UNIVERSITY OF AMSTERDAM

UvA-DARE (Digital Academic Repository)

Ray-Distance Volume Rendering for Neural Scene Reconstruction

Yin, R.; Chen, Y.; Karaoglu, S.; Gevers, T.

DOI
10.1007/978-3-031-72630-9_22

Publication date
2025

Document Version
Final published version

Published in
Computer Vision – ECCV 2024

License
Article 25fa Dutch Copyright Act (https://www.openaccess.nl/en/policies/open-access-in-dutch-copyright-law-taverne-amendment)

Link to publication

Citation for published version (APA):
Yin, R., Chen, Y., Karaoglu, S., & Gevers, T. (2025). Ray-Distance Volume Rendering for Neural Scene Reconstruction. In A. Leonardis, E. Ricci, S. Roth, O. Russakovsky, T. Sattler, & G. Varol (Eds.), *Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024 : proceedings* (Vol. XIV, pp. 377–394). (Lecture Notes in Computer Science; Vol. 15072). Springer. https://doi.org/10.1007/978-3-031-72630-9_22

General rights
It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use, unless the work is under an open content license (like Creative Commons).

Disclaimer/Complaints regulations
If you believe that digital publication of certain material infringes any of your rights or (privacy) interests, please let the Library know, stating your reasons. In case of a legitimate complaint, the Library will make the material inaccessible and/or remove it from the website. Please Ask the Library: https://uba.uva.nl/en/contact, or a letter to: Library of the University of Amsterdam, Secretariat, P.O. Box 19185, 1000 GD Amsterdam, The Netherlands. You will be contacted as soon as possible.

UvA-DARE is a service provided by the library of the University of Amsterdam (https://dare.uva.nl)

Download date:14 Nov 2026

Ray-Distance Volume Rendering for Neural Scene Reconstruction

Ruihong Yin[1(✉)], Yunlu Chen[2], Sezer Karaoglu[1,3], and Theo Gevers[1,3]

[1] University of Amsterdam, Amsterdam, The Netherlands
{r.yin,Th.Gevers}@uva.nl
[2] Carnegie Mellon University, Pittsburgh, USA
yunluche@andrew.cmu.edu
[3] 3DUniversum, Amsterdam, The Netherlands
s.karaoglu@3duniversum.com

Abstract. Existing methods in neural scene reconstruction utilize the Signed Distance Function (SDF) to model the density function. However, in indoor scenes, the density computed from the SDF for a sampled point may not consistently reflect its real importance in volume rendering, often due to the influence of neighboring objects. To tackle this issue, our work proposes a novel approach for indoor scene reconstruction, which instead parameterizes the density function with the Signed Ray Distance Function (SRDF). Firstly, the SRDF is predicted by the network and transformed to a ray-conditioned density function for volume rendering. We argue that the ray-specific SRDF only considers the surface along the camera ray, from which the derived density function is more consistent to the real occupancy than that from the SDF. Secondly, although SRDF and SDF represent different aspects of scene geometries, their values should share the same sign indicating the underlying spatial occupancy. Therefore, this work introduces a SRDF-SDF consistency loss to constrain the signs of the SRDF and SDF outputs. Thirdly, this work proposes a self-supervised visibility task, introducing the physical visibility geometry to the reconstruction task. The visibility task combines prior from predicted SRDF and SDF as pseudo labels, and contributes to generating more accurate 3D geometry. Our method implemented with different representations has been validated on indoor datasets, achieving improved performance in both reconstruction and view synthesis.

Keywords: Indoor scene reconstruction · neural radiance fields · signed ray distance function

1 Introduction

Indoor 3D scene reconstruction involves using multi-view *RGB* images as input to generate detailed 3D geometry as output. This task is crucial and has

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-031-72630-9_22.

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2025
A. Leonardis et al. (Eds.): ECCV 2024, LNCS 15072, pp. 377–394, 2025.
https://doi.org/10.1007/978-3-031-72630-9_22

widespread applications in virtual/mixed augmentation, robotics, navigation, and so on. Contrary to a single object, indoor environments contain various elements, such as chairs, walls, tables, and cups, which present a challenge for the task of reconstruction. Traditional reconstruction approaches [1,21,27,33,40,44] use 2D and 3D convolutional neural networks (CNNs) [34,35] to extract features. They can be classified as depth-based methods [6,27,37,40] and volumetric methods [1,21,31,33,44]. Depth-based techniques estimate depth for every viewpoint, encountering difficulties like scale uncertainties and inconsistencies in depth across different views. The surfaces created through these techniques can be coarse and imprecise. Unlike depth-based approaches, volumetric techniques utilize 3D CNNs to directly generate 3D geometry, leading to smoother surfaces. Nevertheless, these methods use voxels for scene representation, causing GPU memory usage and computational expenses to grow cubically as voxel resolution increases. Such a limitation restricts the use of volumetric approaches for high-quality scene reconstruction. Moreover, volumetric techniques require 3D annotations for network training, significantly increasing the cost of annotations.

Implicit representations [18,24,29] encode 3D models as functions of coordinates. Utilizing implicit representations and volume rendering, neural radiance fields (NeRF) [19] achieves remarkable performance in novel view synthesis, even without 3D supervision. Some methods [38,42] exploit the capabilities of NeRF in reconstruction tasks. These methods typically parameterize the volume density as a learnable transformation of the Signed Distance Function (SDF), leading to impressive reconstruction performance. Compared to classic volumetric methods, neural implicit scene representation is more memory-efficient, marking a new trend in reconstruction tasks.

In multi-object indoor scenes, the SDF from sampled points considers surface points across the entire scene. However, for neural implicit scene reconstruction, along a camera ray, from the camera center to the intersected surface point, the SDF may exhibit fluctuations, leading to several local minimum values. Figure 1 provides a toy example to illustrate this. In Fig. 1c, volume density yielded from the fluctuating SDF displays several local maximum values, contributing to high weights in Fig. 1d. However, the camera ray $\overrightarrow{OP}$ targets the green rectangle, while the SDF of the near-surface point $\mathbf{Q}$ refers to the surface of the blue cylinder. Besides, $\mathbf{Q}$ is far from the intersected point $\mathbf{P}$, but still corresponds to high density and high weight. Nevertheless, during volume rendering that maps the 3D output to 2D space, it is evident that along a ray, points closer to the intersected surface point may carry a higher weight as they are more relevant to actual observations. Consequently, in the multi-object scenes, using the SDF to model the density function along the camera ray may disrupt this relationship, introduce noise, and lead to inaccuracies in 2D results or 3D geometries.

This paper proposes **RS-Recon**, a novel indoor scene reconstruction method with the ray-specific density function to address the aforementioned issue. Firstly, our method models the density function as a function of the Signed Ray Distance Function (SRDF). Unlike SDF, SRDF measures the shortest distance to the surface along a camera ray, eliminating the influence of surrounding

surfaces that are not located on the given ray. Moreover, the density distribution generated from SRDF attains the local maximum only around the intersected surface, aligning more closely to the concept that points nearer to the surface have more importance. To achieve this, our network predicts both SRDF and SDF, where SRDF encompasses the density function, while SDF mainly focuses on describing the 3D surfaces. Secondly, despite SRDF and SDF being defined differently, both of them produce positive values outside the object and negative values within. Therefore, in this paper, a SRDF-SDF consistency loss is devised to ensure they share the same sign. Thirdly, this work introduces a self-supervised visibility task to enhance the prediction of 3D geometry. The visibility task predicts the probability of whether the sampled points are physically visible along the ray in 3D space. The pseudo visibility ground truth is formed by combining the prior of SRDF and SDF from the network itself, without the reliance on multi-view geometry or any additional annotations.

Our contributions can be summarized as follows, **(1)** This work proposes a novel neural scene reconstruction method that leverages the ray-specific SRDF to model the volume density. Compared to SDF, SRDF better reflects the significance of the real observations. **(2)** A SRDF-SDF consistency loss is proposed to constrain the alignment between SRDF and SDF, ensuring the same sign inside and outside the object. **(3)** To enhance the accuracy of 3D geometry, an additional self-supervised visibility task is introduced to predict the visibility probability of sampled points. This task integrates prior information from both the SDF and SRDF predictions as labels. **(4)** The experiments conducted on both synthetic and real-world indoor datasets illustrate that our method enhances performance in both reconstruction and view synthesis.

2 Related Works

Classic Indoor Scene Reconstruction. Traditional methods for indoor scene reconstruction fall into two classes, *i.e.* depth-based methods [6,16,27,37,40,41,45] and volumetric methods [1,7,12,21,31,33,44]. Depth-based methods predict pixel-level depth for individual frames, and subsequently depth fusion [4] is adopted to generate the 3D scene. To achieve favorable outcomes, these methods often construct a plane sweep cost volume [2,9] at the feature or image level, leveraging information from multiple views to complement the current frame, *e.g.* MVSNet [40], DeepVideoMVS [6], and SimpleRecon [27]. Using the depth map as the intermediate representation, these methods encounter challenges such as depth inconsistency and scale ambiguities. In contrast, volumetric methods utilize 3D CNN to directly regress the 3D geometry, which can generate smooth surfaces and reconstruct unobserved regions. For instance, NeuralRecon [33] designs a learnable TSDF fusion module to integrate features from previous frames and predicts the TSDF for sparse volumes. TransformerFusion [1], based on volumetric representation, adopts Transformer to fuse multi-view features and select frames with higher attention. Yin *et al.* [44] propose to incorporate geometry priors at different levels for the volumetric methods. However, due to the use of

3D CNN in volumetric methods, computational costs escalate significantly with higher resolutions.

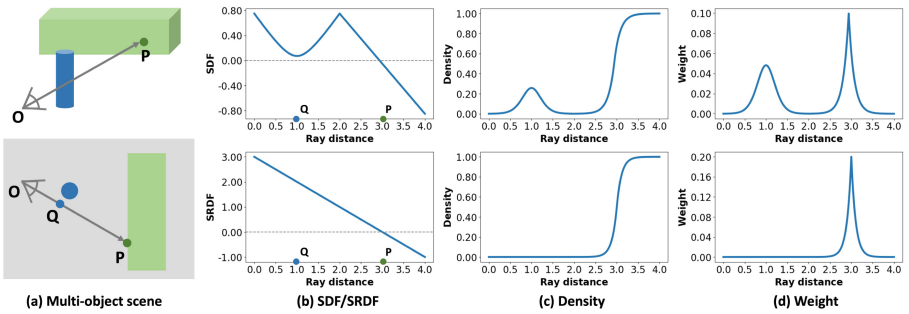

Fig. 1. Analysis of the density function and weight distribution for a toy scene. (a) Frontal and overhead perspectives of a scene with multiple objects, where a ray originates from the camera center **O**, intersecting with the green rectangle at the point **P**. **Q** is the closest point to the blue cylinder along the ray $\overrightarrow{OP}$. (b) The SDF d_Ω /SRDF $\tilde{d}_\Omega$ in Eqs. (2) and (4) along the ray $\overrightarrow{OP}$. (c) The volume density σ generated from the SDF/SRDF in (b) using Eqs. (3) and (5). (d) The weight distribution $T_i\alpha_i$ generated from the density in (c) via Eq. 1. Although **Q** is distant from the surface boundary intersecting with the ray, the density function from the SDF in **Q** generates a high weight in volume rendering, resulting in noisy rendering and reconstruction. (Color figure online)

Neural Implicit Scene Representation. With the success of implicit representation and NeRF [19], neural implicit scene representation has emerged as a hot spot in the past few years. Generally, reconstruction approaches based on neural implicit scene representation first map point positions to a continuous representation of SDF or occupancy by multi-layer perceptrons (MLPs), then transfer SDF or occupancy to volume density, and finally render the 2D results from the 3D outputs. Despite requiring only 2D supervision during training, these methods can reconstruct high-quality surfaces and are more computationally efficient. For example, VolSDF [42] parameterizes the volume density as a Laplace distribution of SDF, while NeuS [38] models it as the logistic distribution. Some extensions [10,36,39,46,47] introduce extra priors to achieve superior outcomes. For instance, Manhattan-SDF [10] employs planar constraints based on Manhattan-world assumption [3]. NeuRIS [36] proposes to combine normal prior for the reconstruction of texture-less regions, while NeuralRoom [39] implements perturbation-residual restriction to reconstruct the flat region better. MonoSDF [46] exploits both depth and normal cues to improve the reconstruction quality. Although additional priors contribute to the accuracy and completeness of reconstructed surfaces, they employ the SDF to model the volume density, potentially resulting in noisy 2D results in the multi-object scenes. Con-

trastingly, our method advocates for parameterizing the density function with the ray-specific SRDF, resulting in enhanced reconstruction performance.

Some NeRF-based methods also utilize multi-view geometry to introduce additional regularization. For instance, RegNeRF [22] regularizes patched-based depth maps from unobserved viewpoints. Geo-NeuS [8] and ConsistentNeRF [11] use multi-view photometric consistency. SRDF and observation consistencies are used by [48]. GenS [25] introduces multi-view and multi-scale feature consistencies for neural surface reconstruction. VIP-NeRF [30] adopts a visibility prior, using the plane sweep volume among multiple views to generate visibility labels. However, they depend on the computation of multi-view geometry, some of which may involve processing multiple frames yielding increased computational demands. In contrast, this paper introduces a self-supervised visibility task that requires minimal computational increases and does not depend on multi-view geometry for ground truth generation.

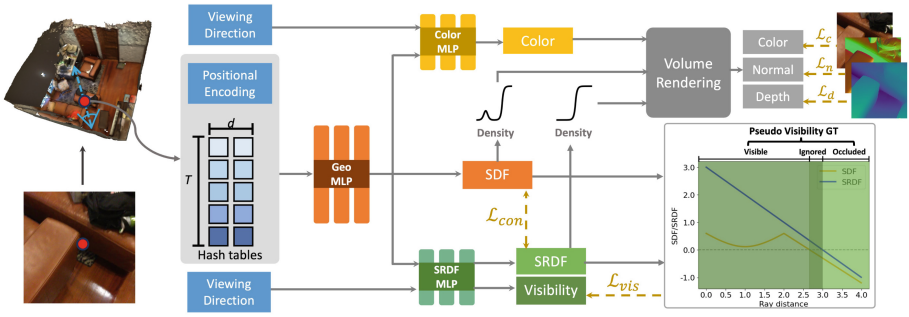

Fig. 2. Our framework. A geometry MLP f_g is adopted to generate the SDF and geometry features $\mathbf{F}_g$, utilizing the encoded position (and optionally grid features) as input. Then, $\mathbf{F}_g$, along with viewing direction, passes through a color MLP f_c to predict the color for each point. Notably, (1) our approach advocates for modeling the density function with the ray-specific SRDF in addition to the SDF. For this purpose, a SRDF MLP f_s is introduced to generate the SRDF. (2) A SRDF-SDF consistency loss $\mathcal{L}_{con}$ is devised to align the signs between the generated SRDF and SDF. (3) To enhance the geometry prediction, a self-supervised visibility task is proposed to integrate geometry priors in both SRDF and SDF predicted by the network and generates the pseudo visibility ground truth. The visibility probability is predicted by the SRDF MLP.

3 Method

In this section, we provide the background on NeRF-based reconstruction methods, which typically model the volume density as a transformation of SDF (refer to Sect. 3.1). To overcome the issue of false local maxima in the density from SDF, our approach covered in Sect. 3.2–Sect. 3.5 proposes to model the density function with the ray-specific SRDF in neural scene reconstruction, which

achieves more accurate reconstruction. Figure 2 presents the framework of our approach.

3.1 Background: Volume Rendering Density as Transformed SDF

Our goal is to recover the scene geometry given a set of projected 2D views of a 3D scene. To achieve this, NeRF [19] optimizes a continuous 5D function to represent the scene, which predicts the volume density $\sigma(\mathbf{p})$ and view-dependent color $\mathbf{c}(\mathbf{p}, \mathbf{r})$ for each sampled point $\mathbf{p} \in \mathbb{R}$ and ray direction $\mathbf{r} \in \mathcal{S}^2$. With $\Omega \subset \mathbb{R}^3$ denoting the scene geometry filled with physical presence, the predicted density $\sigma(\mathbf{p})$ approximates the ground truth scene density $\mathbf{1}_\Omega(\mathbf{p})$, such that $\mathbf{1}_\Omega(\mathbf{p}) = 1$ if $\mathbf{p} \in \Omega$ and 0 otherwise. After that, the classic volume rendering is employed to render the color $\mathbf{C}(\mathbf{r})$, and N points $\{\mathbf{p}_i = \mathbf{o} + z_i\mathbf{r}, i \in [1, N]\}$ are sampled along the camera ray $\mathbf{r}$, with z_i indicating the depth from the camera center $\mathbf{o}$ to the sampled point. The color is accumulated along the ray:

$$\mathbf{C}(\mathbf{r}) = \sum_{i=1}^{N} T_i \alpha_i \mathbf{c}(\mathbf{p}_i, \mathbf{r}), T_i = \prod_{j=1}^{i-1} (1 - \alpha_i), \alpha_i = 1 - \exp(-\sigma(\mathbf{p}_i)\delta_i) \tag{1}$$

where δ_i is the interval of neighboring points.

Despite the great success in novel view synthesis, NeRF has difficulty reconstructing satisfactory actual surfaces from the generated volume density. To address this challenge, recent methods [42,46] propose to predict the SDF and then transform it to volume density, which aligns with the geometry bias that the surface points have a higher density than other non-surface points and corresponds to a higher weight during volume rendering.

For any sampled point $\mathbf{p} \in \mathbb{R}^3$, the SDF $d_\Omega(\mathbf{p})$ is with an absolute value representing the shortest distance from $\mathbf{p}$ to the surface $\mathcal{M} = \partial\Omega$, while the sign indicates whether the point is outside (positive) or inside (negative) the surface. Formally,

$$\text{SDF:} \quad d_\Omega(\mathbf{p}) = (-1)^{\mathbf{1}_\Omega(\mathbf{p})} \min_{\mathbf{p}^* \in \mathcal{M}} ||\mathbf{p} - \mathbf{p}^*||_2, \tag{2}$$

where the indicator function $\mathbf{1}_\Omega(\mathbf{p}) = 1$ if $\mathbf{p} \in \Omega$ and 0 otherwise. This indicates that the SDF considers every surface around the point. To apply SDF to volume rendering, Yariv *et al.* [42] suggest to derive the volume density σ^{SDF} from the signed distance as follows:

$$\sigma^{\text{SDF}}(\mathbf{p}) = \alpha \Psi_\beta(-d_\Omega(\mathbf{p})), \Psi_\beta(s) = \begin{cases} \frac{1}{2}\exp\left(\frac{s}{\beta}\right) & \text{if } s \leq 0 \\ 1 - \frac{1}{2}\exp\left(-\frac{s}{\beta}\right) & \text{if } s > 0 \end{cases} \tag{3}$$

in which α, β are learnable parameters. Ψ_β is the Cumulative Distribution Function (CDF) of the Laplace distribution with zero mean and β scale. Then, $\sigma^{\text{SDF}}(\mathbf{p})$ is applied to Eq. (1) and as the density function σ.

3.2 Accurate Scene Volume Density from Ray-Specific SRDF

We illustrate the issues with volume rendering SDF by a toy example in Fig. 1. Indoor scenes typically include multiple objects. Along the camera ray $\mathbf{r}$, the SDF may be influenced by surfaces not intersecting the ray, resulting in several ambiguous local maxima of false high volume density and weight. This implies that the 2D observations may have a strong correlation to these distant points, contradicting the fact that the points closer to the surface along the camera ray may carry more significance.

To this end, we propose volume rendering scene density function from SRDF, the distance field local to the ray direction. In contrast to the SDF in Eq. (2), the SRDF $\widetilde{d_\Omega}(\mathbf{p}, \mathbf{r})$ [26,48] computes the shortest distance from the point to the surface along the sampled ray $\mathbf{r}$, *i.e.* the SRDF is ray-dependent. Formally, the SRDF $\widetilde{d_\Omega}(\mathbf{p}, \mathbf{r})$ can be represented by

$$\text{SRDF:} \quad \widetilde{d_\Omega}(\mathbf{p}, \mathbf{r}) = (-1)^{\mathbf{1}_\Omega(\mathbf{p})} \min_{(\mathbf{p}+\rho\mathbf{r})\in\mathcal{M};\rho\in\mathbb{R}} |\rho|, \tag{4}$$

where $\mathbf{1}_\Omega(\mathbf{p}) = 1$ if $\mathbf{p} \in \Omega$ and 0 otherwise. In the following sections, d_Ω from Eq. (2) and $\widetilde{d_\Omega}$ are used to denote SDF and SRDF for brevity.

In Fig. 1, it can be seen that the weight computed from SRDF in the volume rendering only reaches a local peak around the surface point, aligning more consistently with the nature of the 2D observations. Therefore, this paper proposes to employ the ray-specific SRDF to yield the volume density instead of the SDF. This work predicts both SDF and SRDF, where SDF is mainly used to model the surface. As depicted in Fig. 2, for a sampled point $\mathbf{p}$, a geometry MLP f_g is applied on the input (including the encoded position or features) to predict SDF d_Ω and geometry feature $\mathbf{F}_g$. By definition, SRDF is view-dependent. Hence, the viewing direction $\mathbf{r}$, geometry feature $\mathbf{F}_g$, and point position $\mathbf{p}$ are concatenated and then passed through a SRDF MLP f_s to yield the SRDF. On the other hand, the viewing direction $\mathbf{r}$, geometry feature $\mathbf{F}_g$, point position $\mathbf{p}$, and 3D unit normal $\mathbf{n} = \nabla d_\Omega$ are used as input to a color MLP f_c to output color $\mathbf{c}$.

For volume rendering, our method adopts the predicted SRDF $\widetilde{d_\Omega}$ to derive the volume density σ. This is defined by:

$$\sigma^{\text{SRDF}}(\mathbf{p}, \mathbf{r}) = \alpha\Psi_\beta\left(-\widetilde{d_\Omega}(\mathbf{p}, \mathbf{r})\right). \tag{5}$$

Then, $\sigma^{\text{SRDF}}(\mathbf{p}, \mathbf{r})$ is applied to the volume rendering in Eq. (1) as a ray-conditioned density function to obtain the color value $\mathbf{C}^{\text{SRDF}}(\mathbf{r})$. Besides, during training, we also render color $\mathbf{C}^{\text{SDF}}(\mathbf{r})$ with density functions from SDF to obtain gradient signals to optimize the prediction of SDF, in which the volume rendering processes use the same rays and points as those for $\mathbf{C}^{\text{SRDF}}(\mathbf{r})$.

For the scene geometry, after the network is trained with NeRF optimization targets, the geometric surface $\mathcal{M} = \partial\Omega$ is extracted as watertight meshes by the marching cubes algorithm [15,24] from the SDF results d_Ω, which only depends on the query location $\mathbf{p}$.

3.3 SRDF-SDF Consistency Loss

Following the definitions in Eq. (4), despite SRDF and SDF being defined differently, they share the same sign, which indicates the spatial occupancy - positive outside the surface and negative inside the surface. However, in this paper, where SDF and SRDF are predicted from separate branches, sign consistency is not guaranteed. To deal with this issue, a SRDF-SDF consistency loss is proposed to enforce this constraint.

Although the sign function effectively computes the output's sign, it cannot be differentiated during back-propagation. To a certain degree, the values derived from the sigmoid function approximate those calculated by the sign function. Thus, in our SRDF-SDF consistency loss, the sigmoid function is first employed to generate the sign of SRDF and SDF as the approximated sign function $\varsigma(\cdot)$, after which a ℓ_2 loss is adopted to quantify the sign difference, as defined by:

$$\begin{aligned} \mathcal{L}_{con} &= \frac{1}{N_r} \sum_{\mathbf{p},\mathbf{r}} M_{con} ||\varsigma(\widetilde{d_\Omega}) - \varsigma(d_\Omega)||_2 \\ \varsigma(d) &= \texttt{Sigmoid}(k \cdot d), M_{con} = [\widetilde{d_\Omega} \cdot d_\Omega < 0] \end{aligned} \tag{6}$$

in which k is a hyperparameter, controlling the slope of the sigmoid function. The indicator function $[\widetilde{d_\Omega} \cdot d_\Omega < 0]$ is leveraged to identify whether the generated SDF and SRDF have opposite signs. Therefore, the consistency loss imposes penalties when there is a difference between the signs of SRDF and SDF. N_r represents the number of sampled points in a minibatch that satisfy $[\widetilde{d_\Omega} \cdot d_\Omega < 0]$.

Taking the derivative of $\mathcal{L}_{con}$ with respect to SDF d_Ω (as an example), is computed as follows:

$$\frac{\partial \mathcal{L}_{con}}{\partial d_\Omega} = -\frac{2}{N_r} M_{con} ||\varsigma(\widetilde{d_\Omega}) - \varsigma(d_\Omega)||_1 \cdot \frac{\partial \varsigma(d_\Omega)}{\partial d_\Omega} \tag{7}$$

The derivative in Eq. (7) shows that our consistency loss has two advantages: (1) Unlike the sign function, which assigns the same penalty for outputs with different signs of SRDF and SDF, our consistency loss can adjust the penalty according to the extent of the sign discrepancy between SRDF and SDF, referring to $||\varsigma(\widetilde{d_\Omega}) - \varsigma(d_\Omega)||_1$. (2) The derivative of the sigmoid function $\frac{\partial \varsigma(d_\Omega)}{\partial d_\Omega}$ peaks at zero and decreases from 0 to $+\infty/-\infty$. Consequently, predictions in proximity to zero (indicating the surface) may exhibit higher absolute gradient values, providing effective supervision for points near the surface.

3.4 Self-supervised Visibility Task

Theoretically, with visibility attributes of 3D points, surface location becomes straightforward. Therefore, to enhance 3D geometry prediction, we propose a self-supervised visibility task in this paper. Along the trajectory of a camera ray, points sampled prior to reaching the first surface point are categorized as visible, whereas points sampled beyond the first surface point are classified as occluded.

To find the first surface point along the ray, using SRDF as an example, the multiplication of SRDF between adjacent sampled points is computed. When $\widetilde{d_\Omega}^{\,i} \cdot \widetilde{d_\Omega}^{\,i+1} \leq 0$ and $\widetilde{d_\Omega}^{\,m} \cdot \widetilde{d_\Omega}^{\,m+1} > 0, m \in [1, i-1]$, it indicates that the first surface is within the interval $[\mathbf{o}+z_i\mathbf{r}, \mathbf{o}+z_{i+1}\mathbf{r}]$. Consequently, the points between $[\mathbf{p}_1, \mathbf{p}_i]$ are considered physically visible while points between $[\mathbf{p}_{i+1}, \mathbf{p}_N]$ are regarded as occluded. However, the SRDF or SDF predicted by the network may be inaccurate at some points, introducing potential noise in the visibility task. To improve the accuracy of the visibility labeling, both SDF and SRDF are utilized in determining visibility. Points are labeled as visible or occluded only when there is agreement between SRDF and SDF on their classification. If the classifications from SRDF and SDF diverge, suggesting difficulties in accurately categorizing these points for the network, such points are omitted from the visibility task. The method for assigning visibility labels is defined as follows

$$V_{gt} = \begin{cases} 1, & \text{if } (V^{\text{SRDF}} = 1) \& (V^{\text{SDF}} = 1) \\ 0, & \text{if } (V^{\text{SRDF}} = 0) \& (V^{\text{SDF}} = 0) \end{cases} \tag{8}$$

where $V^{\text{SRDF}}/V^{\text{SDF}}$ are the visibility labels determined based on the prediction of SRDF/SDF. 1/0 means label for visible/occluded points. The formation of the visibility label can be inferred as self-supervised, thus eliminating the necessity for multi-view geometry and additional annotations. Additionally, it incorporates information from both SRDF and SDF.

From the analysis above, similar to SRDF, the visibility task also relies on the view direction. Therefore, a visibility probability V_{pred} is predicted from the SRDF branch. It is a binary classification task, *i.e.* visible or occluded, so the binary cross-entropy loss is employed as visibility loss $\mathcal{L}_{vis}$ during optimization.

The visibility task can discern the point's visibility along the camera ray, thereby influencing the learning of SRDF. Moreover, the optimization process integrates knowledge of ray-related visibility across the entire space, further impacting the learning of SDF. Notably, unlike 2D supervision, the visibility task provides priors in 3D space, imparting a more direct influence on 3D geometry.

3.5 Optimization

Loss Functions. During training, the overall loss consists of RGB loss $\mathcal{L}_c$, normal loss $\mathcal{L}_n$, depth loss $\mathcal{L}_d$, Eikonal loss $\mathcal{L}_e$, smooth loss $\mathcal{L}_s$, SRDF-SDF consistency loss $\mathcal{L}_{con}$, and visibility loss $\mathcal{L}_{vis}$,

$$\mathcal{L} = \mathcal{L}_c + \lambda_n\mathcal{L}_n + \lambda_d\mathcal{L}_d + \lambda_e\mathcal{L}_e + \lambda_s\mathcal{L}_s + \lambda_{con}\mathcal{L}_{con} + \lambda_{vis}\mathcal{L}_{vis} \tag{9}$$

in which $\lambda_{(\cdot)}$ are the loss weight. Following MonoSDF [46], the 2.5D depth and normal maps are volume rendered using the density function σ^{SRDF} to exploit geometric cues and support training. More details on the computation of loss functions are provided in the supplementary material.

Implementation Details. Theoretically, our method can be applied to replace SDF-based volume rendering in indoor scene reconstruction. To measure our

Table 1. Evaluation of 3D reconstruction meshes on ScanNet. † indicates our reproduced results. The best results are **bold**. ↓ denotes that a lower value is preferable while ↑ means higher value corresponds to better performance.

Method	Acc ↓	Comp ↓	Prec ↑	Recall ↑	F-score ↑
COLMAP [28]	0.047	0.235	0.711	0.441	0.537
UNISURF [23]	0.554	0.164	0.212	0.362	0.267
NeuS [38]	0.179	0.208	0.313	0.275	0.291
VolSDF [42]	0.414	0.120	0.321	0.394	0.346
Manhattan-SDF [10]	0.072	0.068	0.621	0.586	0.602
S3P [43]	0.055	0.059	0.709	0.660	0.683
NeuRIS [36]	0.050	0.049	0.717	0.669	0.692
MonoSDF_Grid [46]	0.072	0.057	0.660	0.601	0.626
MonoSDF_MLP [46]	**0.035**	0.048	0.799	0.681	0.733
HelixSurf [14]	0.038	0.044	0.786	0.727	0.755
Occ_SDF_Hybrid† [17]	0.040	0.041	0.783	0.748	0.765
Ours_Grid	0.074	0.049	0.670	0.703	0.683
Ours_MLP	0.040	**0.040**	**0.809**	**0.779**	**0.794**

method, VolSDF-based reconstruction method *MonoSDF* is considered as our baseline in the following experiments. Furthermore, in the supplementary material, we also verify our method on NeuS-based reconstruction method *NeuRIS*. We used different spatial encoding techniques as input, including hash features [20] and Fourier features [19], referred to as *Grid* and *MLP* respectively in the following experiments. More details are given in the supplementary material.

4 Experiments

4.1 Experimental Setup

Datasets. Our proposed method is evaluated on three indoor datasets, comprising: (1) four scenes from the real-world dataset ScanNet [5], (2) eight scenes from the synthetic dataset Replica [32], and (3) four scenes from the advanced set of the real-world large-scale dataset Tanks and Temples [13].

Metrics. For ScanNet, the assessment of the results involves five reconstruction metrics: accuracy (acc), completeness (comp), precision (prec), recall, and F-score with a threshold of 5cm. For Replica, in alignment with MonoSDF, metrics including normal consistency (normal c.), chamfer L1 distance (Chamfer-L1), and F-score with a threshold of 5cm are utilized. For Tanks and Temples, as evaluated by the official server, results are reported based on the F-score with a 1cm threshold. Additionally, this study assesses the peak signal-to-noise ratio (PSNR) of the rendered 2D *RGBs* to showcase the effectiveness of our approach. Detailed computations for all metrics are given in the supplementary material.

Baselines. Except for the baseline MonoSDF, our method is also compared to (1) the classic MVS method: COLMAP [28], (2) neural implicit reconstruction methods: UNISURF [23], NeuS [38], VolSDF [42], Manhattan-SDF [10], NeuRIS [36], S3P [43], HelixSurf [14], and Occ_SDF_Hybrid [17].

4.2 Evaluation Results

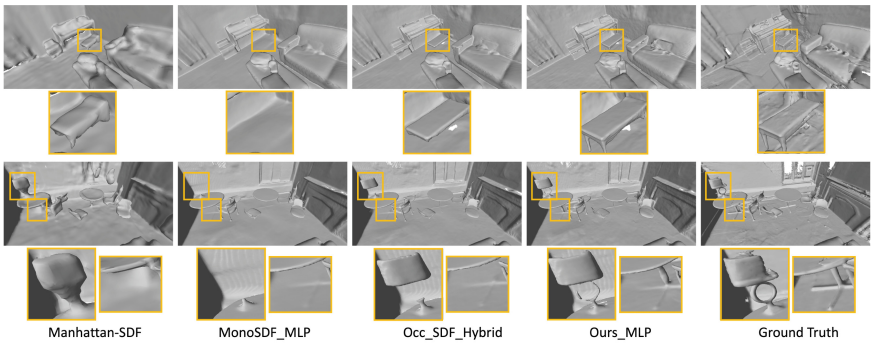

Fig. 3. Qualitative comparisons on ScanNet. It can be seen that our method can reconstruct more surfaces, especially in thin regions.

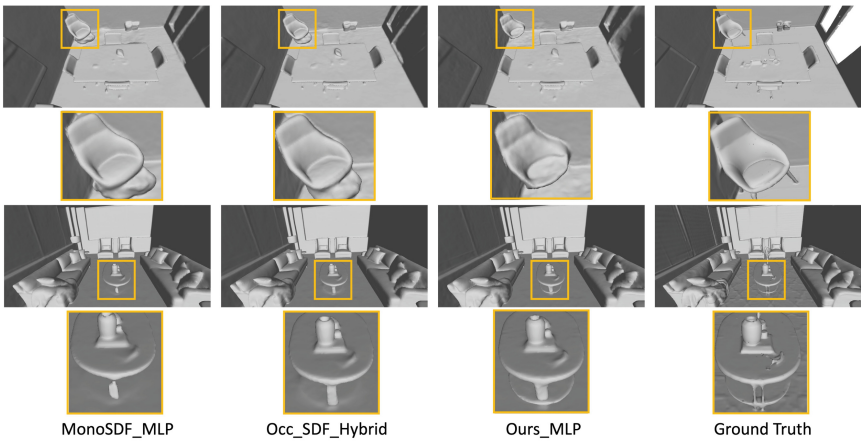

Fig. 4. Qualitative comparisons on Replica. Compared to MonoSDF and Occ_SDF_Hybrid, our method can generate more accurate surfaces.

Table 2. Evaluation of 3D reconstruction meshes on Replica.

	Method	Normal C. ↑	Chamfer-L1 ↓	F-score ↑
Grid	MonoSDF [46]	90.93	3.23	85.91
	Ours	**91.66**	**2.91**	**88.79**
MLP	MonoSDF [46]	92.11	2.94	86.18
	Occ_SDF_Hybrid† [17]	93.34	**2.58**	91.54
	Ours	**93.49**	2.60	**91.72**

Table 3. Evaluation of 3D reconstruction meshes on the advanced set of Tanks and Temples. The F-score is reported in the table.

	Method	Auditorium	Ballroom	Courtroom	Museum	mean
Grid	MonoSDF [46]	3.17	3.70	13.75	5.68	6.58
	Occ_SDF_Hybrid† [17]	3.76	3.58	14.04	7.27	7.16
	Ours	**4.84**	**4.31**	**14.18**	**7.60**	**7.73**
MLP	MonoSDF [46]	3.09	2.47	10.00	5.10	5.17
	Occ_SDF_Hybrid† [17]	3.21	3.34	10.29	4.66	5.38
	Ours	**4.24**	**4.69**	**10.79**	**5.91**	**6.41**

Reconstruction on ScanNet [5]**.** Table 1 compares our method against other state-of-the-art (SOTA) methods on ScanNet. When contrasted with the baseline MonoSDF, our method consistently shows better performance across all metrics, regardless of whether utilizing MLP or grid representation. For example, employing grid representation, our method exhibits a 5.7% higher F-score than MonoSDF. Using the MLP representation, our method outperforms MonoSDF by 1.0% in recall, 9.8% in precision, and 6.1% in F-score. In addition, our method with the MLP representation surpasses other SOTA methods on almost all metrics, explaining the effectiveness of our method on reconstruction. Notably, our method achieves a 2.9% higher F-score compared to Occ_SDF_Hybrid.

Table 4. Evaluation of the rendering quality. PSNR is given in this table.

	Method	ScanNet	Replica	Tanks and Temples
Grid	MonoSDF [46]	28.96	40.25	28.45
	Ours	**30.09**	**42.04**	**29.52**
MLP	MonoSDF [46]	26.40	34.45	24.13
	Occ_SDF_Hybrid† [17]	26.98	35.50	24.72
	Ours	**27.77**	**36.06**	**25.47**

Reconstruction on Replica [32]**.** The results on the synthetic dataset Replica are shown in Table 2. Using the grid representation, our method reports a notable improvement of 9.91% in chamfer distance and 2.88% in F-score. Similar increases are also illustrated in our method with the MLP representation across all metrics. Our approach also achieves on-par results with Occ_SDF_Hybrid.

Reconstruction on Tanks and Temples [13]**.** Table 3 shows quantitative outcomes on the challenging large-scale dataset Tanks and Temples. The F-score for each scene - Auditorium, Ballroom, Courtroom, and Museum - along with their mean, is reported. Our method, with both the MLP and grid representations, outperforms Occ_SDF_Hybrid and MonoSDF. In particular, our F-score is 1.15% higher (grid representation) and 1.24% higher (MLP representation) than MonoSDF across all scenes.

Visualization of Reconstruction Results. Figure 3 and Fig. 4 report qualitative results on ScanNet and Replica respectively. In contrast to other methods, our approach can produce superior meshes and recall more regions, particularly for thin regions/small objects, such as chair legs and table legs. Although our method relies on the SRDF for improved density in volume rendering while the surface geometry is extracted by the SDF, both the proposed SRDF-SDF consistency loss and self-supervised visibility task guide the optimization of the SDF to be more consistent with the SRDF, resulting in more accurate surface boundary. Additional visualization is provided in the supplementary material.

View Synthesis. Table 4 also provides a comparison of view synthesis performance on training views. Notably, 2D views are rendered with density derived from SRDF. On the three datasets, our method outperforms the baseline and Occ_SDF_Hybrid in rendering training views. Overall, our method not only enhances surface reconstruction but also generates more accurate views. The visualization can be found in the supplementary material.

4.3 Ablation Study

Table 5. Ablation study of our method on ScanNet.

	Method	Acc ↓	Comp ↓	Prec ↑	Recall ↑	F-score ↑
a	Baseline	**0.035**	0.048	0.799	0.681	0.733
b	RS. Den.	0.040	0.044	0.772	0.720	0.745
c	RS. Den. + Con. L.	0.039	0.041	0.794	0.760	0.776
d	RS. Den. + Con. L. + Vis.(SDF)	0.039	**0.040**	**0.809**	0.772	0.789
e	RS. Den. + Con. L. + Vis.(SRDF)	0.042	**0.040**	0.804	0.774	0.788
f	RS. Den. + Con.L. + Vis.	0.040	**0.040**	**0.809**	**0.779**	**0.794**

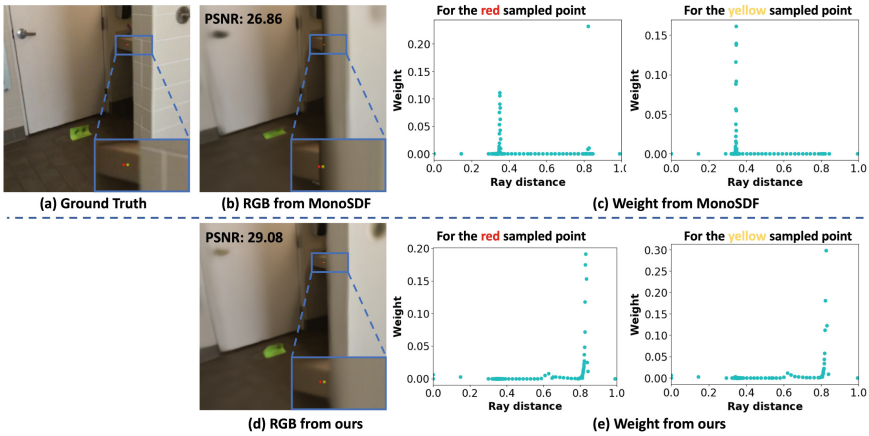

Fig. 5. Comprison of the yielded weight and image. The weight and image generated by our approach align more closely with the actual observations.

Table 5 presents an ablation study to evaluate the effectiveness of our design on ScanNet with MLP representation. The ablation experiments consist of: (a) *Baseline*: MonoSDF. (b) *RS. Den.*: this structure models the volume density using the ray-specific SRDF, without the SRDF-SDF consistency loss and self-supervised visibility task. (c) *RS. Den. + Con. L.*: Building upon the structure in (b), this setup adopts the proposed SRDF-SDF consistency loss. (d) *RS. Den. + Con. L. + Vis.(SDF)*: based on (c), this configuration adds the self-supervised visibility task, in which only SDF is used to compute the visibility label. (e) *RS. Den. + Con. L. + Vis.(SRDF)*: Different from structure (d), this method relies solely on SRDF to compute the visibility label. (f) *RS. Den. + Con. L. + Vis.*: this structure represents the comprehensive model of our method, wherein the visibility labels integrate priors from both SRDF and SDF predictions.

The comparison between configurations (a) and (b) highlights that our design using SRDF to model the density, without any additional constraints, results in a better performance compared to MonoSDF. The SRDF-SDF consistency loss in (c) surpasses the structure in (b) by 3.1% in F-score. This shows the significance of sign consistency. Compared to (c), the self-supervised visibility task in (d)-(f) enhances the reconstruction performance. In particular, structure (f) obtains a better F-score than structures (d) and (e). These explain the effectiveness of our self-supervised visibility task and the computation of visibility labels integrating both SRDF and SDF priors. The visualization of the ablation study is given in the supplementary material.

Furthermore, Fig. 5 compares the weight distribution for the red and yellow sampled points generated by MonoSDF and our method. (1) The analysis of the yellow sampled point demonstrates the benefits of our method for reconstruction: MonoSDF generates high weights and negative SDF values when sampled rays are near the white wall, resulting in false surfaces that significantly deviate from

the ground truth. In contrast, our method yields more accurate surfaces and thus enhances both completeness and recall. (2) The analysis of the red sampled point explains the reason why our method can generate accurate views: MonoSDF produces weights with two local maxima, causing an inaccurate RGB value. This occurs because the SDF along the camera ray fluctuates, resulting in varying density and weight. In comparison, our method produces uni-modal weights that peak only around the surface, resulting in more accurate views. The quantitative PSNR in Fig. 5b and Fig. 5d confirms that our rendered image is more accurate compared to MonoSDF. These findings align with our motivation to achieve a more consistent and accurate weight distribution, thereby reconstructing better surfaces and generating high-quality views. This underscores the effectiveness of using ray-specific SRDF to model volume density in our approach.

5 Conclusion

This work proposes a novel method for neural indoor scene reconstruction, modeling density function with ray-specific SRDF. Firstly, this study analyzed that using the SDF to parameterize the volume density may introduce noises in multi-object indoor scenes, and the SRDF exhibits stronger relationships with the actual 2D observations. Motivated by this, this work employed SRDF to model the volume density, while SDF mainly focuses on the representation of the 3D surface. Secondly, a SRDF-SDF consistency loss has been introduced to align the sign between SRDF and SDF. Thirdly, a self-supervised visibility task has been designed to distinguish whether the sampled 3D point is physically visible or occluded. It combined the prior from SDF and SRDF as visibility labels to improve geometry reconstruction without any other annotations. Experimental results on real-world and synthetic datasets showed that our method excels not only in reconstructing better surfaces but also in generating more accurate views.

References

1. Bozic, A., Palafox, P., Thies, J., Dai, A., Nießner, M.: Transformerfusion: monocular RGB scene reconstruction using transformers. Adv. Neural. Inf. Process. Syst. **34**, 1403–1414 (2021)
2. Collins, R.T.: A space-sweep approach to true multi-image matching. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 358–363. IEEE (1996)
3. Coughlan, J.M., Yuille, A.L.: Manhattan world: Compass direction from a single image by Bayesian inference. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, vol. 2, pp. 941–947. IEEE (1999)
4. Curless, B., Levoy, M.: A volumetric method for building complex models from range images. In: Proceedings of the 23rd Annual Conference on Computer Graphics and Interactive Techniques, pp. 303–312 (1996)
5. Dai, A., Chang, A.X., Savva, M., Halber, M., Funkhouser, T., Nießner, M.: ScanNet: richly-annotated 3D reconstructions of indoor scenes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5828–5839 (2017)

6. Duzceker, A., Galliani, S., Vogel, C., Speciale, P., Dusmanu, M., Pollefeys, M.: Deepvideomvs: multi-view stereo on video with recurrent spatio-temporal fusion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 15324–15333 (2021)
7. Feng, Z., Yang, L., Guo, P., Li, B.: Cvrecon: rethinking 3D geometric feature learning for neural reconstruction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 17750–17760 (2023)
8. Fu, Q., Xu, Q., Ong, Y.S., Tao, W.: Geo-neus: geometry-consistent neural implicit surfaces learning for multi-view reconstruction. Adv. Neural. Inf. Process. Syst. **35**, 3403–3416 (2022)
9. Gallup, D., Frahm, J.M., Mordohai, P., Yang, Q., Pollefeys, M.: Real-time plane-sweeping stereo with multiple sweeping directions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1–8. IEEE (2007)
10. Guo, H., et al.: Neural 3D scene reconstruction with the manhattan-world assumption. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5511–5520 (2022)
11. Hu, S., et al.: Consistentnerf: Enhancing neural radiance fields with 3D consistency for sparse view synthesis. arXiv preprint arXiv:2305.11031 (2023)
12. Ju, J., Tseng, C.W., Bailo, O., Dikov, G., Ghafoorian, M.: Dg-recon: Depth-guided neural 3D scene reconstruction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 18184–18194 (2023)
13. Knapitsch, A., Park, J., Zhou, Q.Y., Koltun, V.: Tanks and temples: benchmarking large-scale scene reconstruction. ACM Trans. Graph. **36**(4), 1–13 (2017)
14. Liang, Z., Huang, Z., Ding, C., Jia, K.: Helixsurf: a robust and efficient neural implicit surface learning of indoor scenes with iterative intertwined regularization. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 13165–13174 (2023)
15. Lorensen, W.E., Cline, H.E.: Marching cubes: a high resolution 3D surface construction algorithm. In: Seminal Graphics: Pioneering Efforts that Shaped the Field, pp. 347–353 (1998)
16. Luo, K., Guan, T., Ju, L., Huang, H., Luo, Y.: P-mvsnet: learning patch-wise matching confidence aggregation for multi-view stereo. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 10452–10461 (2019)
17. Lyu, X., et al.: Learning a room with the OCC-SDF hybrid: Signed distance function mingled with occupancy aids scene representation. arXiv preprint arXiv:2303.09152 (2023)
18. Mescheder, L., Oechsle, M., Niemeyer, M., Nowozin, S., Geiger, A.: Occupancy networks: learning 3D reconstruction in function space. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4460–4470 (2019)
19. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: NeRF: representing scenes as neural radiance fields for view synthesis. Commun. ACM **65**(1), 99–106 (2021)
20. Müller, T., Evans, A., Schied, C., Keller, A.: Instant neural graphics primitives with a multiresolution hash encoding. ACM Trans. Graphics **41**(4), 1–15 (2022)
21. Murez, Z., van As, T., Bartolozzi, J., Sinha, A., Badrinarayanan, V., Rabinovich, A.: Atlas: end-to-end 3D scene reconstruction from posed images. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12352, pp. 414–431. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58571-6_25

22. Niemeyer, M., Barron, J.T., Mildenhall, B., Sajjadi, M.S., Geiger, A., Radwan, N.: RegNeRF: regularizing neural radiance fields for view synthesis from sparse inputs. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5480–5490 (2022)
23. Oechsle, M., Peng, S., Geiger, A.: Unisurf: unifying neural implicit surfaces and radiance fields for multi-view reconstruction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 5589–5599 (2021)
24. Park, J.J., Florence, P., Straub, J., Newcombe, R., Lovegrove, S.: DeepsDF: learning continuous signed distance functions for shape representation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 165–174 (2019)
25. Peng, R., Gu, X., Tang, L., Shen, S., Yu, F., Wang, R.: GENs: generalizable neural surface reconstruction from multi-view images. In: Advances in Neural Information Processing Systems (2023)
26. Ren, Y., Zhang, T., Pollefeys, M., Süsstrunk, S., Wang, F.: Volrecon: volume rendering of signed ray distance functions for generalizable multi-view reconstruction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 16685–16695 (2023)
27. Sayed, M., Gibson, J., Watson, J., Prisacariu, V., Firman, M., Godard, C.: Simplerecon: 3D reconstruction without 3D convolutions. In: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T. (eds.) ECCV 2022. LNCS, vol. 13693, pp. 1–19. Springer, Cham (2022)
28. Schönberger, J.L., Zheng, E., Frahm, J.-M., Pollefeys, M.: Pixelwise view selection for unstructured multi-view stereo. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9907, pp. 501–518. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46487-9_31
29. Sitzmann, V., Martel, J., Bergman, A., Lindell, D., Wetzstein, G.: Implicit neural representations with periodic activation functions. Adv. Neural. Inf. Process. Syst. **33**, 7462–7473 (2020)
30. Somraj, N., Soundararajan, R.: Vip-NeRF: visibility prior for sparse input neural radiance fields. arXiv preprint arXiv:2305.00041 (2023)
31. Stier, N., Rich, A., Sen, P., Höllerer, T.: Vortx: volumetric 3D reconstruction with transformers for voxelwise view selection and fusion. In: 2021 International Conference on 3D Vision, pp. 320–330. IEEE (2021)
32. Straub, J., et al.: The replica dataset: a digital replica of indoor spaces. arXiv preprint arXiv:1906.05797 (2019)
33. Sun, J., Xie, Y., Chen, L., Zhou, X., Bao, H.: Neuralrecon: real-time coherent 3D reconstruction from monocular video. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 15598–15607 (2021)
34. Tan, M., et al.: MnasNet: platform-aware neural architecture search for mobile. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2820–2828 (2019)
35. Tang, H., et al.: Searching efficient 3D architectures with sparse point-voxel convolution. In: Vedaldi, A., Bischof, H., Brox, T., Frahm, J.-M. (eds.) ECCV 2020. LNCS, vol. 12373, pp. 685–702. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58604-1_41
36. Wang, J., et al.: NeuRIS: neural reconstruction of indoor scenes using normal priors. In: Avidan, S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T. (eds.) ECCV 2022. LNCS, vol. 13692, pp. 139–155. Springer, Cham (2022)
37. Wang, K., Shen, S.: MVDepthNet: real-time multiview depth estimation neural network. In: 2018 International Conference on 3D Vision, pp. 248–257. IEEE (2018)

38. Wang, P., Liu, L., Liu, Y., Theobalt, C., Komura, T., Wang, W.: Neus: learning neural implicit surfaces by volume rendering for multi-view reconstruction. arXiv preprint arXiv:2106.10689 (2021)
39. Wang, Y.: Neuralroom: geometry-constrained neural implicit surfaces for indoor scene reconstruction. ACM Trans. Graphics **41**(6), 1–15 (2022)
40. Yao, Y., Luo, Z., Li, S., Fang, T., Quan, L.: MvsNet: depth inference for unstructured multi-view stereo. In: Proceedings of the European Conference on Computer Vision, pp. 767–783 (2018)
41. Yao, Y., Luo, Z., Li, S., Shen, T., Fang, T., Quan, L.: Recurrent MVSNet for high-resolution multi-view stereo depth inference. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 5525–5534 (2019)
42. Yariv, L., Gu, J., Kasten, Y., Lipman, Y.: Volume rendering of neural implicit surfaces. Adv. Neural. Inf. Process. Syst. **34**, 4805–4815 (2021)
43. Ye, B., Liu, S., Li, X., Yang, M.H.: Self-supervised super-plane for neural 3D reconstruction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 21415–21424 (2023)
44. Yin, R., Karaoglu, S., Gevers, T.: Geometry-guided feature learning and fusion for indoor scene reconstruction. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 3652–3661 (2023)
45. Yu, Z., Gao, S.: Fast-MVSNet: sparse-to-dense multi-view stereo with learned propagation and gauss-newton refinement. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1949–1958 (2020)
46. Yu, Z., Peng, S., Niemeyer, M., Sattler, T., Geiger, A.: MonoSDF: exploring monocular geometric cues for neural implicit surface reconstruction. Adv. Neural. Inf. Process. Syst. **35**, 25018–25032 (2022)
47. Zhu, Z., et al.: Nicer-slam: neural implicit scene encoding for RGB slam. arXiv preprint arXiv:2302.03594 (2023)
48. Zins, P., Xu, Y., Boyer, E., Wuhrer, S., Tung, T.: Multi-view reconstruction using signed ray distance functions (SRDF). In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 16696–16706 (2023)