Sparis: Neural Implicit Surface Reconstruction of Indoor Scenes from Sparse Views

Yulun Wu[1, 2*], **Han Huang**[1, 2*], **Wenyuan Zhang**[1, 2], **Chao Deng**[1, 2], **Ge Gao**[1, 2†], **Ming Gu**[1, 2], **Yu-Shen Liu**[2]

[1]Beijing National Research Center for Information Science and Technology (BNRist), Tsinghua University, Beijing, China
[2]School of Software, Tsinghua University, Beijing, China
{wu-yl22, h-huang20, zhangwen21, dengc23}@mails.tsinghua.edu.cn, {gaoge, guming, liuyushen}@tsinghua.edu.cn

Abstract

In recent years, reconstructing indoor scene geometry from multi-view images has achieved encouraging accomplishments. Current methods incorporate monocular priors into neural implicit surface models to achieve high-quality reconstructions. However, these methods require hundreds of images for scene reconstruction. When only a limited number of views are available as input, the performance of monocular priors deteriorates due to scale ambiguity, leading to the collapse of the reconstructed scene geometry. In this paper, we propose a new method, named *Sparis*, for indoor surface reconstruction from sparse views. Specifically, we investigate the impact of monocular priors on sparse scene reconstruction, introducing a novel prior based on inter-image matching information. Our prior offers more accurate depth information while ensuring cross-view matching consistency. Additionally, we employ an angular filter strategy and an epipolar matching weight function, aiming to reduce errors due to view matching inaccuracies, thereby refining the inter-image prior for improved reconstruction accuracy. The experiments conducted on widely used benchmarks demonstrate superior performance in sparse-view scene reconstruction.

Introduction

Reconstructing indoor 3D geometry from multi-view images is a significant task in the field of computer vision and graphics. Due to the sparse nature of indoor image acquisition, traditional Multi-View Stereo (MVS) methods (Yao et al. 2018; Ding et al. 2022) often face challenges in generating satisfactory results when overlap is limited.

Recently, with the emergence of Neural Radiance Fields (NeRF) (Mildenhall et al. 2020) technology, implicit scene representations have injected new vitality into multi-view reconstruction of 3D scenes. Several works (Wang et al. 2021; Yariv et al. 2021) utilize Signed Distance Functions (SDF) as a geometric representation and employ a neural rendering pipeline to accurately learn the geometry of scenes from multi-view images. Although they have gained considerable advancement in indoor scene reconstruction, it is still challenged by texture-less regions (e.g., walls, floors, ceilings) and complex object layouts. To solve these issues, subsequent works leverage structural constraints (Guo et al. 2022; Ye et al. 2023; Wang et al. 2024) or general-purpose monocular priors (Yu et al. 2022; Wang et al. 2022a; Liang et al. 2023) to provide more comprehensive supervision of depth and normal, further enhancing the quality of reconstruction. However, reliable reconstruction results always rely on dense input views. When only sparse views are provided, the performance of these methods significantly decreases.

Two categories of approaches have provided inspiration for addressing the problem of indoor sparse-view reconstruction, yet cannot resolve this issue. Indoor sparse novel view synthesis methods (Roessle et al. 2022; Uy et al. 2023; Song et al. 2023) improve rendering quality by employing dense depth priors or refined monocular priors, but fail to capture accurate geometry for lacking of clear geometric representation. Object level sparse reconstruction methods (Long et al. 2022; Ren et al. 2023; Wu, Graikos, and Samaras 2023; Huang et al. 2023) enhance the feature extraction capabilities of neural fields while underperform in large and complex indoor scenes.

In this work, we adopt SDF for geometric representation, revisiting the paradigms of prior-based neural implicit learning under sparse settings. We notice that enforcing monocular depth supervision diminishes the reconstruction quality due to the inability to calibrate depth scale within sparse views. In addressing this challenge, we leverage matching information between images to obtain more reliable absolute depth prior. Additionally, to further ensure consistency between views, we introduce a reprojection loss, which optimizes the reconstruction geometry surface based on matching relationships. As our matching relationships are entirely determined by the matching network, matching errors may impact the accuracy of our priors. We designed a matching mechanism consisting of a matching angle filter and an epipolar weight function. The matching angle filter calculates the angular score between views and can filter out severe bias introduced by matching errors in nearby perspectives. The epipolar weight function calculates the Sampson Distance for matched pixels within the corresponding images, and quantitatively assesses their correspondence in 3D space, enhancing the overall accuracy of reconstruction. As shown in Figure 1, our method can achieve more complete and detailed surface reconstruction, compared with previous

*These authors contributed equally.
†Corresponding author.

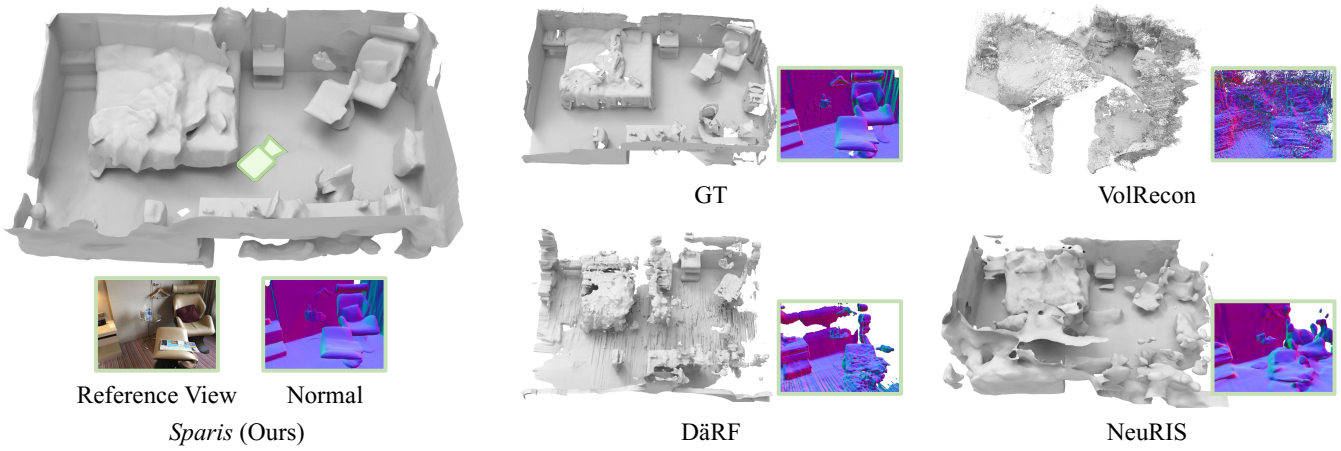

Figure 1: Surface reconstruction results from sparse views of an indoor scene. Our method Sparis outperforms in addressing challenges such as missing reconstruction details (NeuRIS), uneven surface (DäRF), and spatial noise (VolRecon).

approaches. We highlight our key contributions as follows.

- We propose *Sparis*, a novel surface reconstruction method that utilizes correspondence information between images for indoor sparse-view reconstruction. Our method leverages pixel-pair information for depth optimization and reprojection losses to refine the surface.
- We develop matching optimization strategies aimed at minimizing the effects of matching inaccuracies, ensuring more reliable depth and reprojection alignments.
- Our extensive evaluations on both real-world and synthetic datasets show that *Sparis* achieves superior performance over current leading indoor reconstruction methods with sparse views.

Related Works

Novel View Synthesis for Indoor Scenes

Synthesizing images from novel viewpoints of a scene within a set of images has long attracted attention in the field of computer vision. Recently, Neural Radiance Fields (NeRF) (Mildenhall et al. 2020) as a neural implicit representation method, achieves high-quality and view-dependent rendering through a volume rendering pipeline. Based on NeRF, many studies have made improvements in rendering speed (Reiser et al. 2021; Yu et al. 2021; Sun, Sun, and Chen 2022; Müller et al. 2022; Zhang et al. 2023a), quality (Barron et al. 2021, 2022; Wang et al. 2022b; Han et al. 2024), and generalizability (Chen et al. 2021; Johari, Lepoittevin, and Fleuret 2022; Cong et al. 2023). Apart from architectural improvements in universal conditions, some researchers have focused on specific categories of scene reconstruction, such as indoor (Ying et al. 2023; Gao, Cao, and Shan 2023), outdoor (Zhang et al. 2023b; Irshad et al. 2023), underwater (Levy et al. 2023), satellite (Marí, Facciolo, and Ehret 2022), and urban environments (Tancik et al. 2022; Rematas et al. 2022; Turki, Ramanan, and Satyanarayanan 2022), aiming to achieve higher rendering quality within these distinct settings. However, in indoor scenes, challenges often arise due to a limited number of images and small coverage areas. DDP-NeRF (Roessle et al. 2022) trained a dense depth prior from a large indoor dataset to constrain the small-convergence NeRF optimization. DäRF (Song et al. 2023) and SCADE (Uy et al. 2023) improved the ambiguity and scale issues of the SOTA monocular depth model priors, leading to more accurate depth supervision and thereby enhancing the rendering effects. Although significant progress has been made in synthesizing novel views for sparse indoor scenes, these results fail to achieve the reconstruction geometry under sparse views due to the lack of accurate geometric representations, such as Signed Distance Functions (SDF).

Geometry Reconstruction for Indoor Scenes

Reconstructing geometric surfaces from multiple viewpoints is relatively straightforward for a single object with dense views. Inspired by NeRF, NeuS (Wang et al. 2021) and VolSDF (Yariv et al. 2021) utilized a volumetric rendering pipeline to learn the neural implicit surfaces of objects from multi-view images. HelixSurf (Liang et al. 2023) and Neus2 (Wang et al. 2023) adopted dense grid feature coding like instant-ngp (Müller et al. 2022) to accelerate the reconstruction process. However, they are generic, object-centric methods that perform poorly in scenes with many untextured areas and significant lighting variations. To tackle the challenges of indoor scene reconstruction, MonoSDF (Yu et al. 2022) and NeuRIS (Wang et al. 2022a) introduced 2D pretrained models as priors, effectively dealing with the issues in reconstructing untextured areas. Manhattan-SDF (Guo et al. 2022) and S^3P (Ye et al. 2023) did not employ geometric priors from 2D images; instead, they drew upon the laws of the physical world to design constraints on surface normals for indoor scenes. However, current indoor reconstruction works still demand a high number of images, often requiring hundreds of images to achieve satisfactory reconstruction results. Recently, some works (Long et al. 2022; Ren et al. 2023; Huang et al. 2023; Wu, Graikos, and Samaras 2023; Xu et al. 2023) have attempted to perform im-

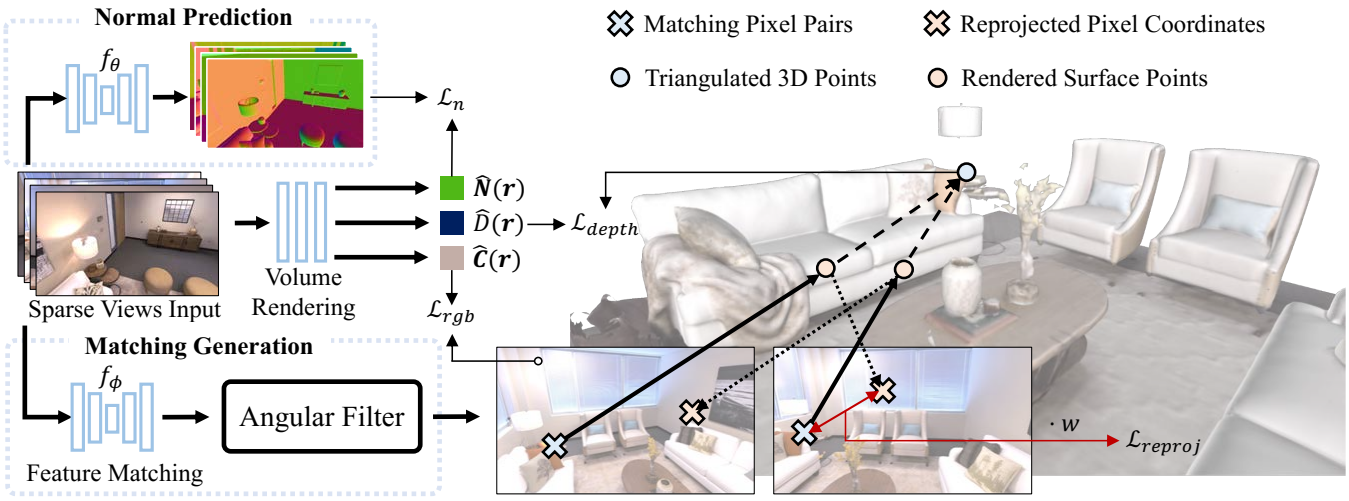

Figure 2: The overview of Sparis. Given sparse indoor images, the reconstruction of 3D surfaces is achieved via a 2-stage process: (1) Pre-processing: estimated normal maps and matching pixel pairs are derived respectively using a pre-trained normal prediction network f_θ and a feature matching network f_ϕ; (2) Training with priors: the neural rendering procedure is optimized with inter-image depth priors, cross-view reprojection and monocular normal priors, generating complete and detailed geometry.

plicit surface reconstruction with sparse views. Yet, these studies primarily concentrate on object-centric reconstruction with few views, overlooking indoor scenes. With more objects and less view overlap in the scene, reconstructing under sparse views grows increasingly difficult.

Method

In this study, we aim to reconstruct the fidelity surface $\mathcal{S}$ of an indoor scene from a limited set of images $\mathcal{I} = \{I_i \mid i \in 1, \dots, M\}$ and camera poses $\mathcal{T} = \{T_i \mid i \in 1, \dots, M\}$. We introduce *Sparis*, a neural surface reconstruction approach optimized for sparse view inputs, as illustrated in Figure 2.

Neural Implicit Surface Volume Rendering

We model both geometry and appearance using SDF and color fields, learned by the differentiable rendering pipeline. Defining the surface geometry of the indoor scene as the zero-level set of SDF $\mathcal{S} = \{\boldsymbol{x} \in \mathbb{R}^3 \mid f(\boldsymbol{x}) = 0\}$, we then adopt VolSDF (Yariv et al. 2021) as our baseline. This allows for the transformation of SDF into volumetric density for volume rendering, with both SDF and color parameterized by two MLPs as VolSDF.

Given a pixel from one image, the ray could be denoted as $\{\boldsymbol{r}(t_i) = \boldsymbol{o} + t\boldsymbol{d} \mid t > 0\}$, where $\boldsymbol{o}$ is the camera center and $\boldsymbol{d}$ is the direction of the ray. The rendered color is accumulated by volume rendering with N discrete points:

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

where T_i is the accumulated transmittance, α_i is the opacity values, as denoted by

$$T_i = \prod_{j=1}^{i-1} (1 - \alpha_i) \,, \quad \alpha_i = 1 - \exp\left(-\sigma_i \delta_i\right). \tag{2}$$

Following VolSDF, we transform SDF values s to density values σ using a learnable parameter β:

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

Subsequently, we calculate the depth $\hat{D}(\boldsymbol{r})$ and normal $\hat{\boldsymbol{N}}(\boldsymbol{r})$ at the intersection of the surface with the current ray using the following expressions:

$$\hat{D}(\boldsymbol{r}) = \sum_{i=1}^{N} T_i \, \alpha_i \, t_i \,, \quad \hat{\boldsymbol{N}}(\boldsymbol{r}) = \sum_{i=1}^{N} T_i \, \alpha_i \, \hat{\boldsymbol{n}}_i \,. \tag{4}$$

Inter-Image Depth Loss

MonoSDF (Yu et al. 2022) represents a foundational work in multi-view indoor reconstruction, introducing the Omnidata (Eftekhar et al. 2021) depth prior that supplies a wealth of geometric information. To enforce the consistency between the render depth $\hat{D}(\boldsymbol{r})$ and monocular depth $\bar{D}(\boldsymbol{r})$, It employs a loss function that is invariant to scale:

$$\mathcal{L}_{mono\ depth} = \sum_{\boldsymbol{r} \in \mathcal{R}} \left\| \left(w\hat{D}(\boldsymbol{r}) + q \right) - \bar{D}(\boldsymbol{r}) \right\|^2. \tag{5}$$

This means that the relative depth from monocular input needs to be scaled to an absolute scale for geometry supervision. Scale w and shift q are solved with the least-squares criterion in the rendering process.

However, this strategy can lead to severe depth ambiguity problems, ultimately resulting in the collapse of the reconstructed geometry. This arises from sparse views training process, where the small overlap between sparse views results in only a limited amount of rendering depth being

correctly scaled. Global scale and shift are miscalculated, ultimately leading to errors in depth supervision scale.

To resolve this issue, we introduce a 2D feature points matching network to compute correspondence information between sparse views, utilizing this inter-image information along with image poses to acquire more accurate depth priors. Given a pair of images captured from different viewpoints of current scene, marked as $\{I_a, I_b\}$, we can directly obtain the matching pixel pairs $(\boldsymbol{p}_a, \boldsymbol{p}_b)$ along with an associated matching uncertainty by employing the feature matching network f_ϕ:

$$\left\{(\boldsymbol{p}_a^i, \boldsymbol{p}_b^i, u_{a,b}^i) \mid i \in 1, \ldots, H\right\} = f_\phi(I_a, I_b)\,. \tag{6}$$

Here, H denotes the quantity of matching pixel pairs. Matching uncertainty $u_{a,b}$ is quantified within the range of $[0, 1]$, indicating the confidence of the matching results.

By leveraging the camera poses alongside these matching pixel pairs, it becomes feasible to triangulate the estimated world coordinates $\boldsymbol{x}$, thus inferring absolute depth priors $\widetilde{D}(\boldsymbol{r})$, as demonstrated in Figure 3 (a). Throughout the training phase, for a given reference view I_r, we systematically sample a batch of rays $\{\boldsymbol{r}_r^i\}$ and rays of their corresponding matching pixels $\{\boldsymbol{r}_s^i\}$ from a source view I_s. Consequently, the inter-image depth loss can be expressed as

$$\mathcal{L}_{depth} = \sum_i \frac{1}{\widetilde{D}(\boldsymbol{r}_r^i)}(1 - u_{r,s}^i)\left|\hat{D}(\boldsymbol{r}_r^i) - \widetilde{D}(\boldsymbol{r}_r^i)\right|. \tag{7}$$

Cross-View Reprojection Loss

During the optimization process of neural rendering, we only compute the depth loss for the current image, ensuring one-way accuracy of inter-image information. When the depth loss converges well, we can approximately assume that the correspondence in inter-image relationships has been ensured. However, in each iteration of the neural rendering pipeline, only a small number of pixels from one view are selected, making it challenging to synchronize the depths of one-to-one corresponding pixels in inter-image relationships. To tackle this challenge, we introduce reprojection for optimization.

As it is shown in Figure 3 (a), considering that the point $\boldsymbol{x}_a'$ on which a ray intersect with the surface estimated by neural rendering is not coincident with the triangulated point $\boldsymbol{x}$, since the error between rendered surfaces and real surfaces exists, an offset is also introduced between the reprojected coordinate $\boldsymbol{p}_b'$ and $\boldsymbol{p}_b$ on another view. Given a reference view I_r and a source view I_s, the reprojected coordinate $\boldsymbol{p}_s'$ from the rendered 3D point of reference view to the source view can be calculated as

$$\boldsymbol{p}_s' = K P_s^{-1}\left(\boldsymbol{o}_r + \hat{D}(\boldsymbol{r}_r)\cdot\boldsymbol{d}(\boldsymbol{r}_r)\right), \tag{8}$$

where K denotes camera intrinsic matrix, P represents the camera pose, $\boldsymbol{d}$ is the normalized direction of $\boldsymbol{r}$, and $\boldsymbol{o}$ indicates the world coordinate of the camera viewpoint.

Then, the cross-view reprojection loss is calculated as

$$\mathcal{L}_{reproj} = \sum_i (1 - u_{r,s}^i)\left\|\boldsymbol{p}_s^i - \boldsymbol{p}_s^{i\prime}\right\|_1, \tag{9}$$

$\|\cdot\|_1$ represents the L1 norm.

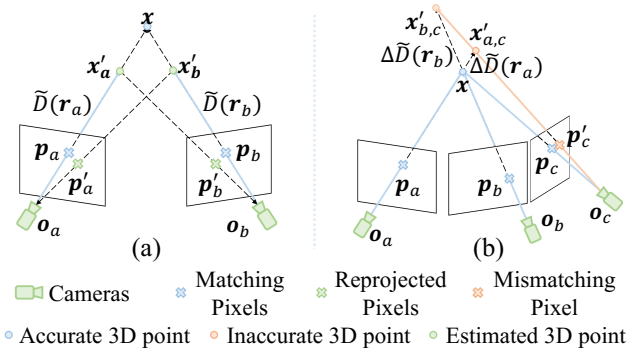

Figure 3: Illustration of matching priors. (a) Using matching pixel pairs, we obtain the triangulated depth $\widetilde{D}$ and reprojected coordinates $\boldsymbol{p}'$ from the rendered 3D surface points; (b) Mismatches cause depth estimation errors, especially under minimal translation and angular variations.

Matching Optimization Strategies

Matching networks inherently introduce certain errors, which can lead to geometric inaccuracies and spatial noise. To mitigate these issues, we develop two optimization strategies: angular filter for refining image matching pairs and epipolar weight function for enhancing pixel matching pairs.

Angular Filter. In multi-view geometry, triangulation errors are strongly influenced by the angles between ray pairs. As illustrated in Figure 3 (b), smaller angles result in greater relative errors in depth estimation when mismatches occur. Therefore, relying solely on the number of matching pairs as a metric for source view selection can lead to more inaccurate estimations. To mitigate this, we compute the certainty-weighted average of the normalized direction vectors of the rays at each matching pixel. The score for translation and angular variations of the views is then calculated as

$$S_{a,b} = 1 - \cos\left(\sum_i (1 - u_{a,b}^i)\boldsymbol{d}_a^i\,,\ \sum_i (1 - u_{a,b}^i)\boldsymbol{d}_b^i\right), \tag{10}$$

where $\boldsymbol{d}$ are the normalized direction vectors of rays. For reference view I_r, the source view is picked as

$$I_s = \arg\max([S_{r,i} - \epsilon > 0]\cdot H_{r,i})\,,\ i \neq r\,. \tag{11}$$

H indicates the number of matching pixel pairs. $[\cdot]$ represents the Iverson bracket.

Epipolar Weight Function. Feature matching networks, as data-driven models operating at the image level, often lack verification of multi-view geometric consistency within the scene. Consequently, the matched pixels may not adhere to the correct spatial-geometric relationships, failing to meet the scene's geometric constraints. Ideally, during triangulation, the rays corresponding to a matched pixel pair should intersect at a single point, with the pixels lying on the epipolar lines. To mitigate this limitation, we introduce an epipolar weight, which can be computed as

$$w_{r,s}^i = \frac{1}{2}\left(1 - \text{Sigmoid}\left(\gamma\cdot d_s(\boldsymbol{p}_r^i, \boldsymbol{p}_s^i)\right)\right). \tag{12}$$

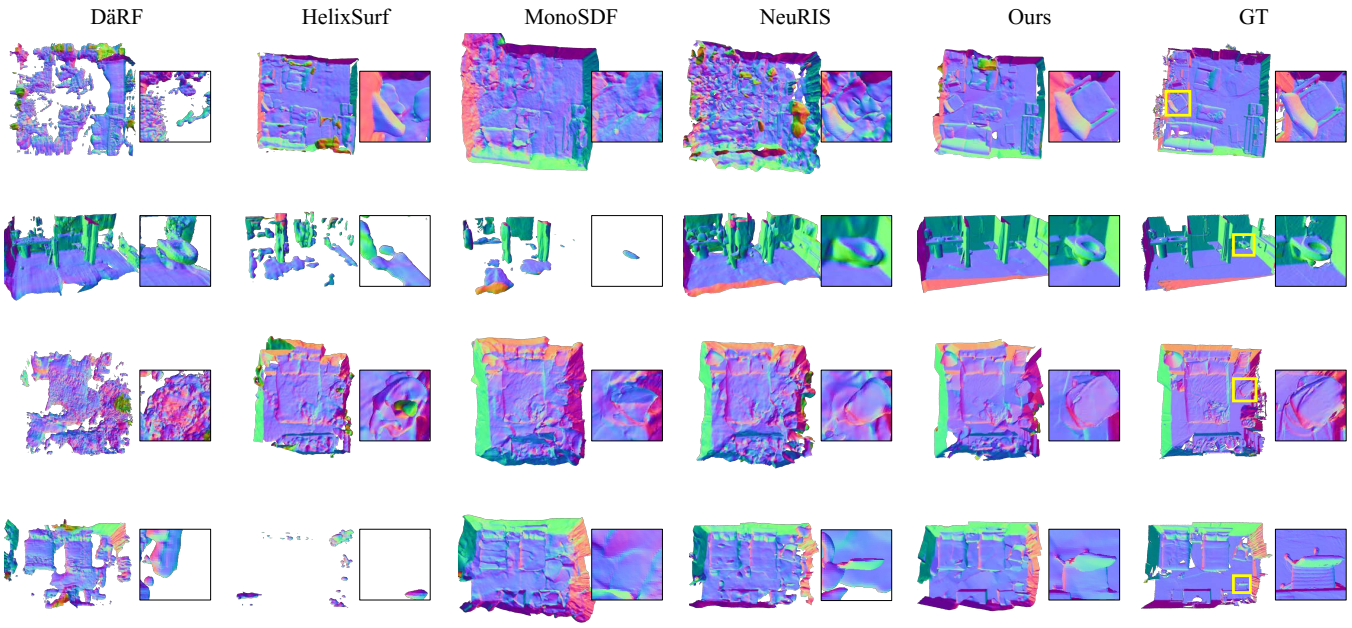

Figure 4: Visual comparisons of 3D reconstruction results on ScanNet with sparse views. The overall top views and the zoom-in views of the marked areas show that our approach produces more complete and fine-grained geometry.

d_s represents the Sampson Distance, calculated as

$$d_s(\boldsymbol{p}_r, \boldsymbol{p}_s) = \frac{(\boldsymbol{p}_s^\top F \boldsymbol{p}_r)^2}{(F\boldsymbol{p}_r)_1^2 + (F\boldsymbol{p}_r)_2^2 + (F^\top \boldsymbol{p}_s)_1^2 + (F^\top \boldsymbol{p}_s)_2^2}, \tag{13}$$

where F denotes the fundamental matrix between reference view I_r and source view I_s. $(\cdot)_k$ represents the k-th element of the vector. Thus, with the consideration of the epipolar weights, $\mathcal{L}_{depth}$ and $\mathcal{L}_{reproj}$ can be rewrote as:

$$\mathcal{L}_{depth} = \sum_i \frac{1}{\widetilde{D}(\boldsymbol{r}_r^i)} (1u_{r,s}^i) w_{r,s}^i \left| \hat{D}(\boldsymbol{r}_r^i) - \widetilde{D}(\boldsymbol{r}_r^i) \right|, \tag{14}$$

$$\mathcal{L}_{reproj} = \sum_i (1 - u_{r,s}^i) w_{r,s}^i \left\| \boldsymbol{p}_s^i - \boldsymbol{p}_s^{i\prime} \right\|_1. \tag{15}$$

Loss Functions

The overall loss functions are:

$$\mathcal{L} = \mathcal{L}_{rgb} + \lambda_1 \mathcal{L}_{depth} + \lambda_2 \mathcal{L}_{reproj} + \lambda_3 \mathcal{L}_n + \lambda_4 \mathcal{L}_{eik}, \tag{16}$$

where $\mathcal{L}_{depth}$ and $\mathcal{L}_{reproj}$ are the inter-image depth loss and cross-view reprojection loss defined above.

$\mathcal{L}_{rgb}$ is the difference between the rendered and ground-truth pixel colors:

$$\mathcal{L}_{rgb} = \frac{1}{|\mathcal{R}|} \sum_{\boldsymbol{r} \in \mathcal{R}} \left\| \boldsymbol{C}(\boldsymbol{r}) - \hat{\boldsymbol{C}}(\boldsymbol{r}) \right\|_1. \tag{17}$$

Similar to NeuRIS (Wang et al. 2022a), we utilize a pre-trained network f_θ to predict monocular normals $\bar{\boldsymbol{N}}(\boldsymbol{r})$, which are then applied to the Normal loss:

$$\mathcal{L}_n = \frac{1}{|\mathcal{R}|} \sum_{\boldsymbol{r} \in \mathcal{R}} \left\| \hat{\boldsymbol{N}}(\boldsymbol{r}) - \bar{\boldsymbol{N}}(\boldsymbol{r}) \right\|_1 + \left\| 1 - \hat{\boldsymbol{N}}(\boldsymbol{r})^\top \bar{\boldsymbol{N}}(\boldsymbol{r}) \right\|_1. \tag{18}$$

In line with the previous approaches, we introduce an Eikonal regularization term (Gropp et al. 2020) on the random sample points $\mathcal{Y}$ for the SDF field $f(\boldsymbol{x})$:

$$\mathcal{L}_{eik} = \frac{1}{|\mathcal{Y}|} \sum_{\boldsymbol{x} \in \mathcal{Y}} (\|\nabla f(\boldsymbol{x})\| - 1)^2. \tag{19}$$

Experiments and Anaysis

Datasets

ScanNet. ScanNet (Dai et al. 2017), a comprehensive real-world dataset, encompasses over 2.5 million views across 1513 scenes, each annotated with 3D camera poses and surface reconstructions. To evaluate the performance of our algorithm, we adopted the sparse setting used by DDP-NeRF (Roessle et al. 2022), sampling 15 to 20 images per scene at a resolution of 624×468 for surface reconstruction.

Replica. The Replica dataset (Straub et al. 2019) is notable for its high-quality reconstructions of various indoor environments. To further ascertain the robustness of our approach, we followed the scene selection strategy outlined in (Yu et al. 2022), opting for 8 distinct scenes. From each scene, 10 images are uniformly sampled out of 2000, at a resolution of 600×340 for our experimental dataset.

Implementation Details

We adopt a similar model architecture as VolSDF (Yariv et al. 2021). RoMa (Edstedt et al. 2023), a robust network for dense matching, is adopted as network f_ϕ to compute priors between images. We utilize the pre-trained Omnidata (Eftekhar et al. 2021) as our normal estimation network f_θ to generate monocular normal priors. All the experiments

Method	F-score↑	Acc.↓	Comp.↓	Prec.↑	Recal.↑
COLMAP (Schonberger and Frahm 2016)	0.161	0.179	0.583	0.284	0.124
TransMVSNet (Ding et al. 2022)	0.119	0.352	0.473	0.142	0.102
DDP-NeRF (Roessle et al. 2022)	0.287	0.280	0.080	0.202	0.539
DäRF (Song et al. 2023)	0.295	0.273	0.127	0.242	0.393
NeuS (Wang et al. 2021)	0.132	0.300	0.665	0.185	0.105
VolRecon (Ren et al. 2023)	0.155	0.225	0.284	0.174	0.144
HelixSurf (Liang et al. 2023)	0.238	0.341	0.249	0.249	0.229
S^3P (Ye et al. 2023)	0.277	0.300	0.177	0.274	0.285
MonoSDF (Yu et al. 2022)	0.328	0.328	0.152	0.320	0.341
NeuRIS (Wang et al. 2022a)	0.464	0.180	0.082	0.445	0.488
Ours	**0.647**	**0.056**	**0.060**	**0.666**	**0.631**

Table 1: Quantitative comparisons of room-scale surface reconstruction results over 10 scenes of ScanNet with 15-20 input views. The best and the second best results are denoted as bold and underlined, respectively.

Method	F-score↑	Acc.↓	Comp.↓	Prec.↑	Recal.↑
HelixSurf (Liang et al. 2023)	0.028	0.558	0.927	0.035	0.019
S^3P (Ye et al. 2023)	0.018	0.271	2.733	0.152	0.010
MonoSDF (Yu et al. 2022)	0.454	0.081	0.139	0.497	0.423
NeuRIS (Wang et al. 2022a)	0.431	0.074	0.147	0.489	0.387
Ours	**0.825**	**0.031**	**0.073**	**0.881**	**0.777**

Table 2: Quantitative comparisons of room-scale surface reconstruction results over 8 scenes of Replica with 10 input views. The best and the second best results are denoted as bold and underlined, respectively.

are conducted on an NVIDIA RTX3090 GPU. More experimental settings and metrics calculations are provided in the supplementary materials.

Comparison

ScanNet. We compare our approach with various types of indoor reconstruction methods on ScanNet dataset: (1) MVS reconstruction methods: COLMAP (Schonberger and Frahm 2016), TransMVSNet (Ding et al. 2022); (2) Novel view synthesis methods for sparse-view indoor scenes: DDP-NeRF (Roessle et al. 2022), DäRF (Song et al. 2023); (3) Neural implicit surface methods for sparse-view reconstruction: VolRecon (Ren et al. 2023); (4) Neural implicit surface methods for indoor scenes: NeuS (Wang et al. 2021), NeuRIS (Wang et al. 2022a), MonoSDF (Yu et al. 2022), HelixSurf (Liang et al. 2023), S^3P (Ye et al. 2023).

To ensure a fair comparison, we fine-tune the experimental setups for specific baselines to maximize their performance. For COLMAP and TransMVSNet, we employ Poisson Reconstruction (Kazhdan, Bolitho, and Hoppe 2006) to generate surface meshes from the densely matched point cloud outputs. In the cases of DDP-NeRF and DäRF, we utilize the Marching Cube algorithm to create meshes from the learned density fields, applying an appropriately adjusted threshold for optimal results. MonoSDF, under its default hyper-parameter configuration, was unable to produce valid meshes; thus, we modified the weight of the monocular depth loss to 0.001 (originally 0.1) for a more equitable comparison. The quantitative outcomes of this assessment are presented in Table 1. Notably, NeuS was unable to generate valid meshes for 4 scenes, and HelixSurf for 1; hence, we specifically report results for the successfully reconstructed scenes for these methods. Our methodology surpasses all benchmarks, demonstrating a significant improvement. Concurrently, as depicted in Figure 4, previous methods could only generate fragmented and noisy surfaces. In contrast, our technique delivers more visually complete geometries, characterized by smoother surfaces and more refined details.

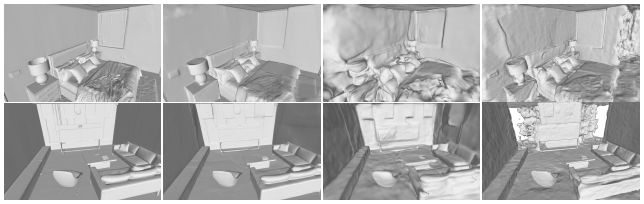

Figure 5: Visual comparisons of 3D reconstruction results on Replica with sparse views.

Replica. As a complementary experiment to validate the robustness on different datasets, we compare our approach with MonoSDF (Yu et al. 2022), NeuRIS (Wang et al. 2022a), HelixSurf (Liang et al. 2023) and S^3P (Ye et al. 2023). The quantitative comparisons are listed in Table 2,

$\mathcal{L}_n$	$\mathcal{L}_{depth}$	$\mathcal{L}_{reproj}$	F-score↑	Acc.↓	Comp.↓	Prec.↑	Recall↑
			0.244	0.177	0.319	0.308	0.212
√			0.253	0.375	0.225	0.250	0.258
√	√		0.598	0.061	0.074	0.624	0.577
√		√	0.560	0.083	0.243	0.518	0.549
√	√	√	**0.647**	**0.056**	**0.060**	**0.666**	**0.631**

Table 3: Ablation studies of each component of our method over 10 scenes of ScanNet.

while the visual comparisons are shown in Figure 5. The most effective indoor reconstruction methods, MonoSDF and NeuRIS, produce uneven surfaces due to the lack of accurate depth guidance. Our approach exhibits a more pronounced advantage on Replica dataset, characterized by minimal occlusion and precise poses, clearly surpassing several neural indoor surface reconstruction methods.

Ablation Study

To evaluate the effectiveness of the components of our proposed priors, we conduct ablation studies on 5 different settings: (1) Naive neural rendering framework without any introduced prior; (2) Neural rendering framework with normal priors; (3) Ours without cross-view reprojection loss; (4) Ours without inter-image depth loss; (4) Ours: neural rendering framework with normal priors, inter-image depth loss and cross-view reprojection loss.

Table 3 demonstrates that the monocular normal, as a commonly used form of supervision information for indoor reconstruction, can also improve the reconstruction quality with sparse view inputs. Our inter-image depth loss provides accurate geometric constraints, significantly enhancing the reconstruction quality of fine local details. Furthermore, by ensuring a one-to-one correspondence of matching pixels, the cross-view reprojection loss offers a relaxed yet stable form of supervision. This guarantees inter-view consistency, reduces overfitting in sparse view reconstruction, and ultimately enhances reconstruction quality. The observation readily suggests that the simultaneous application of both constraints not only enhances the geometric quality in each view but also mitigates overfitting in scenarios with few views, leading to a markedly significant improvement compared to the baseline.

Epipolar weight	Angular filter	F-score↑
×	√	0.617
√	×	0.624
√	√	**0.647**

Table 4: Quantitative results of ablation study on epipolar weight and angular filter.

To validate the effectiveness of the matching optimization strategies we propose, we conducted ablation studies on epipolar weight function and angular filter, respectively. As shown in Table 4, both strategies significantly enhance geometric reconstruction. This demonstrates that our method

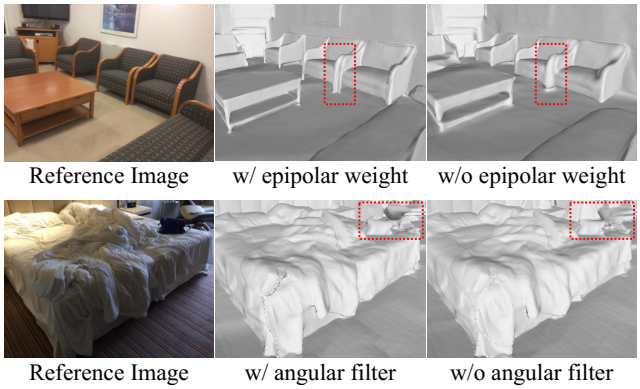

Figure 6: Our reconstruction results with or without matching optimization strategies.

is resilient to noise in the matching network. To intuitively demonstrate the effectiveness of our strategies, we visualize the ablation experiments for a scene from ScanNet, as shown in Figure 6. Owing to the similar textures of the chair legs, the matching network features exhibit high similarity. Without the incorporation of the epipolar weight, the geometric structure of the chairs in the reconstruction degrades, leading to difficulties in distinguishing between different chairs. And it is evident that angular filter is capable of improving the quality of reconstruction at local details by eliminating view pairs with significant error influences.

Conclusion

We introduce a novel neural implicit surface reconstruction approach for 3D indoor scenes from sparse views. Our method exploits inter-image matching information and utilizes triangulation to provide more accurate depth information than monocular depth, thereby enhancing the stability of the reconstruction process. In addition, we design a projection loss based on pixel-to-pixel matching relationships in the images to ensure consistency across views. To refine accuracy further, we design an angular filter and an epipolar weight function. This helps remove wrong potential matches that might harm the final results. Extensive experiments demonstrate that our method outperforms all existing indoor reconstruction approaches. With only a limited number of views available, we achieve satisfactory reconstruction results on both real and synthetic datasets.

Acknowledgments

The corresponding author is Ge Gao. This work was supported by Beijing Science and Technology Program (Z231100001723014).

References

Barron, J. T.; Mildenhall, B.; Tancik, M.; Hedman, P.; Martin-Brualla, R.; and Srinivasan, P. P. 2021. Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 5855–5864.

Barron, J. T.; Mildenhall, B.; Verbin, D.; Srinivasan, P. P.; and Hedman, P. 2022. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5470–5479.

Chen, A.; Xu, Z.; Zhao, F.; Zhang, X.; Xiang, F.; Yu, J.; and Su, H. 2021. Mvsnerf: Fast generalizable radiance field reconstruction from multi-view stereo. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 14124–14133.

Cong, W.; Liang, H.; Wang, P.; Fan, Z.; Chen, T.; Varma, M.; Wang, Y.; and Wang, Z. 2023. Enhancing nerf akin to enhancing llms: Generalizable nerf transformer with mixture-of-view-experts. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3193–3204.

Dai, A.; Chang, A. X.; Savva, M.; Halber, M.; Funkhouser, T.; and Nießner, M. 2017. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 5828–5839.

Ding, Y.; Yuan, W.; Zhu, Q.; Zhang, H.; Liu, X.; Wang, Y.; and Liu, X. 2022. Transmvsnet: Global context-aware multi-view stereo network with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 8585–8594.

Edstedt, J.; Sun, Q.; Bökman, G.; Wadenbäck, M.; and Felsberg, M. 2023. RoMa: Revisiting Robust Losses for Dense Feature Matching. *arXiv preprint arXiv:2305.15404*.

Eftekhar, A.; Sax, A.; Malik, J.; and Zamir, A. 2021. Omnidata: A Scalable Pipeline for Making Multi-Task Mid-Level Vision Datasets From 3D Scans. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 10786–10796.

Gao, Y.; Cao, Y.-P.; and Shan, Y. 2023. SurfelNeRF: Neural Surfel Radiance Fields for Online Photorealistic Reconstruction of Indoor Scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 108–118.

Gropp, A.; Yariv, L.; Haim, N.; Atzmon, M.; and Lipman, Y. 2020. Implicit geometric regularization for learning shapes. *arXiv preprint arXiv:2002.10099*.

Guo, H.; Peng, S.; Lin, H.; Wang, Q.; Zhang, G.; Bao, H.; and Zhou, X. 2022. Neural 3d scene reconstruction with the manhattan-world assumption. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5511–5520.

Han, L.; Zhou, J.; Liu, Y.-S.; and Han, Z. 2024. Binocular-Guided 3D Gaussian Splatting with View Consistency for Sparse View Synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Huang, H.; Wu, Y.; Zhou, J.; Gao, G.; Gu, M.; and Liu, Y. 2023. NeuSurf: On-Surface Priors for Neural Surface Reconstruction from Sparse Input Views. *arXiv preprint arXiv:2312.13977*.

Irshad, M. Z.; Zakharov, S.; Liu, K.; Guizilini, V.; Kollar, T.; Gaidon, A.; Kira, Z.; and Ambrus, R. 2023. Neo 360: Neural fields for sparse view synthesis of outdoor scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9187–9198.

Johari, M. M.; Lepoittevin, Y.; and Fleuret, F. 2022. Geonerf: Generalizing nerf with geometry priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18365–18375.

Kazhdan, M.; Bolitho, M.; and Hoppe, H. 2006. Poisson surface reconstruction. In *Proceedings of the fourth Eurographics symposium on Geometry processing*, volume 7, 0.

Levy, D.; Peleg, A.; Pearl, N.; Rosenbaum, D.; Akkaynak, D.; Korman, S.; and Treibitz, T. 2023. SeaThru-NeRF: Neural Radiance Fields in Scattering Media. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 56–65.

Liang, Z.; Huang, Z.; Ding, C.; and Jia, K. 2023. 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*, 13165–13174.

Long, X.; Lin, C.; Wang, P.; Komura, T.; and Wang, W. 2022. Sparseneus: Fast generalizable neural surface reconstruction from sparse views. In *European Conference on Computer Vision*, 210–227. Springer.

Marí, R.; Facciolo, G.; and Ehret, T. 2022. Sat-nerf: Learning multi-view satellite photogrammetry with transient objects and shadow modeling using rpc cameras. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1311–1321.

Mildenhall, B.; Srinivasan, P. P.; Tancik, M.; Barron, J. T.; Ramamoorthi, R.; and Ng, R. 2020. NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis. In Vedaldi, A.; Bischof, H.; Brox, T.; and Frahm, J.-M., eds., *Computer Vision – ECCV 2020*, 405–421. Cham: Springer International Publishing. ISBN 978-3-030-58452-8.

Müller, T.; Evans, A.; Schied, C.; and Keller, A. 2022. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Trans. Graph.*, 41(4).

Reiser, C.; Peng, S.; Liao, Y.; and Geiger, A. 2021. Kilonerf: Speeding up neural radiance fields with thousands of tiny mlps. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 14335–14345.

Rematas, K.; Liu, A.; Srinivasan, P. P.; Barron, J. T.; Tagliasacchi, A.; Funkhouser, T.; and Ferrari, V. 2022. Urban radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12932–12942.

Ren, Y.; Zhang, T.; Pollefeys, M.; Süsstrunk, S.; and Wang, F. 2023. 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*, 16685–16695.

Roessle, B.; Barron, J. T.; Mildenhall, B.; Srinivasan, P. P.; and Nießner, M. 2022. Dense depth priors for neural radiance fields from sparse input views. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12892–12901.

Schonberger, J. L.; and Frahm, J.-M. 2016. Structure-from-motion revisited. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4104–4113.

Song, J.; Park, S.; An, H.; Cho, S.; Kwak, M.-S.; Cho, S.; and Kim, S. 2023. DäRF: Boosting Radiance Fields from Sparse Inputs with Monocular Depth Adaptation. arXiv:2305.19201.

Straub, J.; Whelan, T.; Ma, L.; Chen, Y.; Wijmans, E.; Green, S.; Engel, J. J.; Mur-Artal, R.; Ren, C.; Verma, S.; et al. 2019. The Replica dataset: A digital replica of indoor spaces. *arXiv preprint arXiv:1906.05797*.

Sun, C.; Sun, M.; and Chen, H.-T. 2022. Direct voxel grid optimization: Super-fast convergence for radiance fields reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5459–5469.

Tancik, M.; Casser, V.; Yan, X.; Pradhan, S.; Mildenhall, B.; Srinivasan, P. P.; Barron, J. T.; and Kretzschmar, H. 2022. Block-nerf: Scalable large scene neural view synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8248–8258.

Turki, H.; Ramanan, D.; and Satyanarayanan, M. 2022. Mega-nerf: Scalable construction of large-scale nerfs for virtual fly-throughs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12922–12931.

Uy, M. A.; Martin-Brualla, R.; Guibas, L.; and Li, K. 2023. SCADE: NeRFs from Space Carving with Ambiguity-Aware Depth Estimates. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16518–16527.

Wang, J.; Wang, P.; Long, X.; Theobalt, C.; Komura, T.; Liu, L.; and Wang, W. 2022a. Neuris: Neural reconstruction of indoor scenes using normal priors. In *European Conference on Computer Vision*, 139–155. Springer.

Wang, P.; Liu, L.; Liu, Y.; Theobalt, C.; Komura, T.; and Wang, W. 2021. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. *arXiv preprint arXiv:2106.10689*.

Wang, X.; Dong, S.; Zheng, Y.; and Yang, Y. 2024. InfoNorm: Mutual Information Shaping of Normals for Sparse-View Reconstruction. *arXiv preprint arXiv:2407.12661*.

Wang, Y.; Han, Q.; Habermann, M.; Daniilidis, K.; Theobalt, C.; and Liu, L. 2023. Neus2: Fast learning of neural implicit surfaces for multi-view reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 3295–3306.

Wang, Y.; Li, Y.; Liu, P.; Dai, T.; and Xia, S.-T. 2022b. NeXT: Towards High Quality Neural Radiance Fields via Multi-skip Transformer. In *European Conference on Computer Vision*, 69–86. Springer.

Wu, H.; Graikos, A.; and Samaras, D. 2023. S-VolSDF: Sparse Multi-View Stereo Regularization of Neural Implicit Surfaces. *arXiv preprint arXiv:2303.17712*.

Xu, L.; Guan, T.; Wang, Y.; Liu, W.; Zeng, Z.; Wang, J.; and Yang, W. 2023. C2F2NeUS: Cascade Cost Frustum Fusion for High Fidelity and Generalizable Neural Surface Reconstruction. *arXiv preprint arXiv:2306.10003*.

Yao, Y.; Luo, Z.; Li, S.; Fang, T.; and Quan, L. 2018. Mvsnet: Depth inference for unstructured multi-view stereo. In *Proceedings of the European conference on computer vision (ECCV)*, 767–783.

Yariv, L.; Gu, J.; Kasten, Y.; and Lipman, Y. 2021. Volume rendering of neural implicit surfaces. *Advances in Neural Information Processing Systems*, 34: 4805–4815.

Ye, B.; Liu, S.; Li, X.; and Yang, M.-H. 2023. Self-Supervised Super-Plane for Neural 3D Reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 21415–21424.

Ying, H.; Jiang, B.; Zhang, J.; Xu, D.; Yu, T.; Dai, Q.; and Fang, L. 2023. PARF: Primitive-Aware Radiance Fusion for Indoor Scene Novel View Synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 17706–17716.

Yu, A.; Li, R.; Tancik, M.; Li, H.; Ng, R.; and Kanazawa, A. 2021. Plenoctrees for real-time rendering of neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 5752–5761.

Yu, Z.; Peng, S.; Niemeyer, M.; Sattler, T.; and Geiger, A. 2022. Monosdf: Exploring monocular geometric cues for neural implicit surface reconstruction. *Advances in neural information processing systems*, 35: 25018–25032.

Zhang, W.; Xing, R.; Zeng, Y.; Liu, Y.-S.; Shi, K.; and Han, Z. 2023a. Fast Learning Radiance Fields by Shooting Much Fewer Rays. *IEEE Transactions on Image Processing*, 32: 2703–2718.

Zhang, X.; Kundu, A.; Funkhouser, T.; Guibas, L.; Su, H.; and Genova, K. 2023b. Nerflets: Local radiance fields for efficient structure-aware 3d scene representation from 2d supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8274–8284.