NeRFPrior: Learning Neural Radiance Field as a Prior for Indoor Scene Reconstruction

Anonymous CVPR submission

Paper ID 9045

Abstract

Recently, it has shown that priors are vital for neural implicit functions to reconstruct high-quality surfaces from multi-view RGB images. However, current priors require large-scale pre-training, and merely provide geometric clues without considering the importance of color. In this paper, we present NeRFPrior, which adopts a neural radiance field as a prior to learn signed distance fields using volume rendering for surface reconstruction. Our NeRF prior can provide both geometric and color clues, and also get trained fast under the same scene without additional data. Based on the NeRF prior, we are enabled to learn a signed distance function (SDF) by explicitly imposing a multi-view consistency constraint on each ray intersection for surface inference. Specifically, at each ray intersection, we use the density in the prior as a coarse geometry estimation, while using the color near the surface as a clue to check its visibility from another view angle. For the textureless areas where the multi-view consistency constraint does not work well, we further introduce a depth consistency loss with confidence weights to infer the SDF. Our experimental results outperform the state-of-the-art methods under the widely used benchmarks. The source code will be publicly available.

1. Introduction

3D surface reconstruction from multi-view images is a long-standing challenge in computer vision and graphics. Traditional methods, like multi-view-stereo (MVS) [15, 31, 42], estimate 3D geometry by first extracting a sparse point cloud and then applying dense reconstruction on it. The latest reconstruction methods [28, 37, 43] learn implicit functions from multiple images via volume rendering using neural networks. These methods require learning priors [18, 38, 46, 47] from an additional large-scale dataset to reveal accurate geometry and structure. However, these data-driven priors do not generalize well to other kinds of scenes which are different from the pretrained datasets, which drastically degenerates the performance.

Instead, some methods [4, 9, 12] introduce overfitting based priors to improve the generalization, since these priors can be learned by directly overfitting a single scene. Methods like MVS are widely adopted to extract overfitting priors, which use the photometric consistency to overfit a scene. However, these priors can merely provide geometric information and do not provide photometric information which is important for the network to predict colors in volume rendering.

To address this issue, we propose NeRFPrior, which introduces a neural radiance field as a prior to learn signed distance functions (SDF) to reconstruct smooth and high-quality surfaces from multi-view images. Thanks for current advanced training techniques for radiance fields [5, 11, 21, 27, 33], we are able to train a radiance field by overfitting multi-view images of a scene in minutes. Although more recent 3DGS methods [21] present a very promising solution for learning radiance fields with explicit 3D Gaussians, it is still a challenge to recover continuous SDFs from discrete, scattered, or even sparse 3D Gaussians. Per this, we adopt NeRF and leverage the trained NeRF as a prior to provide the geometry and color information of the scene itself. This enables us to learn a more precise SDF by explicitly imposing a multi-view consistency constraint on each ray intersection for its SDF inference.

Specifically, to get the prior geometry, we query the density from the NeRF prior as an additional supervision for our neural implicit networks. With the predicted density at each sample along a ray, we find the intersection with the surface, and then, we use the prior color to determine whether this intersection is visible from another view. If it is visible, our multi-view constraint is triggered to make this intersection participate in the rendering along the two rays for better surface inference. For the textureless areas where the multi-view consistency constraint does not work well, we further introduce a depth consistency loss with confidence weights to improve the completeness and smoothness of the surface. Our method does not require additional datasets to learn priors or suffer from a generalization issue. Our experimental results outperform the state-of-the-art methods under widely used benchmarks. Our contributions are listed below.

- We propose NeRFPrior to reconstruct accurate and smooth scene surfaces by exploiting NeRF as a prior. Such prior is learned by merely overfitting the scene to be reconstructed, without requiring any additional large-scale datasets.
- We introduce a novel strategy to impose a multi-view consistency constraint using our proposed NeRFPrior, which reveals more accurate surfaces.
- We propose a novel depth consistency loss with confidence weights to improve the smoothness and completeness of reconstructed surfaces for textureless areas in the real-world scenes.

2. Related Work

2.1. Multi-view Reconstruction

Multi-view reconstruction aims at reconstructing 3D surfaces from a given set of uncalibrated multi-view images. Traditional multi-view reconstruction pipeline is split into two stages: the structure-from-motion (SFM) [31] and the multi-view-stereo (MVS) [13, 15]. MVSNet [42] is the first to introduce the learning-based idea into traditional MVS methods. It applies 3D CNN on a plane-swept cost volume for depth estimation and outperforms the classical traditional methods [22]. Many following studies improve MVSNet from different aspects, such as training speed [39, 45], memory consumption [16, 40] and network structure [7, 10].

2.2. Neural Surface Reconstruction

Recently, NeRF [26] has achieved impressive results in novel view synthesis and has attracted lots of follow-up work. NeRF uses a single continuous 5D coordinate to represent the scene and predicts per-point density and color to render novel views using volume rendering algorithm. Following studies develop the potential of NeRF in various aspects, such as generation [25, 49], relighting [30, 41], human [6, 14] and so on. More and more strategies have been recognized and applied in improving NeRF performances, such as integrated positional encoding [2], voxelization [5, 33] and patch loss [12]. Among the studies improving the rendering performance of NeRF, Mip-NeRF [2] and Mip-NeRF-360 [3] aim at avoiding aliased images in novel view synthesis by considering the conical frustum area instead of the ray interval. PixelNeRF [44] makes use of image features to improve NeRF on the condition of sparse inputs. Other methods improve NeRF in generalization ability [24, 34] and some of them [20, 24] seek to exclude features of invisible image pixels to avoid providing misleading image priors to NeRF training.

Recent works [28, 37] investigate learning neural implicit fields from multi-view image inputs by differentiable ray marching. They replace the density field in NeRF with an implicit SDF field or occupancy field, which greatly improves the ability of reconstructing 3D geometry. More recently, many methods focus on variant kinds of priors to improve the reconstruction quality, for example, depth prior from MVS [4, 9], ground truth depth [1], estimated normals from pre-trained models [36, 38] and pre-trained semantic segmentation [50]. Latest methods infer SDF fields from 3D Gaussians [17, 19, 48]. However, they struggle to produce plausible surfaces because the geometry and color in 3D is not continuous with 3D Gaussians.

We notice that although the above-mentioned priors can improve the reconstruction quality to some extent, there still exist various shortcomings. Data-driven based priors used by the previous methods do not generalize well to different kinds of datasets, while overfitting priors can not provide photometric information for the network. To address the above problems, we propose NeRFPrior, which introduces a neural radiance field as a prior to learn implicit functions to reconstruct accurate surfaces without requiring any additional information from large-scale datasets.

3. Method

Given a set of posed images captured from a scene, we aim to learn neural implicit functions to reconstruct the scene without requiring any additional information from other datasets. We represent the geometry in the scene as a signed distance field and then extract the mesh using marching cubes algorithm. In this section, we first discuss the insight of adopting neural radiance field as a prior. Then we introduce multi-view consistency constraint and the depth consistency loss with confidence weights as two of our contributions to improve the reconstruction quality. An overview of our framework is provided in Fig. 1.

3.1. Neural Radiance Field Prior

NeRF [26] models a static scene using a continuous 5D function which takes a 3D coordinate and a corresponding viewing direction as input and outputs per-point density σ and color $\mathbf{c}$. Specifically, let $\mathbf{x}_i$ denotes the i-th sampled point along the ray $\mathbf{r}$, and $\mathbf{d}$ denotes the viewing direction. The predicted ray color $\hat{C}(\mathbf{r})$ is obtained by volume rendering below:

$$\hat{C}(\mathbf{r}) = \sum_{i=1}^{N} T_i(1 - \exp(-\sigma_\theta(\mathbf{x}_i)\delta_i))\mathbf{c}_\phi(\mathbf{x}_i, \mathbf{d}) \\ T_i = \exp(-\sum_{k=1}^{i-1} \sigma_\theta(\mathbf{x}_k)\delta_k), \tag{1}$$

where δ_i and T_i represent the sampling interval and the accumulated transmittance of the ray $\mathbf{r}$ at i-th sampled point,

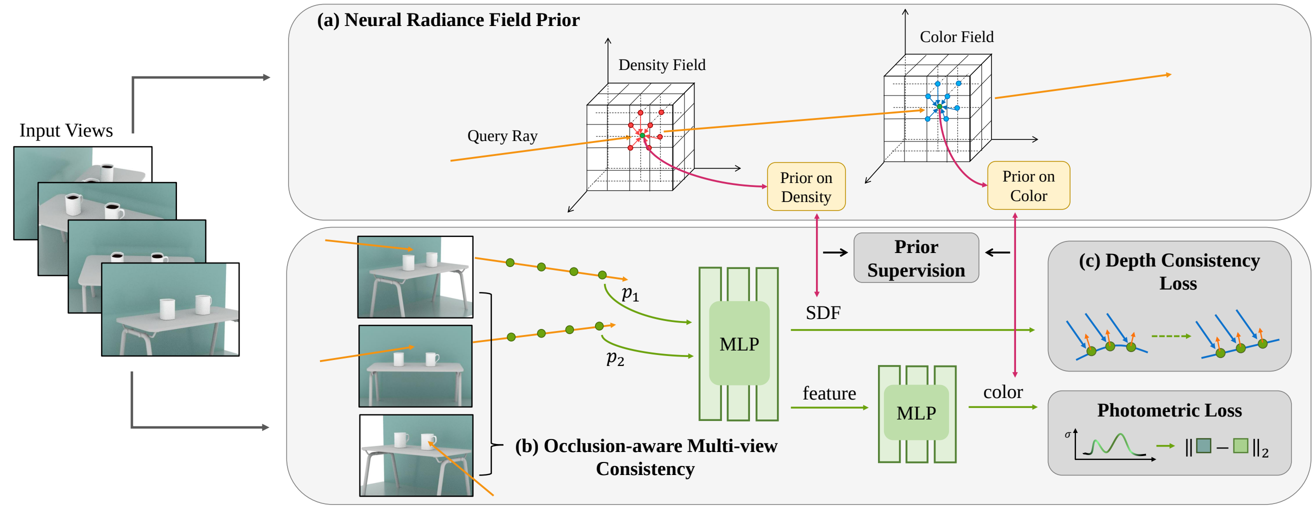


Figure 1. An overview of our NeRFPrior method. Given multi-view images of a scene as input, we first train a grid-based NeRF to obtain the density field and color field as priors. We then learn a signed distance function by imposing a multi-view consistency constraint using volume rendering. For each sampled point on the ray, we query the prior density and prior color as additional supervision of the predicted density and color, respectively. To improve the smoothness and completeness of textureless areas in the scene, we propose a depth consistency loss, which forces surface points in the same textureless plane to have similar depths.

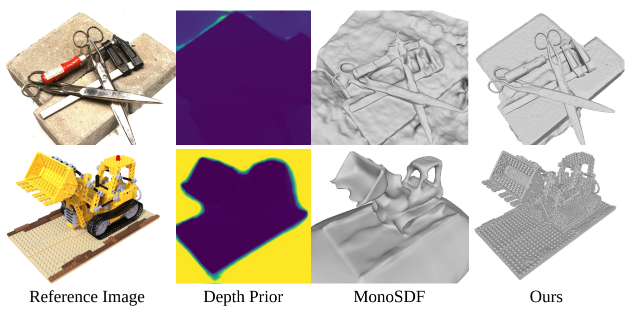


Figure 2. Comparison on object-surrounding scenes between MonoSDF and ours. The performance of MonoSDF drastically degenerates because the depth prior cannot generalize well to different kinds of datasets.

respectively. θ and ϕ are the parameters of the density and color networks, respectively.

Recently, there has been a number of studies combining NeRF framework and implicit functions to reconstruct 3D surfaces. However, the advanced NeRF techniques [5, 11, 27, 33] inspire us that NeRF itself can serve as a prior for surface reconstruction. Compared to NeRF-based surface reconstruction methods [28, 37, 50], we have the ability to explicitly use geometry and color information from the field for visibility check and imposing multi-view depth consistency constraints. This design has two main advantages. Firstly, our NeRF prior is able to provide color cues for optimization, which is missing in other methods combining priors [12, 46].

Secondly, our prior is easily accessed compared to existing prior acquisition methods. Data-driven priors such as depth and normal priors [36, 38, 46], need days of pre-training on large-scale datasets. Additionally, data-driven priors do not generalize well to different kinds of scenes, as shown in Fig. 2. The prior of MonoSDF is pretrained on indoor scene datasets, so the quality of prior degenerates while generalizing to object-surrounding datasets. On the other hand, overfitting priors such as sparse depth and sparse point cloud from COLMAP argorithm [12, 18, 38], are sparse and incontinuous that most pixels or points cannot be supervised. And it lacks the supervision of color. Thanks for the advance in NeRF training acceleration, we can optimize a grid-based NeRF, which can be trained in minutes. Additionally, the grid-based structure has advantages in perceiving high-frequency surface details, which is beneficial to our accurate reconstruction.

As shown in Fig. 1 (a), to obtain the neural radiance field prior from multi-view images, we firstly construct a pair of density grid $F_\sigma \in \mathbb{R}^{[N_1,N_2,N_3,1]}$ and color feature grid $F_\mathbf{c} \in \mathbb{R}^{[N_1,N_2,N_3,d]}$, where N_1, N_2, N_3 are the resolutions of the feature grids, and d is the feature length of color grid. For a 3D point $\mathbf{x}$ sampled along the rendering ray with viewing direction $\mathbf{d}$, the density and color are interpolated from the feature grids of the trained NeRF, as denoted by

$$\begin{aligned} \sigma_{prior}(\mathbf{x}) &= \mathrm{act}(\mathrm{interp}(F_\sigma, \mathbf{x})) \\ \mathbf{c}_{prior}(\mathbf{x}, \mathbf{d}) &= \mathrm{act}(\mathrm{MLP}(\mathrm{interp}(F_\mathbf{c}, \mathbf{x}), \mathbf{d})), \end{aligned} \tag{2}$$

where the operation act represents activation function and interp represents trilinear interpolation, respectively. For color prediction, we use an additional shallow MLP to take viewing direction into consider. The network is trained using volume rendering and then frozen as our NeRF prior.

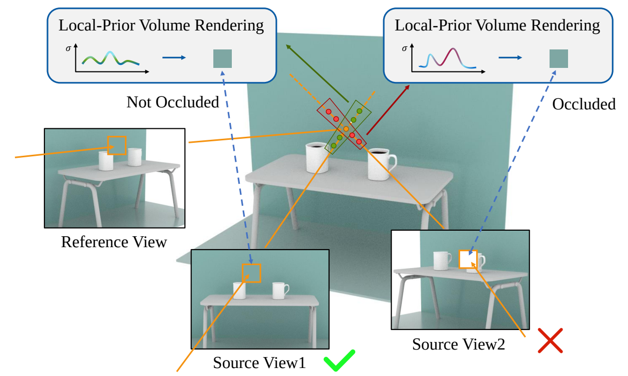


Figure 3. An illustration of our multi-view consistency constraint. To judge the visibility of the intersection, we conduct a local-prior volume rendering around the intersection and compare the rendering color with the projection color. The ray from source view is participated in training only if the intersection is visible along this ray.

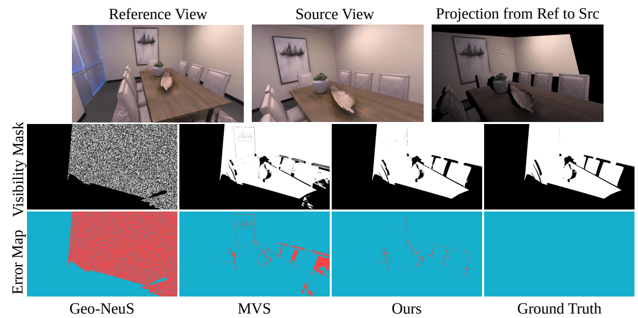


Figure 4. A comparison on the accuracy of visibility check. The first row shows the ground truth result of projecting pixels from reference view to source view. The second row shows the visibility mask, indicating which points in the reference view are visible after projection. The third row is the error map of visibility check.

Following [37], we further integrate the signed distance field into neural surface reconstruction by learning SDF to represent density in volume rendering:

$$\sigma(\mathbf{x}) = \max\left(\frac{-\Phi'(f_s(\mathbf{x}))}{\Phi(f_s(\mathbf{x}))}, 0\right), \tag{3}$$

where $\mathbf{x}$ represents the sampled point along the ray. Φ and f_s represent sigmoid function and SDF network, respectively. To combine the prior field and the signed distance field together, we query the density and color of each sampled point from the prior fields and use them as supervision of the predicted density and color from neural implicit network:

$$\begin{aligned}\mathcal{L}_\sigma &= \|\hat{\sigma}(\mathbf{x}) - \sigma_{prior}(\mathbf{x})\|_1 \\ \mathcal{L}_c &= \|\hat{\mathbf{c}}(\mathbf{x}, \mathbf{d}) - \mathbf{c}_{prior}(\mathbf{x}, \mathbf{d})\|_1.\end{aligned} \tag{4}$$

We notice that the prior density field is usually noisy, which may mislead the neural implicit network. Therefore, we use a threshold to filter out the fuzzy density value and apply supervision only if the density value is convincing. The filtering strategy will be discussed in the supplementary in detail. Benefiting from the NeRF prior, we are able to learn the signed distance field to reconstruct accurate 3D geometry details at a fast speed.

3.2. Multi-view Consistency Constraint

Multi-view consistency is a key intuition for geometry extraction because the photometric consistency information existed in the multi-view images is a powerful prompt to help revealing the surface. To reconstruct accurate 3D surfaces, we explicitly impose a multi-view consistency constraint on each ray for its SDF inference. Specifically, for an emitted ray $\mathbf{r}_m$ from a reference view I_m, we firstly apply root finding [28] to locate the intersection point $\mathbf{p}^*$ where the ray hits the surface. Then we select several nearby images as source views. For each source view, we emit an additional ray from the camera viewpoint to the intersection $\mathbf{p}^*$. The ray from reference view and the rays from source views are gathered and fed into volume rendering in parallel. An intuition of this idea is that the network is enabled to inference the zero-level-set of the intersection from the photometric difference of multi-view images, as shown in Fig. 1 (b) and detailed in Fig. 3. While emitting multi-view rays towards an intersection, some rays may be blocked by some objects in front of the intersection. To resolve this issue, we use our prior field to conduct a local-prior volume rendering for visibility check. Specifically, to determine the visibility of intersection $\mathbf{p}^*$ from source view I_s with viewing direction $\mathbf{r}_s$, we sample M points in a small interval $[d_s^* - \Delta, d_s^* + \Delta]$ centered at $\mathbf{p}^*$ along $\mathbf{r}_s$, where d_s^* is the distance between $\mathbf{p}^*$ and the viewpoint of I_s. Next we apply volume rendering on the sampled points using the queried prior density and prior color:

$$\mathbf{c}_s^* = \sum_{k=1}^{M} T_k(1 - \exp(-\sigma_{prior}(\mathbf{x}_k)\delta))\mathbf{c}_{prior}(\mathbf{x}_k, \mathbf{d}(\mathbf{r}_s)),$$
$$T_k = \exp(-\sum_{q=1}^{k-1} \sigma_{prior}(\mathbf{x}_q)\delta), \tag{5}$$

where $\mathbf{d}(\mathbf{r}_s)$ represents the viewing direction of $\mathbf{r}_s$. In practice, we typically set $\Delta = 0.1$, $M = 64$ and $\delta = 0.003$. The rendered color $\mathbf{c}_s^*$ is compared with the pixel color $\mathbf{c}_s^{proj}$, which is the projection of $\mathbf{p}^*$ on the source view I_s. If the two colors differ a lot, we consider that $\mathbf{p}^*$ is invisible from I_s, otherwise visible:

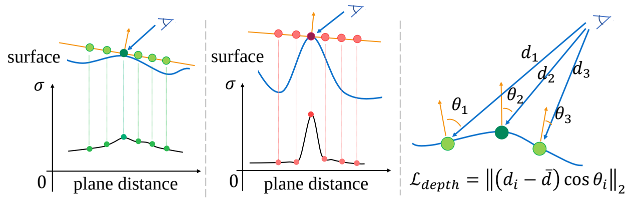
$$\mathbf{p}^* = \begin{cases} \text{visible} & |\mathbf{c}_s^* - \mathbf{c}_s^{proj}| < t_0 \\ \text{invisible} & |\mathbf{c}_s^* - \mathbf{c}_s^{proj}| \geq t_0 \end{cases} \tag{6}$$

If $\mathbf{p}^*$ is visible, we then emit the ray $\mathbf{r}_s$ for volume rendering together with the ray $\mathbf{r}_m$ from the reference view.

CVPR #9045

Our visibility check is more robust than traditional MVS methods which directly match the projection color on two views, since the color of projections is significantly biased on illumination. Our NeRFPrior resolves this issue by predicting view-dependent color. Although the standard volume rendering needs sampling in a fairly long interval, we observe that due to the pulse characteristics of density, only a small interval is enough for volume rendering to get accurate color in the pretrained NeRF. Fig. 4 provides an example. Comparing to Geo-NeuS [12] which uses patched normalization cross correlation (NCC) to judge visibility and MVS [31] which depends on projection color to judge visibility, our method achieves significantly more accurate results.

3.3. Depth Consistency Loss



(a) Plane Detected **(b) Plane Not Detected** **(c) Depth Consistency Loss**

Figure 5. An illustration of our depth consistency loss. We calculate the density variance of the intersection and its neighboring points on the tangent plane. If (a) the variance is small, we constrain these points to maintain the same depth on normal directions as in (c). Otherwise, (b) we do not impose depth constraints.

It is hard for neural implicit functions to infer accurate surfaces in textureless areas in indoor scenes such as walls and floors, due to the lack of distinctive color information. We further propose a depth consistency loss with confidence weights to improve the smoothness and completeness in textureless areas. We observe that continuous textureless areas usually have consistent or continuously varying colors, and are ususally composed of planes [38]. Hence, we use density distribution as a clue to determine whether the neighboring area of an intersection is a plane, and then add depth consistency constraints if it is the case, as shown in Fig. 1 (c) and detailed in Fig. 5.

In order to impose depth consistency constraints on surface points, two prerequisites are needed: (i) the intersection and its neighboring points have similar colors on the projection view, (ii) the intersection and its neighboring points are nearly on a plane. For (i), we calculate the color variance of each pixel and its neighboring pixels on the input views. For (ii), we calculate density variance of the intersection and its neighboring points as a confidence to judge whether a surface is a plane. If the density variance and the color variance are both small, we assume that the ray hits a plane. Then we constrain the neighborhood points to maintain the same depth on their normal directions. Otherwise, we do not impose depth constraints. Formally, let $\mathbf{p}^*$ be the intersection between ray $\mathbf{r}$ and the object surface, $\mathbf{c}^{proj}$ be the projection pixel color of $\mathbf{p}^*$ on the source view. The depth loss can be written as following:

$$\mathcal{L}_{depth} = \sum_{\mathbf{r}\in\mathcal{R}} \|(\hat{D}(\mathbf{r}) - \bar{D}) \cos\langle \mathbf{n}, \mathbf{r}\rangle\|_2 * \mathrm{sgn}_c * \mathrm{sgn}_\sigma \quad (7)$$

$$\begin{aligned} \mathrm{sgn}_c &= \begin{cases} 1 & \mathrm{var}(\mathbf{c}^{proj}) < t_1 \\ 0 & \mathrm{var}(\mathbf{c}^{proj}) \geq t_1 \end{cases} \\ \mathrm{sgn}_\sigma &= \begin{cases} 1 & \mathrm{var}(\sigma(\mathbf{p}^*)) < t_2 \\ 0 & \mathrm{var}(\sigma(\mathbf{p}^*)) \geq t_2 \end{cases} \end{aligned} \quad (8)$$

where $\hat{D}(\mathbf{r})$ is the rendered depth of ray $\mathbf{r}$ and $\bar{D}$ is the mean depth in a batch of rays $\mathcal{R}$, which are emitted from some neighboring pixels. $\mathbf{n}$ is the rendered normal vector of ray $\mathbf{r}$, and var represents the variation. In a word, only when the intersection is on a plane and it is in the textureless areas of the image, we constrain the depth of the intersection to keep similar with the depth of its neighboring intersections.

3.4. Loss Function

We render the color of each ray using Eq. (1) and measure the error between rendered color and ground truth pixel color:

$$\mathcal{L}_{rgb} = \sum_{\mathbf{r}\in\mathcal{R}} \|\hat{C}(\mathbf{r}) - C(\mathbf{r})\|_1, \quad (9)$$

where $\mathcal{R}$ denotes all of the rays in a training batch. Following [37], we add an Eikonal term on the sampled points to regularize the SDF field by

$$\mathcal{L}_{reg} = \frac{1}{N}\sum_i \|\nabla f_s(\mathbf{p}_i) - 1\|_2, \quad (10)$$

where $\mathbf{p}_i$ is the sampled point on the ray and N is the number of sampled points.

With our additional prior field supervision (Eq. 4) and depth loss (Eq. 7), the overall loss function can be written as

$$\mathcal{L} = \mathcal{L}_{rgb} + \lambda_1\mathcal{L}_\sigma + \lambda_2\mathcal{L}_c + \lambda_3\mathcal{L}_{reg} + \lambda_4\mathcal{L}_{depth}. \quad (11)$$

4. Experiments

4.1. Implementation Details

To train a neural radiance field as our NeRF prior, we adopt the grid-based architecture of TensoRF [5]. We train the prior NeRF for each scene in 30k iterations, which takes about 30 minutes per scene. For our implicit surface function, we adopt the architecture of NeuS [37], where the signed distance function and color function are modeled by an MLP with 8 and 6 hidden layers, respectively. We train our implicit surface function for 200k iterations in total. The multi-view consistency constraint is applied after

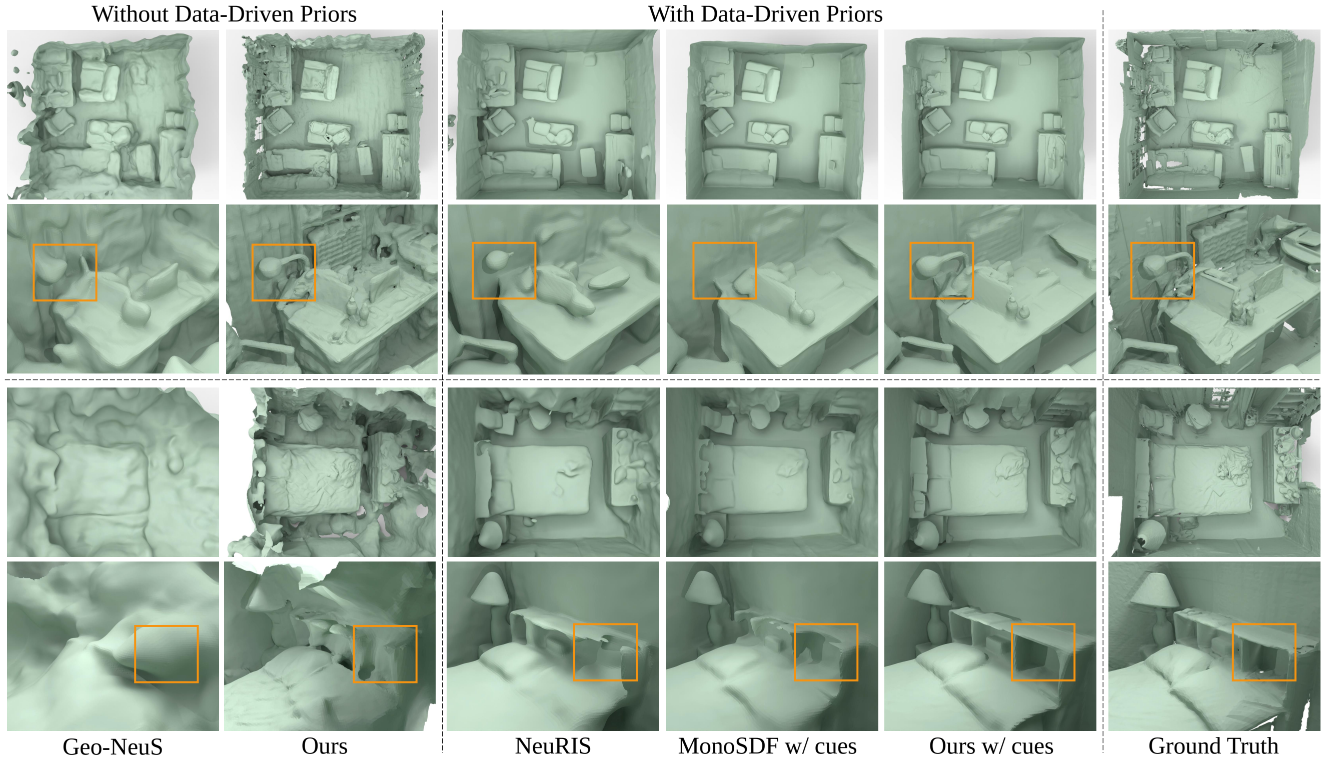


Figure 6. Visualization comparison on ScanNet Dataset.

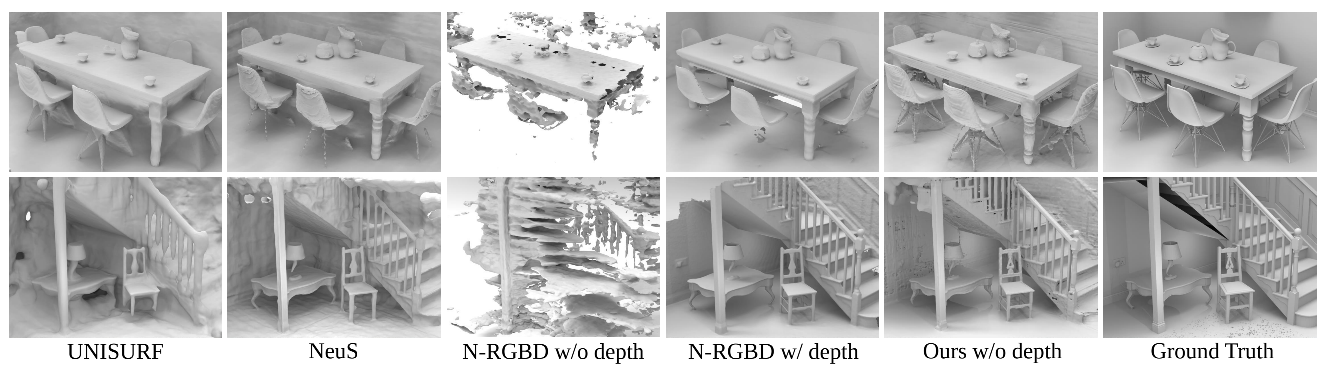


Figure 7. Visualization comparison on BlendSwap Dataset.

100k iterations and the depth consistency loss is applied after 150k iterations. We adopt such strategy based on the observation that the multi-view consistency and depth loss may mislead the network at the early training stage when the surface is noisy and ambiguous. We set $t_0 = 0.02$ in Eq. (6), $t_1 = 0.04$ and $t_2 = 0.1$ in Eq. (8), $\lambda_1 = \lambda_2 = 0.1$ and decreases exponentially to 0, $\lambda_3 = 0.05$ and $\lambda_4 = 0.5$ in Eq. (11). The choice of hyperparameters and thresholds will be discussed in supplementary in details. All the experiments are conducted on a single NVIDIA RTX 3090Ti GPU.

4.2. Experimental Settings

Datasets. We evaluate our method quantitatively and qualitatively on real-captured dataset ScanNet [8]. Following previous works [46], we use 4 scenes from ScanNet for our evaluation. We also evaluate our method under two synthetic scene datasets, including BlendSwap [1] and Replica [32], each of which contains 8 indoor scenes.

Baselines. We compare our method with the following state-of-the-art methods: (1) Classic MVS method: **COLMAP** [31]. (2) Neural raidance field methods without data-driven priors: **NeRF** [26], **UNISURF** [28], **NeuS** [37], **Geo-NeuS** [12], **PermutoSDF** [29], **NeuralAngelo** [23]. (3) Neural implicit reconstruction methods with data-

Figure 8. Visualization comparison on Replica Dataset.

Table 1. Evaluation results on ScanNet dataset. MonoSDF* represents MonoSDF with its monocular depth and normal cues.

Methods	Acc ↓	Comp ↓	Prec ↑	Recall ↑	F1 ↑
NeRF[26]	0.735	0.177	0.131	0.291	0.176
NeuS[37]	0.179	0.208	0.313	0.275	0.291
Geo-Neus[12]	0.236	0.206	0.282	0.313	0.291
MonoSDF[46]	0.214	0.180	0.297	0.325	0.310
PermutoSDF[29]	0.143	0.219	**0.448**	0.209	0.285
NeuralAngelo[23]	0.245	0.272	0.274	0.311	0.292
Ours	**0.133**	**0.120**	0.439	**0.429**	**0.433**
Manhattan[18]	0.072	0.068	0.621	0.586	0.602
NeuRIS[36]	0.054	0.052	0.729	0.684	0.705
MonoSDF*[46]	0.042	0.049	0.760	0.707	0.732
Ours (+monocular cues)	**0.037**	**0.042**	**0.799**	**0.766**	**0.782**
Go-Surf[35]	0.048	0.021	0.880	0.894	0.887
Ours (+depth)	**0.027**	**0.020**	**0.931**	**0.928**	**0.930**

Table 2. Evaluation results on BlendSwap dataset. Results are averaged among the 8 scenes.

Methods	CD ↓	NC ↑	Prec ↑	Recall ↑	F1 ↑
COLMAP[31]	0.420	0.556	0.429	0.353	0.387
UNISURF[28]	0.213	0.710	0.610	0.413	0.484
NeuS[37]	0.180	0.731	0.526	0.454	0.483
N-RGBD[1]	0.380	0.423	0.266	0.219	0.292
Ours	**0.088**	**0.813**	**0.651**	**0.594**	**0.621**

Table 3. Evaluation results on Replica dataset. Results are averaged among the 8 scenes.

Methods	CD ↓	NC ↑	Prec ↑	Recall ↑	F1 ↑
COLMAP[31]	0.232	0.468	0.455	0.408	0.430
UNISURF[28]	0.110	0.769	0.566	0.449	0.496
NeuS[37]	0.066	0.883	0.709	0.626	0.665
MonoSDF[46]	0.075	0.867	0.657	0.609	0.632
Ours	**0.038**	**0.912**	**0.833**	**0.795**	**0.813**

Table 4. Comparison of the total time of training pipeline.

Methods	Getting Priors	Training	Total
COLMAP[31]	10.7h	-	8.7h
NeuS[37]	-	7.2h	7.2h
Neural RGB-D[1]	-	10.3h	10.3h
Geo-NeuS[12]	1.5h	7.5h	9.0h
MonoSDF[46]	-	10.6h	10.6h
Ours	37min	4.2h	**4.7h**

driven priors: **Neural RGB-D** [1], **Manhattan-SDF** [18], **NeuRIS** [36], **MonoSDF** [46], **GO-Surf** [35].

Evaluation Metrics. For ScanNet dataset, following [46], we adopt Accuracy, Completeness, Precision, Recall and F1-score as evaluation metrics. For synthetic dataset, following [1], we adopt Chamfer Distance (CD), Normal Consistency (NC), Precision, Recall and F1-score as evaluation metrics. Please refer to the supplementary for more details on these metrics.

4.3. Quantitative and Qualitative Comparison

Evaluation on ScanNet Dataset. We report our evaluation on ScanNet datset in Tab. 1 and Fig. 6. The comparison is splitted into three parts. The first part is the comparison with the methods that do not use data-driven priors, including NeRF, NeuS, Geo-NeuS, MonoSDF without cues, PermutoSDF. The second part is the comparison with the methods that use data-driven priors, including Manhattan with pretrained segmentation priors, NeuRIS with pretrained normal priors, MonoSDF with estimated depth and normal cues (marked as "MonoSDF*"), and our results integrated with MonoSDF cues. The third part is the comparison with the methods that use ground truth depth supervision, including Go-Surf and our results with depth supervision. Our method exceeds other baselines without data-driven priors. On the other hand, integrated with monocular cues or ground truth depth supervision, our method also achieves the best performance comparing to other methods with priors. Visual comparisons in Fig. 6 show that our method is able to reconstruct complete and smooth surfaces and captures more scene details, such as the lamp and the bedside cupboard.

Evaluation on BlendSwap Dataset. We report our evaluation on BlendSwap dataset in Tab. 2 and Fig. 7. We compare our method with state-of-the-art methods that do not use

Table 5. Ablation study on each module of our method.

Base	NeRF prior	Multi-view	Depth loss	Reg term	CD ↓	NC ↑	F1 ↑
✓				✓	0.083	0.832	0.619
✓		✓	✓	✓	0.051	0.893	0.781
	✓			✓	0.049	0.763	0.673
✓	✓			✓	0.050	0.887	0.744
✓	✓	✓		✓	0.044	0.897	0.773
✓	✓	✓	✓		0.043	0.873	0.794
✓	✓	✓	✓	✓	**0.038**	**0.912**	**0.813**

data-driven priors, including COLMAP, UNISURF, NeuS and Neural-RGBD without ground truth depth supervision (marked as "N-RGBD"). The results show our brilliant ability of inferring implicit representations from multi-view images. Additionally, our advantages over our baseline "NeuS" highlight the benefits we get from the NeRF prior. Visual comparisons in Fig. 7 show that our reconstruction does not have artifacts, and contains more details with much higher accuracy than other methods.

Evaluation on Replica Dataset. We evaluate our method on Replica dataset, as shown in Tab. 3 and Fig. 8. We report comparisons with the latest methods, including COLMAP, UNISURF, NeuS and MonoSDF without cues. Qualitative results in Fig. 8 further demonstrate the advantages of our method on reconstructing complete, smooth and high fidelity surfaces.

Optimization Time. We evaluate the total time of training pipeline of different methods, including the time of obtaining priors and the time of training, as reported in Tab. 4. Benefiting from the advance in NeRF training acceleration [5], we are able to obtain our NeRF prior in half an hour, comparing to COLMAP which takes a long time in dense reconstruction. With the guidance of the NeRF prior, our network is able to converge fast in the early stage of training, which reduces the total training time by about 50% compared to current neural implicit function methods.

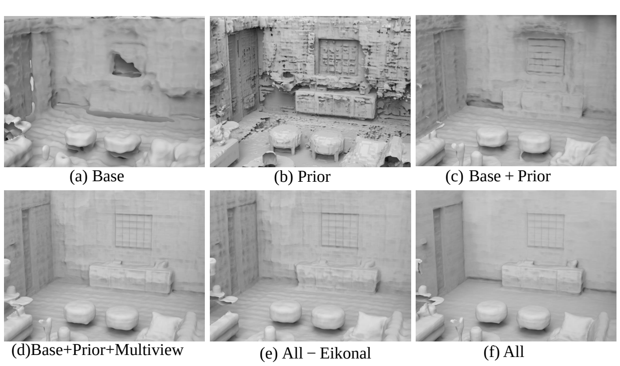


Figure 9. Ablation study on each module of our method.

4.4. Ablation Study

To demonstrate the effectiveness of our proposed components, we conduct ablation studies on Replica dataset, as reported in Tab. 5 and Fig. 9. We report our visualization and quantification results on 6 different settings: **(a)** only the base implicit function network, **(b)** only the NeRF prior, **(c)** the base network with our NeRF prior, **(d)** the base network with NeRF prior and the multi-view consistency constraint, **(e)** the complete method without eikonal regularization term, **(f)** our complete method. Our NeRF prior is able to perceive geometric details but shows very poor performance on consistency and smoothness, as shown in Fig. 9 (b). With the help of multi-view consistency constraint and depth consistency loss, we can reconstruct high fidelity scene surfaces.

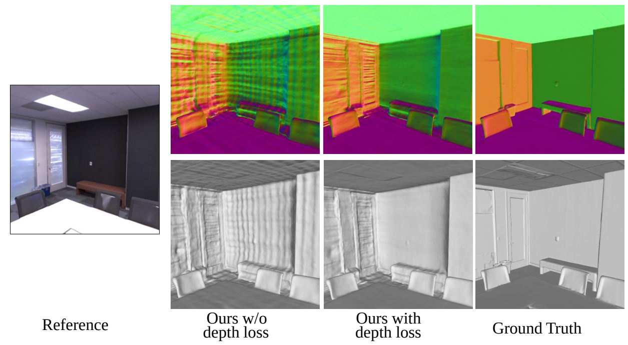


Figure 10. A visualization of the ablation on depth consistency loss. The first line is the normal map and the second line is the reconstructed mesh.

We further conduct an ablation study on depth consistency loss, as shown in Fig. 10. We select a room corner, where the input views contain lots of textureless areas. Our depth consistency loss greatly improves the consistency of surface normals and the smoothness of the textureless surfaces.

5. Conclusion

We propose NeRFPrior for reconstructing indoor scenes from multi-view images. We introduce to learn a NeRF as a prior which can be trained very fast to sense the geometry and color of a scene. With NeRF prior, we are enabled to use view-dependent color to check visibility, impose multi-view consistency constraints to infer SDF on the surface through volume rendering, and introduce a confidence weighted depth consistency loss to infer planes from textureless areas. Our method provides a novel perspective to learn neural implicit representations from multi-view images through volume rendering, which is much different from the latest methods merely using geometry prior learned in a data-driven or overfitting manner. Our method successfully learns more accurate implicit representations which produces smoother, sharper and more complete surfaces than the state-of-the-art methods. Our experimental results justify the effectiveness and superior of our method.

References

[1] Dejan Azinović, Ricardo Martin-Brualla, Dan B Goldman, Matthias Nießner, and Justus Thies. Neural RGB-D surface reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6290–6301, 2022. 2, 6, 7

[2] Jonathan T. Barron, Ben Mildenhall, Matthew Tancik, Peter Hedman, Ricardo Martin-Brualla, and Pratul P. Srinivasan. Mip-NeRF: A multiscale representation for anti-aliasing neural radiance fields. *Proceedings of International Conference on Computer Vision (ICCV)*, 2021. 2

[3] Jonathan T. Barron, Ben Mildenhall, Dor Verbin, Pratul P. Srinivasan, and Peter Hedman. Mip-NeRF 360: Unbounded anti-aliased neural radiance fields. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2022. 2

[4] Anpei Chen, Zexiang Xu, Fuqiang Zhao, Xiaoshuai Zhang, Fanbo Xiang, Jingyi Yu, and Hao Su. MVSNeRF: Fast generalizable radiance field reconstruction from multi-view stereo. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14124–14133, 2021. 1, 2

[5] Anpei Chen, Zexiang Xu, Andreas Geiger, Jingyi Yu, and Hao Su. Tensorf: Tensorial radiance fields. In *European Conference on Computer Vision (ECCV)*, pages 333–350. Springer, 2022. 1, 2, 3, 5, 8

[6] Mingfei Chen, Jianfeng Zhang, Xiangyu Xu, Lijuan Liu, Yujun Cai, Jiashi Feng, and Shuicheng Yan. Geometry-guided progressive nerf for generalizable and efficient neural human rendering. In *European Conference on Computer Vision*, pages 222–239. Springer, 2022. 2

[7] Rui Chen, Songfang Han, Jing Xu, and Hao Su. Point-based multi-view stereo network. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1538–1547, 2019. 2

[8] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. ScanNet: Richly-annotated 3D reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5828–5839, 2017. 6

[9] Kangle Deng, Andrew Liu, Jun-Yan Zhu, and Deva Ramanan. Depth-Supervised NeRF: Fewer views and faster training for free. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12882–12891, 2022. 1, 2

[10] Yikang Ding, Wentao Yuan, Qingtian Zhu, Haotian Zhang, Xiangyue Liu, Yuanjiang Wang, and Xiao Liu. TransMVSNet: Global context-aware multi-view stereo network with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8585–8594, 2022. 2

[11] Sara Fridovich-Keil, Alex Yu, Matthew Tancik, Qinhong Chen, Benjamin Recht, and Angjoo Kanazawa. Plenoxels: Radiance fields without neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5501–5510, 2022. 1, 3

[12] Qiancheng Fu, Qingshan Xu, Yew Soon Ong, and Wenbing Tao. Geo-NeuS: Geometry-consistent neural implicit surfaces learning for multi-view reconstruction. *Advances in Neural Information Processing Systems*, 35:3403–3416, 2022. 1, 2, 3, 5, 6, 7

[13] Yasutaka Furukawa and Jean Ponce. Accurate, dense, and robust multiview stereopsis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(8):1362–1376, 2009. 2

[14] Chen Geng, Sida Peng, Zhen Xu, Hujun Bao, and Xiaowei Zhou. Learning neural volumetric representations of dynamic humans in minutes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8759–8770, 2023. 2

[15] Michael Goesele, Brian Curless, and Steven M Seitz. Multi-view stereo revisited. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, pages 2402–2409. IEEE, 2006. 1, 2

[16] Xiaodong Gu, Zhiwen Fan, Siyu Zhu, Zuozhuo Dai, Feitong Tan, and Ping Tan. Cascade cost volume for high-resolution multi-view stereo and stereo matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2495–2504, 2020. 2

[17] Antoine Guédon and Vincent Lepetit. SuGaR: Surface-aligned gaussian splatting for efficient 3d mesh reconstruction and high-quality mesh rendering. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. 2

[18] Haoyu Guo, Sida Peng, Haotong Lin, Qianqian Wang, Guofeng Zhang, Hujun Bao, and Xiaowei Zhou. Neural 3D scene reconstruction with the Manhattan-world assumption. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5511–5520, 2022. 1, 3, 7

[19] Binbin Huang, Zehao Yu, Anpei Chen, Andreas Geiger, and Shenghua Gao. 2D Gaussian Splatting for Geometrically Accurate Radiance Fields. In *SIGGRAPH 2024 Conference Papers*. Association for Computing Machinery, 2024. 2

[20] M. M. Johari, Y. Lepoittevin, and F. Fleuret. GeoNeRF: Generalizing NeRF with geometry priors. *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition*, 2022. 2

[21] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3D gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4):1–14, 2023. 1

[22] Vladimir Kolmogorov and Ramin Zabih. Multi-camera scene reconstruction via graph cuts. In *European conference on computer vision*, pages 82–96. Springer, 2002. 2

[23] Zhaoshuo Li, Thomas Müller, Alex Evans, Russell H Taylor, Mathias Unberath, Ming-Yu Liu, and Chen-Hsuan Lin. Neuralangelo: High-fidelity neural surface reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8456–8465, 2023. 6, 7

[24] Yuan Liu, Sida Peng, Lingjie Liu, Qianqian Wang, Peng Wang, Christian Theobalt, Xiaowei Zhou, and Wenping Wang. Neural rays for occlusion-aware image-based rendering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2022. 2

[25] Gal Metzer, Elad Richardson, Or Patashnik, Raja Giryes, and Daniel Cohen-Or. Latent-nerf for shape-guided generation of 3d shapes and textures. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12663–12673, 2023. 2

[26] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. NeRF: Representing scenes as neural radiance fields for view synthesis. In *European Conference on Computer Vision (ECCV)*, pages 405–421. Springer, 2020. 2, 6, 7

[27] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Transactions on Graphics (ToG)*, 41(4):1–15, 2022. 1, 3

[28] Michael Oechsle, Songyou Peng, and Andreas Geiger. UNISURF: Unifying neural implicit surfaces and radiance fields for multi-view reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5589–5599, 2021. 1, 2, 3, 4, 6, 7

[29] Radu Alexandru Rosu and Sven Behnke. PermutoSDF: Fast multi-view reconstruction with implicit surfaces using permutohedral lattices. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8466–8475, 2023. 6, 7

[30] Viktor Rudnev, Mohamed Elgharib, William Smith, Lingjie Liu, Vladislav Golyanik, and Christian Theobalt. Nerf for outdoor scene relighting. In *European Conference on Computer Vision*, pages 615–631. Springer, 2022. 2

[31] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4104–4113, 2016. 1, 2, 5, 6, 7

[32] Julian Straub, Thomas Whelan, Lingni Ma, Yufan Chen, Erik Wijmans, Simon Green, Jakob J Engel, Raul Mur-Artal, Carl Ren, Shobhit Verma, et al. The replica dataset: A digital replica of indoor spaces. *arXiv preprint arXiv:1906.05797*, 2019. 6

[33] Cheng Sun, Min Sun, and Hwann-Tzong Chen. Direct voxel grid optimization: Super-fast convergence for radiance fields reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5459–5469, 2022. 1, 2, 3

[34] Alex Trevithick and Bo Yang. GRF: Learning a general radiance field for 3D representation and rendering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15182–15192, 2021. 2

[35] Jingwen Wang, Tymoteusz Bleja, and Lourdes Agapito. GoSurf: Neural feature grid optimization for fast, high-fidelity rgb-d surface reconstruction. In *2022 International Conference on 3D Vision (3DV)*, pages 433–442. IEEE, 2022. 7

[36] Jiepeng Wang, Peng Wang, Xiaoxiao Long, Christian Theobalt, Taku Komura, Lingjie Liu, and Wenping Wang. NeuRIS: Neural reconstruction of indoor scenes using normal priors. In *European Conference on Computer Vision*, 2022. 2, 3, 7

[37] Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. NeuS: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. *Advances in Neural Information Processing Systems*, 34, 2021. 1, 2, 3, 4, 5, 6, 7

[38] Yusen Wang, Zongcheng Li, Yu Jiang, Kaixuan Zhou, Tuo Cao, Yanping Fu, and Chunxia Xiao. NeuralRoom: Geometry-constrained neural implicit surfaces for indoor scene reconstruction. *ACM Transactions on Graphics (SIGGRAPH Asia)*, 2022. 1, 2, 3, 5

[39] Rafael Weilharter and Friedrich Fraundorfer. HighRes-MVSNet: A fast multi-view stereo network for dense 3D reconstruction from high-resolution images. *IEEE Access*, 9: 11306–11315, 2021. 2

[40] Jianfeng Yan, Zizhuang Wei, Hongwei Yi, Mingyu Ding, Runze Zhang, Yisong Chen, Guoping Wang, and Yu-Wing Tai. Dense hybrid recurrent multi-view stereo net with dynamic consistency checking. In *European Conference on Computer Vision*, pages 674–689. Springer, 2020. 2

[41] Siqi Yang, Xuanning Cui, Yongjie Zhu, Jiajun Tang, Si Li, Zhaofei Yu, and Boxin Shi. Complementary intrinsics from neural radiance fields and cnns for outdoor scene relighting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16600–16609, 2023. 2

[42] Yao Yao, Zixin Luo, Shiwei Li, Tian Fang, and Long Quan. MVSNet: Depth inference for unstructured multi-view stereo. In *European Conference on Computer Vision (ECCV)*, pages 767–783, 2018. 1, 2

[43] Lior Yariv, Jiatao Gu, Yoni Kasten, and Yaron Lipman. Volume rendering of neural implicit surfaces. *Advances in Neural Information Processing Systems*, 34:4805–4815, 2021. 1

[44] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. pixelNeRF: Neural radiance fields from one or few images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2021. 2

[45] Zehao Yu and Shenghua Gao. 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*, pages 1949–1958, 2020. 2

[46] Zehao Yu, Songyou Peng, Michael Niemeyer, Torsten Sattler, and Andreas Geiger. MonoSDF: Exploring monocular geometric cues for neural implicit surface reconstruction. *Advances in neural information processing systems*, 35:25018–25032, 2022. 1, 3, 6, 7

[47] Jingyang Zhang, Yao Yao, Shiwei Li, Tian Fang, David McKinnon, Yanghai Tsin, and Long Quan. Critical regularizations for neural surface reconstruction in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6270–6279, 2022. 1

[48] Wenyuan Zhang, Yu-Shen Liu, and Zhizhong Han. Neural signed distance function inference through splatting 3d gaussians pulled on zero-level set. In *Advances in Neural Information Processing Systems*, 2024. 2

[49] Shaohong Zhong, Alessandro Albini, Oiwi Parker Jones, Perla Maiolino, and Ingmar Posner. Touching a nerf: Leveraging neural radiance fields for tactile sensory data generation. In *Conference on Robot Learning*, pages 1618–1628. PMLR, 2023. 2

[50] Xiaowei Zhou, Haoyu Guo, Sida Peng, Yuxi Xiao, Haotong Lin, Qianqian Wang, Guofeng Zhang, and Hujun Bao. Neural 3d scene reconstruction with indoor planar priors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 2, 3