implement them using point clouds generated from scale-aligned pseudo depth and utilize the provided pretrained models for evaluation. And we directly obtain the results from the main paper of NICER-SLAM [39].

Metrics. All meshes are evaluated by 5 standard metrics defined in [16]: *Accuracy*, *Completeness*, *Precision*, *Recall*, and *F-score*. Their definition will be discussed in the supplementaryary. For the Replica dataset, we also report the normal consistency following [13, 8, 20]. For the Tanks and Temples dataset, we use the official server to evaluate our results and report the F-score for selected scenes.

6.1. Main Results

We compare our method with state-of-the-art methods on three benchmark datasets.

Method	Acc ↓	Comp ↓	C-$\mathcal{L}_1$ ↓	Prec ↑	Recall ↑	F-score ↑
COLMAP [25]	0.047	0.235	0.141	71.1	44.1	53.7
UNISURF [18]	0.554	0.164	0.359	21.2	36.2	26.7
VolSDF [34]	0.414	0.120	0.267	32.1	39.4	34.6
NeUS [32]	0.179	0.208	0.194	31.3	27.5	29.1
Manhattan-SDF [8]	0.072	0.068	0.070	62.1	56.8	60.2
NeuRIS [31]	0.050	0.049	0.050	71.7	66.9	69.2
MonoSDF [36]	**0.035**	0.048	0.042	79.9	68.1	73.3
Ours	0.039	**0.041**	**0.040**	**80.0**	**76.0**	**77.9**

Table 1. Quantitative assessments of the proposed model against previous works on the ScanNet dataset.

Results on ScanNet Dataset. We conducted a comparative analysis of our proposed approach against existing implicit reconstruction methods, including Manhattan-SDF [8], NeuRIS [31] and MonoSDF [36] using the ScanNet dataset. As revealed in Table 1, our proposed method outperforms state-of-the-art methods, with a significant increase in F-score by **4.6**. Additionally, in terms of Recall, our method substantially outperforms MonoSDF by **7.9** without needing extra data. Overall, our approach performs on par with the state-of-art methods in "Acc", "Chamfer-$\mathcal{L}_1$(C-$\mathcal{L}_1$)" and "Prec" and obtain notable performance gains in "Comp", "Recall" and "F-score". This is because these metrics ("Comp" and "Recall") are better metrics in evaluating how complete and accurate in capturing the shape and details of the scene being reconstructed. Further, Fig. 6 reveals that our method can attain more complete reconstructions with details and for low pixel intensities regions.

	Auditorium	Ballroom	Courtroom	Museum	Mean
MonoSDF	3.09	2.47	10.00	5.10	5.165
Ours	**5.22**	**5.42**	**13.99**	**8.59**	**8.305**
MonoSDF*	3.17	3.70	13.75	5.68	6.58
Ours*	**6.19**	**7.33**	**19.80**	**11.85**	**11.295**

Table 2. Quantitative assessments of the proposed model against Monosdf on the Tanks and Temples dataset. The evaluation metrics for the Tanks and Temples dataset are F-score. * means that the hash-grid structure is adopted.

Results on Tanks and Temples Dataset. For challenging large-scale indoor scenes, we conduct experiments on

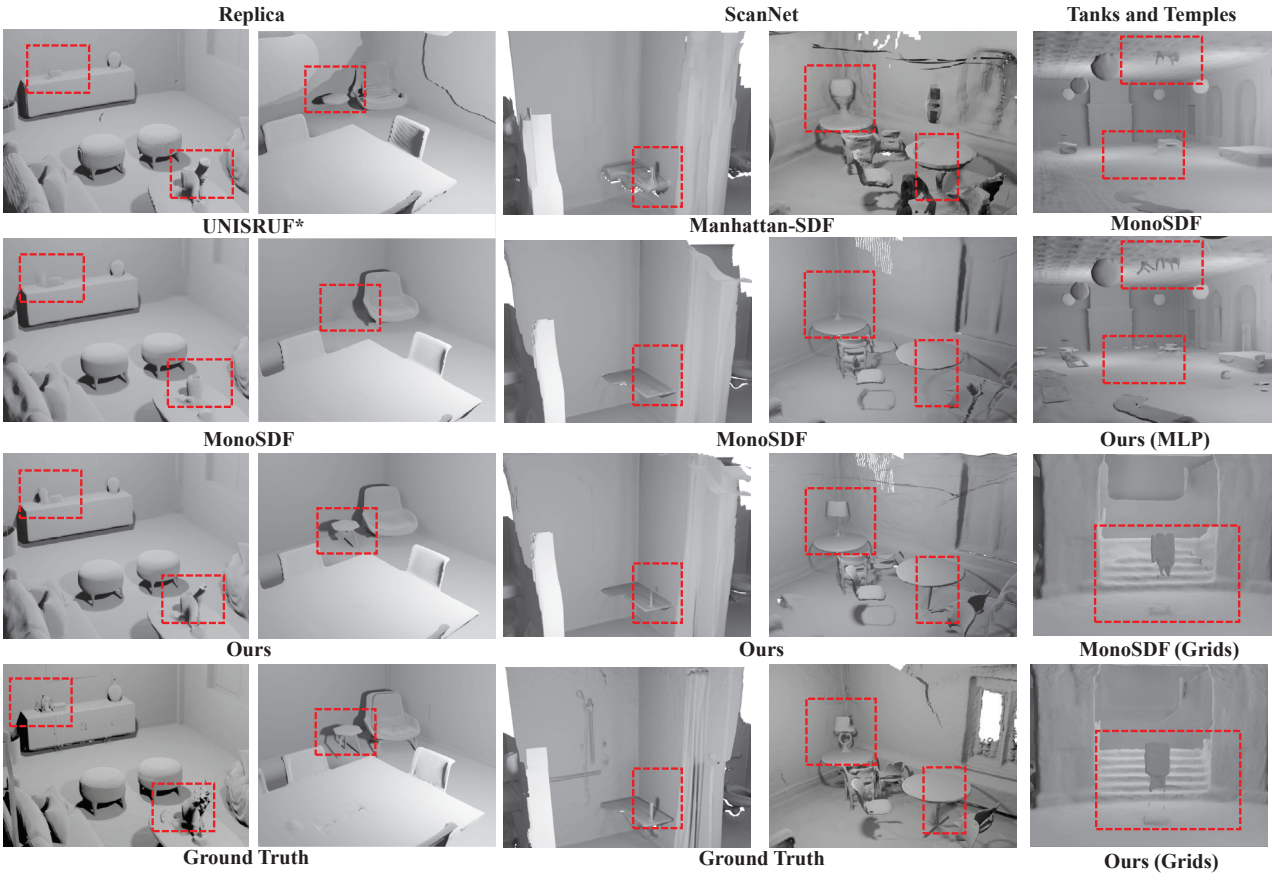


Figure 6. **Reconstruction results on representative datasets** of Replica (left), ScanNet (middle), and Tanks and Temples (right). The ground truth is presented on the bottom-most. Red boxes in sub-figures highlight those areas where distinctive differences can be observed.

the advanced Tanks and Temples dataset [11], which features more complicated structures. As alternative methods of neural reconstruction from images are not assessed on this dataset, we implement the best-performing method MonoSDF and compare with it. Thus, the MonoSDF and two versions of our method are implemented. Specifically, one adopts the pure MLP architecture while the other uses the hash grids as the input representation. The quantitative assessments (Table 2) reveal that our method shows better performance on this dataset, regardless of whether an MLP or hash-grid structure is used. And our hybrid representation exhibits excellent generalization abilities across different implicit structures. Overall, our compelling experimental results on the Tanks and Temples dataset further validate the robustness and versatility of the proposed method in reconstructing complex and challenging indoor scenes.

Results on Replica Dataset. Quantitative assessment results on the Replica dataset are presented in Table 3. For this dataset, we compare our methods with both point-based methods [21, 9] and rendering methods [39, 18, 36]. Ours significantly surpasses existing state-of-the-art neural rendering methods. The results reveal that the SDF-based representation outperforms the occupancy-based ones (*i.e.* UNISURF*). This is because the SDF usually enforces constraints on the distribution of the entire scene, benefiting to suppressing the occurrence of floaters or unnecessary structures in occupancy-based representation. Notably, our Occ-SDF Hybrid method can constrain the distribution of the entire scene with SDF representation meanwhile exploiting the occupancy representation to resolve thin structures and small objects. Qualitative comparisons are shown in Fig. 6.

Method	Normal C.↑	Chamfer-$\mathcal{L}_1$ ↓	F-score ↑
Conv-Occ [21]	85.73	6.43	58.33
LIG [9]	89.56	5.53	65.20
NICER-SLAM [39]	90.27	3.91	-
UNISURF† [18]	90.96	4.93	78.99
MonoSDF† [36]	92.11	2.94	86.18
Ours†	**93.43**	**2.58**	**92.12**

Table 3. Quantitative assessments of the proposed model against prior works on the Replica dataset. Herein, † indicates the use of geometry priors as supervision signals.

6.2. Ablation Study

	Normal C.↑	Chamfer-$\mathcal{L}_1$ ↓	F-score ↑
MonoSDF	92.11	2.94	86.18
+ feature	93.01	2.64	91.01
+ hybrid	93.22	2.77	90.24
full model	**93.43**	**2.58**	**92.12**

Table 4. Ablation study on the Replica dataset [27], where we progressively add different constraints to assess their impacts. The MonoSDF [36] is set as the baseline model.

We conduct ablation studies on the Replica dataset as it provides ground-truth geometry. Four different configurations are investigated to train our model, including (1) MonoSDF with MLP settings (MonoSDF-MLP); (2) MonoSDF-MLP with our feature-based rendering formula; (3) MonoSDF-MLP with our hybrid representation; (4) MonoSDF-MLP with both the feature-based rendering formula and hybrid representation scheme (Full model).

Table 4 shows that all metrics are improved when using the feature rendering to reconstruct this scene. Our proposed feature rendering scheme addresses the difficulties in reconstructing areas of low intensities, resulting in better results. On the other hand, the hybrid representation also leads to significant improvements in all metrics. Notably, it improves the completeness of small objects and thin structures, as evidenced by the results in Fig. 6. By leveraging both components, our model achieves an overall improvement of **5.94** in F-score, along with improved normal consistency and Chamfer-$\mathcal{L}_1$. We attribute this success to our feature-based color rendering formula and our hybrid representation, which addresses the color-bias issue in optimization and difficulties in reconstructing detailed structures. The visualization results in Fig. 6 show our model's excellent reconstruction performance, especially in low-intensity areas and detailed structures. We will add more ablation studies and visualize results in the supplementary.

6.3. Parameters Adjusting

As shown in Eq. (15), our method newly added three different losses $\mathcal{L}_{rgb}^{sdf_F}, \mathcal{L}_{depth}^{occ}, \mathcal{L}_{normal}^{occ}$. In order to verify the sensitivity of our method to hyperparameters, we provided the results of the experiment on the Replica dataset as shown in Fig. 7. It is clear that our method is not very sensitive to hyperparameters, and all the evaluated settings outperform the baseline method (**C-$\mathcal{L}_1$: 2.94, F-score: 86.18**). It is worth noting that our approach demonstrates consistent performance enhancement across diverse datasets using the same parameter set, outperforming the baseline results.

7. Conclusion

We have analyzed the constraints present in current neural scene representation techniques with geometry priors,

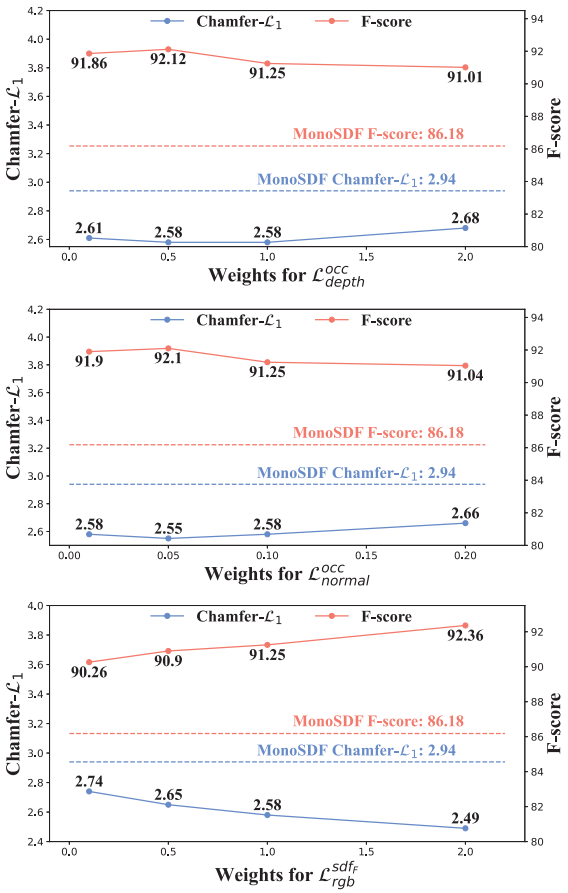


Figure 7. Performance with different hyperparameter choices.

and have identified issues in their ability to reconstruct detailed structures due to a biased optimization towards high color intensities and the complex SDF distribution. As a result, we have developed a feature rendering scheme that balances color regions and have implemented a hybrid representation to address the limitations of the SDF distribution. Our approach has demonstrated the successful reconstruction of room scenes with a high-fidelity surface, including small objects, detailed structures, and low-intensity pixel regions. We envision our results inspire further research on improving neural scene representation for accurate and large-scale surface reconstruction.

Acknowledgement

We thank Xin Kong for his advising during the rebuttal period. This work has been supported by Hong Kong Research Grant Council - Early Career Scheme (Grant No. 27209621 and 27212822), General Research Fund Scheme (Grant No. 17202422), and RGC matching fund scheme (RMGS). Part of the described research work is conducted in the JC STEM Lab of Robotics for Soft Materials funded by The Hong Kong Jockey Club Charities Trust.

References

[1] Motilal Agrawal and Larry S Davis. A probabilistic framework for surface reconstruction from multiple images. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, volume 2, pages II–II. IEEE, 2001. 2

[2] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5828–5839, 2017. 7

[3] Angela Dai, Matthias Nießner, Michael Zollhöfer, Shahram Izadi, and Christian Theobalt. Bundlefusion: Real-time globally consistent 3d reconstruction using on-the-fly surface reintegration. *ACM Transactions on Graphics (ToG)*, 36(4):1, 2017. 7

[4] Frank Dellaert, Steven M Seitz, Charles E Thorpe, and Sebastian Thrun. Structure from motion without correspondence. In *Proceedings IEEE Conference on Computer Vision and Pattern Recognition. CVPR 2000 (Cat. No. PR00662)*, volume 2, pages 557–564. IEEE, 2000. 2

[5] Simon Donne and Andreas Geiger. Learning non-volumetric depth fusion using successive reprojections. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7634–7643, 2019. 2

[6] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network. *Advances in neural information processing systems*, 27, 2014. 5

[7] Amos Gropp, Lior Yariv, Niv Haim, Matan Atzmon, and Yaron Lipman. Implicit geometric regularization for learning shapes. In *International Conference on Machine Learning*, pages 3789–3799. PMLR, 2020. 5, 7

[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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5511–5520, 2022. 2, 3, 7

[9] Chiyu Max Jiang, Avneesh Sud, Ameesh Makadia, Jingwei Huang, Matthias Nießner, and Thomas Funkhouser. Local implicit grid representations for 3d scenes. In *Proceedings IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2020. 3, 7, 8

[10] Oğuzhan Fatih Kar, Teresa Yeo, Andrei Atanov, and Amir Zamir. 3d common corruptions and data augmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18963–18974, 2022. 5

[11] Arno Knapitsch, Jaesik Park, Qian-Yi Zhou, and Vladlen Koltun. Tanks and temples: Benchmarking large-scale scene reconstruction. *ACM Transactions on Graphics*, 36(4), 2017. 7, 8

[12] Paul Merrell, Amir Akbarzadeh, Liang Wang, Philippos Mordohai, Jan-Michael Frahm, Ruigang Yang, David Nistér, and Marc Pollefeys. Real-time visibility-based fusion of depth maps. In *2007 IEEE 11th International Conference on Computer Vision*, pages 1–8. Ieee, 2007. 2

[13] Lars Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 1, 7

[14] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 2, 3, 6

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

[16] Zak Murez, Tarrence van As, James Bartolozzi, Ayan Sinha, Vijay Badrinarayanan, and Andrew Rabinovich. Atlas: End-to-end 3d scene reconstruction from posed images. In *European conference on computer vision*, pages 414–431. Springer, 2020. 3, 7

[17] Pauline C Ng and Steven Henikoff. Sift: Predicting amino acid changes that affect protein function. *Nucleic acids research*, 31(13):3812–3814, 2003. 2

[18] Michael Oechsle, Songyou Peng, and Andreas Geiger. Unisurf: Unifying neural implicit surfaces and radiance fields for multi-view reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5589–5599, 2021. 2, 3, 7, 8

[19] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 165–174, 2019. 1, 2

[20] Songyou Peng, Chiyu "Max" Jiang, Yiyi Liao, Michael Niemeyer, Marc Pollefeys, and Andreas Geiger. Shape as points: A differentiable poisson solver. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 7

[21] Songyou Peng, Michael Niemeyer, Lars Mescheder, Marc Pollefeys, and Andreas Geiger. Convolutional occupancy networks. In *European Conference on Computer Vision*, pages 523–540. Springer, 2020. 1, 7, 8

[22] Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary Bradski. Orb: An efficient alternative to sift or surf. In *2011 International conference on computer vision*, pages 2564–2571. Ieee, 2011. 2

[23] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. Superglue: Learning feature matching with graph neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4938–4947, 2020. 2

[24] Johannes L Schönberger, Enliang Zheng, Jan-Michael Frahm, and Marc Pollefeys. Pixelwise view selection for unstructured multi-view stereo. In *European conference on computer vision*, pages 501–518. Springer, 2016. 2

[25] Johannes Lutz Schönberger, Enliang Zheng, Marc Pollefeys, and Jan-Michael Frahm. Pixelwise View Selection for Unstructured Multi-View Stereo. In *European Conference on Computer Vision (ECCV)*, 2016. 2, 7

[26] Vincent Sitzmann, Eric Chan, Richard Tucker, Noah Snavely, and Gordon Wetzstein. Metasdf: Meta-learning signed distance functions. *Advances in Neural Information Processing Systems*, 33:10136–10147, 2020. 1

[27] 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, 9

[28] Jiaming Sun, Yiming Xie, Linghao Chen, Xiaowei Zhou, and Hujun Bao. Neuralrecon: Real-time coherent 3d reconstruction from monocular video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15598–15607, 2021. 3

[29] Andrew Szot, Alex Clegg, Eric Undersander, Erik Wijmans, Yili Zhao, John Turner, Noah Maestre, Mustafa Mukadam, Devendra Chaplot, Oleksandr Maksymets, Aaron Gokaslan, Vladimir Vondrus, Sameer Dharur, Franziska Meier, Wojciech Galuba, Angel Chang, Zsolt Kira, Vladlen Koltun, Jitendra Malik, Manolis Savva, and Dhruv Batra. Habitat 2.0: Training home assistants to rearrange their habitat. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 7

[30] Benjamin Ummenhofer, Huizhong Zhou, Jonas Uhrig, Nikolaus Mayer, Eddy Ilg, Alexey Dosovitskiy, and Thomas Brox. Demon: Depth and motion network for learning monocular stereo. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5038–5047, 2017. 2

[31] Jiepeng Wang, Peng Wang, Xiaoxiao Long, Christian Theobalt, Taku Komura, Lingjie Liu, and Wenping Wang. Neuris: Neural reconstruction of indoor scenes using normal priors. *arXiv preprint arXiv:2206.13597*, 2022. 2, 3, 4, 5, 7

[32] 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. *arXiv preprint arXiv:2106.10689*, 2021. 2, 3, 7

[33] Yao Yao, Zixin Luo, Shiwei Li, Tian Fang, and Long Quan. Mvsnet: Depth inference for unstructured multi-view stereo. In *Proceedings of the European conference on computer vision (ECCV)*, pages 767–783, 2018. 2

[34] Lior Yariv, Jiatao Gu, Yoni Kasten, and Yaron Lipman. Volume rendering of neural implicit surfaces. *Advances in Neural Information Processing Systems*, 34:4805–4815, 2021. 2, 3, 6, 7

[35] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelnerf: Neural radiance fields from one or few images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4578–4587, 2021. 2, 3

[36] Zehao Yu, Songyou Peng, Michael Niemeyer, Torsten Sattler, and Andreas Geiger. Monosdf: Exploring monocular geometric cues for neural implicit surface reconstruction. *arXiv preprint arXiv:2206.00665*, 2022. 2, 3, 4, 5, 7, 8, 9

[37] Sergey Zagoruyko and Nikos Komodakis. Learning to compare image patches via convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4353–4361, 2015. 2

[38] Shuaifeng Zhi, Tristan Laidlow, Stefan Leutenegger, and Andrew Davison. In-place scene labelling and understanding with implicit scene representation. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2021. 3

[39] Zihan Zhu, Songyou Peng, Viktor Larsson, Zhaopeng Cui, Martin R. Oswald, Andreas Geiger, and Marc Pollefeys. Nicer-slam: Neural implicit scene encoding for rgb slam, 2023. 3, 7, 8