Journal of *Imaging*

MDPI

Article

Three-Dimensional Reconstruction of Indoor Scenes Based on Implicit Neural Representation

Zhaoji Lin [1,*], Yutao Huang [2] and Li Yao [2]

[1] School of Computer Science and Engineering, Sanjiang University, Nanjing 210012, China
[2] School of Computer Science and Engineering, Southeast University, Nanjing 211189, China; 172210501215@stu.just.edu.cn (Y.H.); yao.li@seu.edu.cn (L.Y.)
* Correspondence: lin_zhaoji@sju.edu.cn; Tel.: +86-137-7669-3202

Abstract: Reconstructing 3D indoor scenes from 2D images has always been an important task in computer vision and graphics applications. For indoor scenes, traditional 3D reconstruction methods have problems such as missing surface details, poor reconstruction of large plane textures and uneven illumination areas, and many wrongly reconstructed floating debris noises in the reconstructed models. This paper proposes a 3D reconstruction method for indoor scenes that combines neural radiation field (NeRFs) and signed distance function (SDF) implicit expressions. The volume density of the NeRF is used to provide geometric information for the SDF field, and the learning of geometric shapes and surfaces is strengthened by adding an adaptive normal prior optimization learning process. It not only preserves the high-quality geometric information of the NeRF, but also uses the SDF to generate an explicit mesh with a smooth surface, significantly improving the reconstruction quality of large plane textures and uneven illumination areas in indoor scenes. At the same time, a new regularization term is designed to constrain the weight distribution, making it an ideal unimodal compact distribution, thereby alleviating the problem of uneven density distribution and achieving the effect of floating debris removal in the final model. Experiments show that the 3D reconstruction effect of this paper on ScanNet, Hypersim, and Replica datasets outperforms the state-of-the-art methods.

Keywords: 3D reconstruction; indoor scene; neural radiance fields; signed distance function; normal prior; mesh

Citation: Lin, Z.; Huang, Y.; Yao, L. Three-Dimensional Reconstruction of Indoor Scenes Based on Implicit Neural Representation. *J. Imaging* **2024**, *10*, 231. https://doi.org/10.3390/jimaging10090231

Academic Editor: Daniel Meneveaux

Received: 27 August 2024
Revised: 12 September 2024
Accepted: 14 September 2024
Published: 16 September 2024

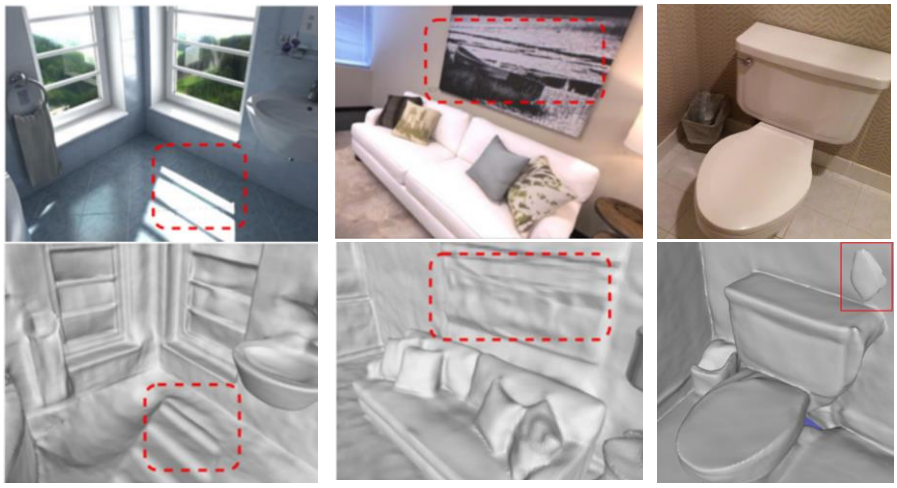
1. Introduction

The goal of 3D reconstruction of indoor scenes is to reconstruct and restore an accurate scene model from 2D images of indoor scenes from multiple angles to reflect the geometry, structure, and appearance characteristics of the actual scene [1]. This process can obtain an explicit 3D model observed from any perspective. This task has been a hot topic and an important task in computer vision and graphics research in recent years, and has broad application prospects in house interior restoration, interior design, virtual reality, augmented reality, indoor navigation, etc. [2].

Unlike object-level reconstruction, indoor environments usually include large and small objects, different materials, and complex spatial layouts, which puts higher demands on feature extraction and scene understanding. Indoor lighting conditions are also complex and changeable, and may include natural light, artificial light, and shadow areas. These factors will affect the quality of the reconstructed model and the difficulty of reconstruction. In addition, occlusion between objects is more common in indoor scenes, which makes it more difficult to obtain complete scene information from a limited perspective. Existing methods often have poor effects on uneven lighting (as shown in Figure 1a) and partial planar texture processing (as shown in Figure 1b). There are many floating debris noises in the air of the reconstructed indoor 3D model (as shown in Figure 1c). The recently

proposed neural radiance fields reconstruction method can achieve good results, but the computation is large, and it cannot directly obtain a 3D mesh model. We strengthen surface learning by adding normal priors and use an SDF, a compact and continuous multi-layer perceptron (MLP), to parameterize the representation of the implicit model, and finally obtain a high-quality 3D mesh model. In summary, this paper has the following main contributions:

(1) It proposes a new indoor 3D reconstruction method that combines NeRF and SDF scene expression, which not only preserves the high-quality geometric information of the NeRF, but also uses the SDF to generate an explicit mesh with a smooth surface.
(2) By adding adaptive normal priors to provide globally consistent geometric constraints, the reconstruction quality of planar texture areas and details is significantly improved.
(3) By introducing a new regularization term, the problem of uneven distribution of NeRF density is alleviated, and the effect of removing floating debris is achieved in the final generated model, which improves the look and feel of the visualization results.



(**a**) (**b**) (**c**)

Figure 1. (**a**) Distortion of reconstructed 3D models under uneven lighting conditions enclosed by the red dashed box; (**b**) distortion of 3D reconstruction of smooth planar texture areas enclosed by the red dashed box; (**c**) floating debris noise in red box in 3D reconstruction.

2. Related Works

2.1. Three-Dimensional Reconstruction of Based on Visual SLAM

Simultaneous Localization and Mapping (SLAM) refers to the process of a moving object carrying a sensor to locate itself during movement and to synchronously map the surrounding environment in an appropriate manner. In 2016, Google open-sourced an indoor SLAM library called Cartographer [3], which is still being updated. Its main application area is indoor reconstruction, providing functions such as positioning, mapping, and loop detection. LOAM [4] is a SLAM algorithm based on a 3D laser sensor. Compared with Cartographer, it can solve indoor and outdoor problems, but it does not have loop detection. Later, based on LOAM, researchers proposed LeGOLOAM [5], a new algorithm derived from the LOAM framework. By introducing a global map and loop detection module, the positioning accuracy and robustness are improved, and at the same time, the entire algorithm is made more lightweight. The advantage of the SLAM system is its fast reconstruction speed. With the improvement of the robustness and accuracy of its algorithm, it makes real-time 3D reconstruction possible. However, due to the need for sensor participation, the reconstruction cost is high, and the process is relatively complicated.

2.2. Three-Dimensional Reconstruction Based on TSDF

The main idea of the 3D reconstruction method based on regression to a truncated signed distance function (TSDF) is to create a voxel grid, initialize the maximum truncated signed distance value for each voxel, fuse the depth map into the voxel through the TSDF [6], calculate the distance from each point to the center of each voxel in the voxel grid, and update the TSDF value of the voxel if the distance is within the truncation range. KinectFusion [7] uses the depth camera Kinect to scan and model indoor spaces and objects in real time, and uses a TSDF to manage spatial information. This method improves the efficiency of data processing and the accuracy of the reconstruction process. Atlas [8] provides an end-to-end reconstruction pipeline, using a 2D CNN to extract features from each image independently, and then using the intrinsic and extrinsic features of the camera to back-project and accumulate into voxel volumes. After accumulation, a 3D CNN refines the accumulated features and predicts the TSDF value. At the same time, the semantic segmentation goal is added to the model to accurately mark the generated surface, which improves the model's ability to handle occlusion and large room scenes. In order to reduce the computational burden, unlike Atlas, which processes the entire image sequence at once, NeuralRecon [9] proposed a coarse-to-fine framework that uses a recursive network to fuse features from previous fragments and reconstructs the entire scene by processing the local surface of each fragment sequence, making progress on datasets with high occlusion and scene complexity. However, due to its design idea of local estimation, TSDF-based 3D reconstruction methods have difficulty obtaining global reconstruction with fine details.

2.3. Three-Dimensional Reconstruction Based on MVS

The traditional multi-view stereo (MVS) method first estimates the depth map of each image based on multiple views, and then performs depth fusion to obtain the final reconstruction result. These methods can reconstruct relatively accurate 3D shapes and have been used in many downstream applications, such as new view synthesis. Schoenberger proposed a new general image 3D reconstruction method, which uses scale-invariant feature transform (SIFT) features and the Fast Library for Approximate Nearest Neighbors (FLANN) matching algorithm to improve the accuracy and efficiency of Structure from Motion (SFM), and open-sourced the project as COLMAP 3.10 [10] software for enthusiasts to use, which can perform dense point cloud reconstruction and surface reconstruction. Based on this, they further proposed the Pixelwise View Selection [11] method, which improved the poor reconstruction effect and efficiency caused by the previous input image specifications (such as different input image sizes, different lighting, etc.). This method regards the view selection problem as a binary classification problem, selects the best view for each pixel, and thus can use the information of all perspectives for better stereo matching. However, these methods perform poorly in large planar texture areas or areas with sparse textures because their optimization is highly dependent on the photometric process. In indoor scenes with planar texture areas, the inherent uniformity makes photometry ineffective, making it difficult to accurately estimate depth [12].

With the development of deep learning, learning-based MVS methods [13–15] have shown good performance in recent years. Mvsnet [16] uses convolutional neural networks to predict depth maps, integrates image information from multiple perspectives into a unified 3D space, and implements depth estimation by constructing a 3D matching cost-volume, which significantly improves the accuracy of depth estimation. Fast-Mvsnet [17] improves on Mvsnet by adopting a more efficient network structure and optimizing the depth map prediction process. By simplifying the representation of cost volume and adopting a lighter network architecture, the processing speed is improved, and the memory consumption is reduced. DeepV2D [18] combines the temporal information of video frames with visual depth estimation and adopts a two-stage network architecture. It first performs motion estimation on the video sequence, and then combines motion information and visual features to predict the depth map of each frame, further improving the quality of depth prediction. DeepMVS [19] uses a deep neural network to preprocess the input

image, extract features, and predict the depth information of each pixel. It introduces a novel image warping and synthesis step to improve the consistency between views. UCS-Net [20] constructs the cost volume in a coarse-to-fine hierarchical manner to obtain higher resolution depth estimation. Although the MVS method based on deep learning has made great progress, it still faces some problems. Since the depth map is estimated separately for each view, there are often some geometric inconsistencies and scale ambiguities, resulting in holes and noise on the surface of the reconstructed result [21]. When encountering areas with relatively scarce textures, these models find it difficult to accurately predict depth information.

2.4. Three-Dimensional Reconstruction Based on Implicit Neural Networks

Coordinate-based implicit neural networks, which encode a field by regressing 3D coordinates to output values via an MLP, have become a popular approach for representing scenes due to their compactness and flexibility. The Dist [22] model proposed a method to learn 3D shapes from 2D images. IDR [23] models rely on view appearance and can be applied to non-Lambertian surface reconstruction. However, they require mask information to obtain reconstructions. NeRFs encode scene geometry via volume density and are suitable for the task of novel view synthesis for volume rendering. However, due to the lack of surface constraints, volume density cannot represent high-fidelity surfaces. Inspired by neural radiance fields, Neus [24] and VolSDF [25] attached volume rendering techniques to IDR and eliminated the need for mask information. Although these methods achieve stunning reconstruction results in small-scale, texture-rich scenes, they often perform poorly in large-scale indoor scenes with planar texture areas. Mip-NeRF360 [26] improves upon the original NeRF's problems of unbalanced details and proportions at near and far distances, as well as the limited nature of synthesized scenes, and can render unbounded scenes more realistically. To address the slow convergence of the original NeRF training process, NSVF [27] uses a sparse voxel octree to assist in spatial modeling, achieving significant improvements in training time. The traditional NeRF requires input from multiple views to estimate volume representations. If multi-view data are insufficient, the generated scene can easily collapse into a plane. To address this problem, Google researchers proposed LOLNeRF [28], which can train a NeRF model from a single viewpoint for the same type of object, without adversarial supervision, thereby enabling a single 2D image to generate a 3D model. All of the above are advances made by NeRFs in synthesizing new viewpoints. In recent years, researchers have gradually shifted their focus to using NeRFs for 3D reconstruction. Guy [29] et al. applied a NeRF to facial reconstruction, achieving the generation of high-quality 3D facial models from a single RGB image. By introducing the time dimension, this method can model and reconstruct the dynamic changes in facial shape and expression. However, all of the above NeRF studies are limited to object-level reconstruction. When the reconstructed objects are expanded to indoor scenes, the generated 3D models often contain a lot of noise and topological errors.

3. Methodology

This paper proposes a method for indoor scene 3D reconstruction that combines NeRF and SDF implicit expression. The volume density of the NeRF is used to provide geometric information for the SDF field, and the learning process is optimized by adding normal priors to strengthen the learning of geometric shapes and surfaces. The overall framework of the method proposed in this paper is shown in Figure 2. Multiple 2D images of indoor scenes are used as input, and the explicit 3D model mesh of the corresponding scene is output. This method mainly includes the following three modules:

1. Normal estimation module: This module uses a spatial rectifier-based method to generate the corresponding normal map for a single RGB image, and prepares data for the prior part of neural implicit reconstruction.
2. NeRF module: The appearance decomposition and feature processing of the scene image are performed through the neural radiant field, and the volume density and

color are obtained. The image under the corresponding perspective is obtained by volume rendering, and the MLP parameters are optimized inversely with the input image loss.

3. SDF field module: The purpose of this module is to learn a high-quality SDF from the network, and at the same time, strengthen the network's understanding of the geometric structure through the normal prior. The implicit-3D-expression SDF is converted into an explicit triangular mesh through the Marching Cubes algorithm.

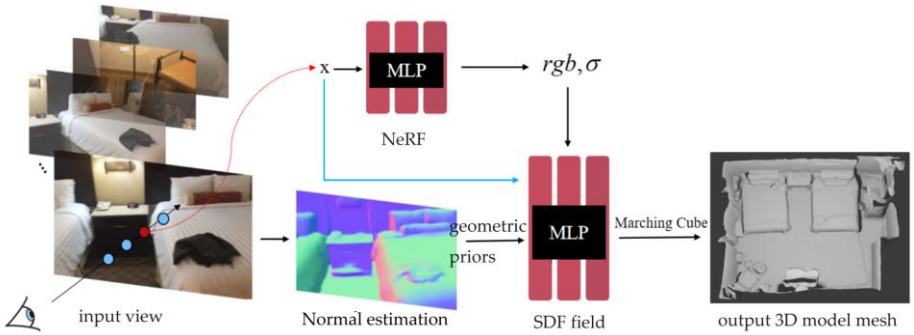


Figure 2. Overall framework of the method.

3.1. Optimization of Indoor 3D Reconstruction Based on Adaptive Normal Prior

A normal map is an important type of image, which represents the surface normal direction of each pixel in the image through the color of the pixel. In 3D graphics and computer vision, the normal is a vector perpendicular to the surface of an object, which can be used to describe the direction and shape of the surface. In the normal map, RGB color channels are generally used to represent the X, Y, and Z components of the normal vector, respectively. Normal information helps to correct errors in the reconstruction process and improve the reconstruction quality.

Currently, many monocular normal estimation methods have achieved high accuracy under a clear image input. However, considering that indoor scene images often have small blur and tilt, this paper selects TiltedSN [30] as the normal estimation module. Because the estimated normal map is usually over-smoothed, there are problems of inaccurate estimation on some fine structures, such as the chair legs in Figure 3a. Therefore, we adopt an adaptive method to use normal priors, using a mechanism based on the consistency of multiple views of the input image to evaluate the reliability of the normal prior. As shown in Figure 3b, this process is also called geometric checking. For areas that do not meet the consistency of multiple images, the normal prior is not applied. Instead, the appearance information is used for optimization to avoid the negative effects of incorrect normal maps that lead to misjudgment in reconstruction.

Given a reference image I_i, evaluate the consistency of the surface observed from pixel q. Define a local 3D plane $\{p|p^T n = dv^T n\}$ in the camera space associated with q, where v is the viewing direction, d is the distance to pixel q, and n is the normal estimate. Next, find a set of adjacent images, assuming that one of the adjacent images is I_j. The homography change from I_i to I_j can be calculated by the following formula:

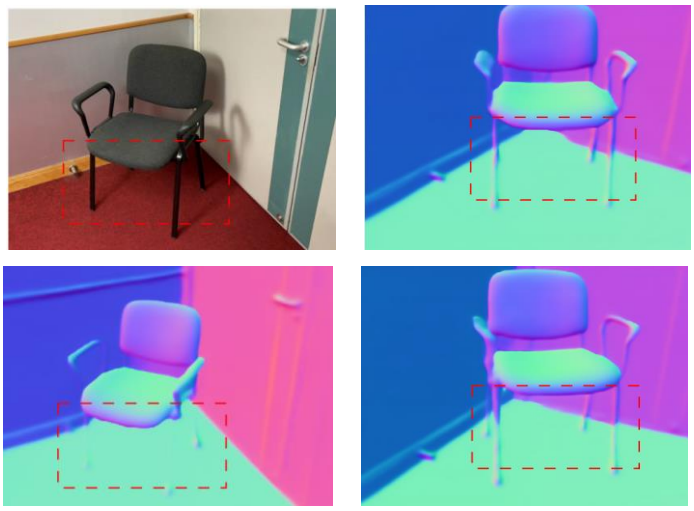
$$H_{n,d} = K_j \left(R_j R_i^{-1} - \frac{(t_i - t_j) n^T}{d v^T n} \right) K_i^{-1} \quad (1)$$

where $\{K_*, R_*, t_*\}$ is the intrinsic parameter matrix, the rotation and translation camera parameters. For pixel q in I_i, find a square block P centered on it as the neighborhood, and warp the block to its adjacent view I_j using the calculated homography matrix. The block matching method (patchmatch) can be used to find similar image blocks on adjacent views, and the normalized cross-correlation (NCC) method is used to evaluate the visual consistency of (n, d). NCC is a method for measuring the similarity between two images. It

evaluates the similarity between the two images by calculating the degree of correlation between them. Compared with simple cross-correlation, normalized cross-correlation is insensitive to changes in brightness and contrast, so it is more reliable in practical applications. The applied NCC formula is as follows:

$$NCC_j(P,n) = \frac{\sum_{q\in P} \hat{I}_i(q)\hat{I}_j(H_{n,d}(q))}{\sqrt{\sum_{q\in P} \hat{I}_i(q)^2 \sum_{q\in P} \hat{I}_j(H_{n,d}(q))^2}} \tag{2}$$

where $\hat{I}_*(q) = I_*(q) - \overline{I}_*(q)$, $I_i(q)$ and $I_j(q)$ represent two image regions to be compared, $\overline{I}_*(q)$ is the average value of $I_*(q)$, $\hat{I}_*(q)$ is the difference between them; the numerator calculates the sum of the products of the differences between the two image blocks, and the denominator calculates the square root of the product of the sum of the squares of the differences between the two image blocks. This process ensures the normalization of the results, making the NCC range within $(-1, 1)$. The closer the NCC value is to 1, the more similar the two image regions are; the closer the NCC value is to -1, the less similar they are; the closer the NCC value is to 0, the less obvious linear relationship there is between them.



(**a**) (**b**)

Figure 3. (**a**) The normal estimation is inaccurate in some fine structures enclosed by red dashed box, such as chair legs, based on TiltedSN normal estimation module; (**b**) we use an adaptive normal prior method to derive accurate normals based on the consistency of adjacent images. In the red dashed box, the fine structures are accurately reconstructed.

If the reconstructed geometry is not accurate at the sampling pixel, it cannot meet the multi-view photometric consistency, which means that its related normal prior cannot provide help for the overall reconstruction. Therefore, a threshold ϵ is set, and by comparing the NCC at the sampling block with ϵ, the following indicator function can be used to adaptively determine the training weight of the normal prior.

$$\Omega_q(\hat{n}) = \begin{cases} 1 & if \sum_j NCC_j(P,\hat{n}) \geq \epsilon \\ 0 & if \sum_j NCC_j(P,\hat{n}) < \epsilon \end{cases} \tag{3}$$

The normal prior is used for supervision only when $\Omega_q(\hat{n}) = 1$. If $\Omega_q(\hat{n}) = 0$, the normal prior of the region is considered unreliable and will not be used in subsequent optimization processes.

3.2. Neural Implicit Reconstruction

Since the NeRF cannot express the surface of the object well, we are looking for a new implicit expression. The SDF can represent the surface information of objects and scenes and achieve better surface reconstruction. Therefore, we choose the SDF as the implicit scene representation.

3.2.1. Scene Representation

We represent the scene geometry as an SDF. An SDF is a continuous function f that, for a given 3D point, returns the distance from that point to the nearest surface:

$$f : \mathbb{R}^3 \to \mathbb{R} \qquad x \mapsto s = f(x) \tag{4}$$

Here, x represents a 3D point and s is the corresponding SDF value, thus completing the mapping from a 3D point to a signed distance. We define the surface S as the zero-level set of the SDF, expressed as follows:

$$S = \{x | f(x) = 0\} \tag{5}$$

By using the Marching Cubes algorithm on the zero horizontal plane of the SDF, that is, the surface of the object or scene, we can obtain a 3D mesh with a relatively smooth surface.

Using the SDF to get the mesh has the following advantages:

(1) Clear surface definition: The SDF provides the distance to the nearest surface for each spatial point, where the surface is defined as the location where the SDF value is zero. This representation is well suited for extracting clear and precise surfaces, making the conversion from SDF to mesh relatively direct and efficient.
(2) Geometric accuracy: The SDF can accurately represent sharp edges and complex topological structures, which can be maintained when converted to meshes, thereby generating high-quality 3D models.
(3) Optimization-friendly: The information provided by the SDF can be directly used to perform geometry optimization and mesh smoothing operations, which helps to further improve the quality of the model when generating the mesh.

3.2.2. Implicit Indoor 3D Reconstruction Based on Normal Prior

The reconstruction process based on the normal prior is shown in Figure 4. The input mainly consists of three parts.

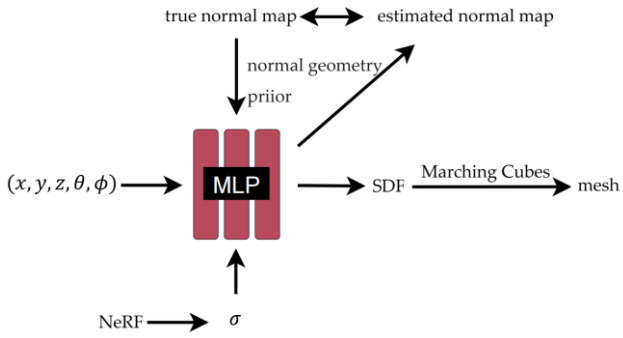


Figure 4. Neural implicit reconstruction process.

The first part is a five-dimensional vector representing the information of the sampling point (x, y, z, θ, ϕ). The second part is the volume density we obtained previously through the NeRF σ. Given that we use the SDF as a surface expression, the volume density obtained by the NeRF can represent more comprehensive scene geometry information. Because the scene contains multiple objects and air, it is difficult to represent a complex scene only through surface information. Traditional SDF-based methods are often limited to object reconstruction. The constraints of NeRF volume density combined with SDF reconstruction

can reconstruct richer scene geometry. The third part is the normal geometry prior. The normal map here is obtained by the monocular normal estimation method mentioned earlier. The normal map provides the orientation information of the object surface, which is conducive to enhancing detail reconstruction.

The network used here is an improved NeRF, which also contains 2 MLPs. Like the NeRF, there is a color network, f_c, and the other grid has become an SDF network, f_{θ_g}, which can get the SDF value of the point through the 3D coordinates of the point.

Here, we will use volume rendering technology to get the predicted image, and optimize it with the input real image through the loss function. Specifically, for each pixel, we sample a set of points along the corresponding emission light, denoted as $p_i = o + d_i v$, where p_i is the sampling point, o is the camera center, and v is the direction of the light. The color value can be accumulated by the following volume rendering formula.

$$\hat{c} = \sum_{i=1}^{n} T_i \alpha_i c(p_i, v) \tag{6}$$

where $T_i = \prod_{j=1}^{i-1}(1 - \alpha_j)$ is the cumulative transmittance, i.e., the probability that with no object occlusion, c is the color value, $\alpha_i = 1 - \exp\left(-\int_{t_i}^{t_{i+1}} \rho(t)dt\right)$ is the discrete opacity, and $\rho(t)$ is the opacity corresponding to the volume density σ in the original NeRF. Since the rendering process is fully differentiable, we learn the weights of f_c and f_{θ_g} by minimizing the difference between the rendering result and the reference image.

In addition to generating the appearance, the above volume rendering scheme can also obtain the normal vector. We can approximate the surface normal observed from a viewpoint by volume accumulation along this ray:

$$\hat{n} = \sum_{i=1}^{n} T_i \alpha_i n_i \tag{7}$$

where $n_i = \nabla f(p_i)$ is the gradient at point p_i.

At this time, we compare the true normal map obtained by the monocular method with the estimated normal map obtained by the volume rendering process, and further optimize the parameters of the MLP by calculating the loss to obtain a more accurate normal map and geometric structure.

3.3. Floating Debris Removal

The NeRF initially acts on the generation of new perspectives on objects, maintaining a high degree of clarity and realism. This is mainly due to volume rendering. However, when the NeRF is used for 3D reconstruction, some floating debris in the air often appears. This floating debris refers to small disconnected areas in the volume space and translucent substances floating in the air. In view synthesis work, this floating debris is often not easy to detect. However, if an explicit 3D model needs to be generated, this floating debris will seriously affect the quality and appearance of the 3D model. Therefore, it is very necessary to remove the floating debris in these incorrectly reconstructed places.

This floating debris often does not appear in object reconstruction. However, in scene-level reconstruction, due to the significant increase in environmental complexity and the lack of relevant constraints on the NeRF, there is a phenomenon of inaccurate local area density prediction. Therefore, this paper proposes a new regularization term to constrain the weight distribution of the NeRF.

First, simulate the sampling process of the NeRF. In a scene, assume that there are only rigid objects, excluding the existence of translucent objects. After a ray is shot out, there will be countless sampling points on this ray. The weight value of the sampling point before the ray encounters the object should be extremely low (close to 0). When the ray contacts the rigid object, the weight value here should soar, much higher than other values. After the object, the weight value returns to a lower range. This is the desired weight distribution in an ideal state, which is a relatively compact unimodal distribution. This method defines a

regularization term, which is a step function defined by a set of standardized ray distances s and the weight w after parameterizing each ray:

$$L_{dist}(s,w) = \iint_{-\infty}^{\infty} w_s(u) w_s(v) |u-v| d_u d_v \tag{8}$$

Here, u and v refer to points on the sampling ray, that is, points on the x-axis in the weight distribution diagram, $|u-v|$ is the distance between the two points, and $w_s(u)$ and $w_s(v)$ are the weight values at point u and point v, respectively. Since all particle combinations from negative infinity to positive infinity need to be exhausted, integration is performed in the front. If you want to make the loss function as small as possible, there are mainly two situations:

(1) If the distance between point u and point v is relatively far, that is, the value of $|u-v|$ is large, if you want to ensure that the value of $L_{dist}(s,w)$ is as small as possible, then either $w_s(u)$ or $w_s(v)$ needs to be small and close to zero. That is, as shown in Figure 5, (A, B), (B, D), (A, D), etc. all satisfy that $|u-v|$ is large and the weight value of at least one point is extremely small (close to zero);

(2) If the values of $w_s(u)$ and $w_s(v)$ are both large, if you want to ensure that the value of $L_{dist}(s,w)$ is as small as possible, then the value of $|u-v|$ needs to be small, that is, the distance between points u and v is very close. As shown in Figure 5, only the combination of (B, C) satisfies the condition that the values of $w_s(u)$ and $w_s(v)$ are both large, and at this time, points u and v just meet the condition that the distance is very close.

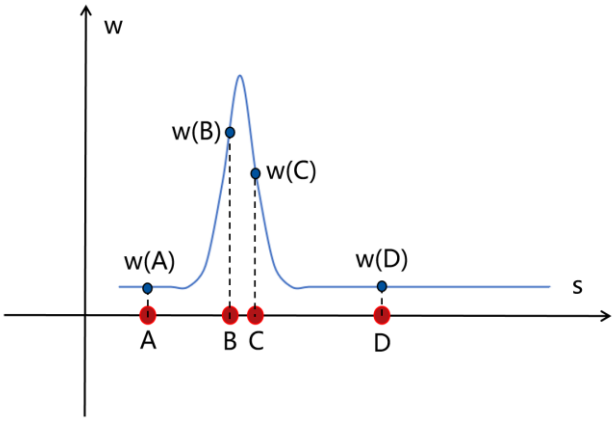


Figure 5. Distribution diagram of distance and weight values between sampling points.

Therefore, through the analysis of the above two cases, it can be found that the properties of the regularization term can constrain the density distribution to the ideal single-peak distribution with the good central tendency proposed before. The purpose of this regularization term is to minimize the sum of the normalized weighted absolute values of all samples along the ray, and encourage each ray to be as compact as possible, which is specifically reflected in the following steps:

1. Minimize the width of each interval;
2. Bring the intervals that are far apart closer to each other;
3. Make the weight distribution more concentrated.

The regularization term above cannot be used directly for calculation because it is in integral form. In order to facilitate calculation and use, it is discretized as the following:

$$L_{dist}(s,w) = \sum_{i,j} w_i w_j \left| \frac{s_i + s_{i+1}}{2} - \frac{s_j + s_{j+1}}{2} \right| + \frac{1}{3} \sum_i w_i^2 (s_{i+1} - s_i)$$

This discretized form also provides a more intuitive understanding of the behavior represented by this regularization, where the first term minimizes the weighted distance

between points in all intervals, and the second term minimizes the weighted size of each individual interval.

3.4. Training and Loss Function

During the training phase, we sample a batch of pixels and adaptively minimize the difference between the color and normal estimates and the true normal map. We sample m pixels $\{q_k\}$ and their corresponding reference colors $\{I(q_k)\}$ and normals $\{N(q_k)\}$ in each iteration. For each pixel, we sample n points in the world coordinate system along the corresponding ray, and the total loss is defined as

$$L = \lambda_c L_c + \lambda_n L_n + \lambda_e L_{eik} + \lambda_d L_{dist} \tag{9}$$

Among them, λ_c, λ_n, λ_e, and λ_d are the hyperparameters of color loss, normal loss, Eikonal loss, and distortion loss, respectively.

Color loss, L_c, is used to measure the difference in color between the reconstructed image and the real image:

$$L_c = \frac{1}{m} \sum_k \parallel I(q_k) - \hat{c}(q_k) \parallel \tag{10}$$

In the training phase, a batch of pixels need to be sampled. Each iteration samples m pixels $\{q_k\}$ and the corresponding reference color $\{I(q_k)\}$, where $\hat{c}(q_k)$ represents the pixel color predicted by volume rendering.

The normal prior loss L_n is to render the reconstructed 3D mesh of the indoor scene as a normal map, and compare it with the real normal map generated by the monocular method to obtain a loss:

$$L_n = \sum_k \parallel N(q_k) - \hat{n}(q_k) \parallel_1 \cdot \Omega_{q_k}(\hat{n}(q_k)) \tag{11}$$

where $N(q_k)$ is the true normal information, and $\hat{n}(q_k)$ is the normal information predicted by the gradient. $\Omega_{q_k}(\hat{n}(q_k))$ is an indicator function used to judge the accuracy of the normal prior. Here, some data with inaccurate normal estimation are eliminated. The normal loss is mainly calculated by cosine similarity, because the direction of the normal vector is more important than its length. Cosine similarity is a good measure of the similarity of the two vectors in direction.

The Eikonal loss L_{eik} [31] of the regularized SDF is

$$L_{eik} = \sum_{x \in \mathrm{X}} \Big(\parallel \nabla f_\theta(x) \parallel_2 -1 \Big)^2 \tag{12}$$

where X is a set of sampling points in the 3D space and the area near the surface, and x represents one of the sampling points. The reason why the gradient $\nabla f_\theta(x)$ needs to be close to 1 is that the ideal SDF represents the shortest distance from the current point to the surface, so the gradient direction is the direction in which the distance field change is the steepest. Assuming that x moves toward the surface along this direction by Δd, the SDF should also change by Δd. Therefore, in the ideal state, $\nabla f_\theta(x) = \partial D(x) = \Delta D(x)/\Delta x = \Delta d/\Delta d = 1$, where $D(x)$ is the original Eikonal equation. By introducing the Eikonal loss, the properties of the SDF can be well constrained, thereby ensuring the smoothness and continuity of the reconstructed surface.

4. Experimentation

4.1. Dataset

This paper conducts experimental analysis on the ScanNet [31], Hypersim [32], and Replica [33] datasets, as shown in Table 1. The ScanNet dataset is the main dataset used for indoor scene 3D reconstruction tasks. Our method selects 10 scenes from the ScanNet dataset. For each scene, a set of equally spaced images (about 150–600 images) are sampled from the corresponding video and the images are adjusted to a resolution of 640 × 480. In addition, a scene is selected from the Hypersim dataset and the Replica dataset to test the

generalization of this method in large-scale scenes. The results verify that this method has good reconstruction effects on other datasets in addition to the large public dataset ScanNet.

Table 1. Selected datasets to test our method.

Dataset	Scene Number	Scenes Selected in This Paper
ScanNet	1500+	10
Hypersim	461	10
Replica	18	10

4.2. Comparative Experiment

Five evaluation indicators are used: accuracy (Acc), completeness (Comp), precision (Prec), recall (Recall), and F1 score (F-score).

Accuracy is an indicator to measure the degree of consistency between the reconstructed mesh and the real scene mesh:

$$Acc = mean_{p \in P} \left(\min_{p^* \in P^*} \parallel p - p^* \parallel \right) \tag{13}$$

where P represents the set of points in the reconstructed grid, P^* is the set of points in the grid of the real scene, p is a point in the set P, and p^* is a point in the set P^*. $\parallel p - p^* \parallel$ represents the Euclidean distance between P and P^*. For each point p in P, find the point p^* in P^* that is closest to it, calculate the distance between them, and average all the distances.

Completeness is used to measure the extent to which the reconstructed model covers the original model or scene:

$$Comp = mean_{p^* \in P^*} \left(\min_{p \in P} \parallel p - p^* \parallel \right) \tag{14}$$

This metric measures completeness by calculating the average distance from each point in the true scene mesh P^* to the nearest point in the reconstructed mesh P.

Precision is a measure of the proportion of the correctly reconstructed part of the reconstructed model to the entire model. The calculation formula for precision is

$$Prec = mean_{p \in P} \left(\min_{p^* \in P^*} \parallel p - p^* \parallel < 0.05 \right) \tag{15}$$

For each point p in the reconstructed mesh P, find the point p^* in the GT mesh P^* that is closest to it, calculate the distance between them, and compare it with the set threshold 0.05 to calculate the proportion less than 0.05.

Recall measures the proportion of points in the GT model that are covered and correctly reconstructed by the reconstruction model:

$$Recall = mean_{p^* \in P^*} \left(\min_{p \in P} \parallel p - p^* \parallel < 0.05 \right) \tag{16}$$

For each point p^* in the reconstructed mesh P^*, find the point p in the GT mesh P that is closest to it, calculate the distance between them, and compare it with the previously set threshold of 0.05 to calculate the proportion less than 0.05.

The F1 score (F-score) is the harmonic mean of precision and recall, which can better measure some unbalanced datasets and provide a more comprehensive evaluation of the overall model:

$$F - score = \frac{2 \times Prec \times Recall}{Prec + Recall} \tag{17}$$

The architecture of the MLP encoding the SDF in our method consists of eight hidden layers of 256 channels. The training process and related parameters of the model are set

as follows: the number of iterations is set to 160,000, the learning rate is set to 4×10^{-4}, the number of rays sampled in each batch of training (batchsize) is set to 1024, and each ray contains 64 coarse sampling points and 64 fine sampling points. In the patchmatching process, the patch size is set to 11×11, the step size is set to 2 for the block matching process, and the NCC threshold ϵ is 0.66. The weights of each loss function, λ_c, λ_n, λ_e and λ_d, are set to 1.0, 1.0, 0.1, and 0.5, respectively.

As shown in Table 2, the method in this paper is significantly better than the state-of-the art methods, and performs well in most indicators, far exceeding the traditional MVS method and the TSDF-based reconstruction method.

Table 2. Quantitative comparison with other methods.

Method	Acc ↓	Comp ↓	Prec ↑	Recall ↑	F-Score ↑
COLMAP [10]	0.047	0.235	0.711	0.441	0.537
Atlas [8]	0.211	0.070	0.500	0.659	0.564
NeuralRacon [11]	0.056	0.081	0.545	0.604	0.572
Neus [24]	0.179	0.208	0.313	0.275	0.291
VolSDF [25]	0.414	0.120	0.321	0.394	0.346
NeRF [34]	0.735	0.177	0.131	0.290	0.176
Manhattan-SDF [35]	0.072	0.068	0.621	0.586	0.602
MonoSDF [36]	**0.035**	**0.048**	<u>0.799</u>	<u>0.681</u>	<u>0.733</u>
I2-SDF [37]	0.066	0.070	0.605	0.575	0.590
Ours	<u>0.037</u>	**0.048**	**0.801**	**0.702**	**0.748**

Bold text indicates the best results and the underlined text indicates the second best results. ↓ indicates that the smaller the value of this indicator, the better; ↑ indicates that the larger the value of this indicator, the better.

Among the neural implicit reconstruction methods, Neus and VolSDF mainly focus on object-level reconstruction, so they perform poorly on the ScanNet dataset; the first-generation NeRF is mainly used for new perspective synthesis, and no algorithm for extracting grids is proposed. Therefore, direct use of Marching Cubes to extract grids will result in large-scale collapse. This article only draws on many modules of the NeRF, so it is compared here; the I2-SDF method has good results in the synthetic dataset proposed in this article, but it does not have generalization ability and does not work well on the ScanNet dataset with more noise and motion blur.

MonoSDF and Manhattan-SDF are second only to this method in terms of quantitative results. Manhattan-SDF is based on the Manhattan World hypothesis, which assumes that the world is mainly composed of planes aligned with the coordinate axes. It is easily affected by the complexity of indoor scenes. Strict reliance on planes aligned with the coordinate axes for reconstruction may lead to missing or inaccurate details. Therefore, the overall accuracy is lower than this method. MonoSDF has achieved good performance in various indicators, especially accuracy. This is because MonoSDF weakens the reconstruction of complex structures that are difficult to handle during the reconstruction process. Therefore, the reconstructed part is highly overlapped with the scene itself, but some parts of the reconstruction will be lost. Therefore, compared with other evaluation indicators, there is some gap between its recall rate and this method, because the recall rate measures the ratio of the real model that is reconstructed.

In addition to quantitative analysis, this chapter visualizes the reconstruction results and makes qualitative comparisons with the most advanced MonoSDF and Manhattan-SDF. As shown in Figure 6a, compared with GT, this method fills some holes that were not scanned at the time. Compared with Manhattan-SDF, this method has higher accuracy and a smoother reconstruction effect on the surface of objects. Compared with MonoSDF, this method has more accurate reconstruction in many structures (shown by the red dashed line). The specific details in the red dashed box in Figure 6a are shown in Figure 6b. Manhattan-SDF has low reconstruction accuracy. After zooming in on the detail image, larger triangular facets can be seen. From the visualization point of view, the details are far inferior to those of this method. MonoSDF has better overall detail reconstruction—especially, the wall area

is relatively smooth—but there are many missing objects. In the right picture of Figure 6b, MonoSDF did not correctly reconstruct the lamp on the table, and only reconstructed a small part of the outline of the objects placed on the bookshelf, while this method restored these details of the real scene well. In the left picture of Figure 6b, MonoSDF and Manhattan-SDF are largely missing chair legs and armrests, while the method in this paper restores the chair structure in the real scene to a large extent.

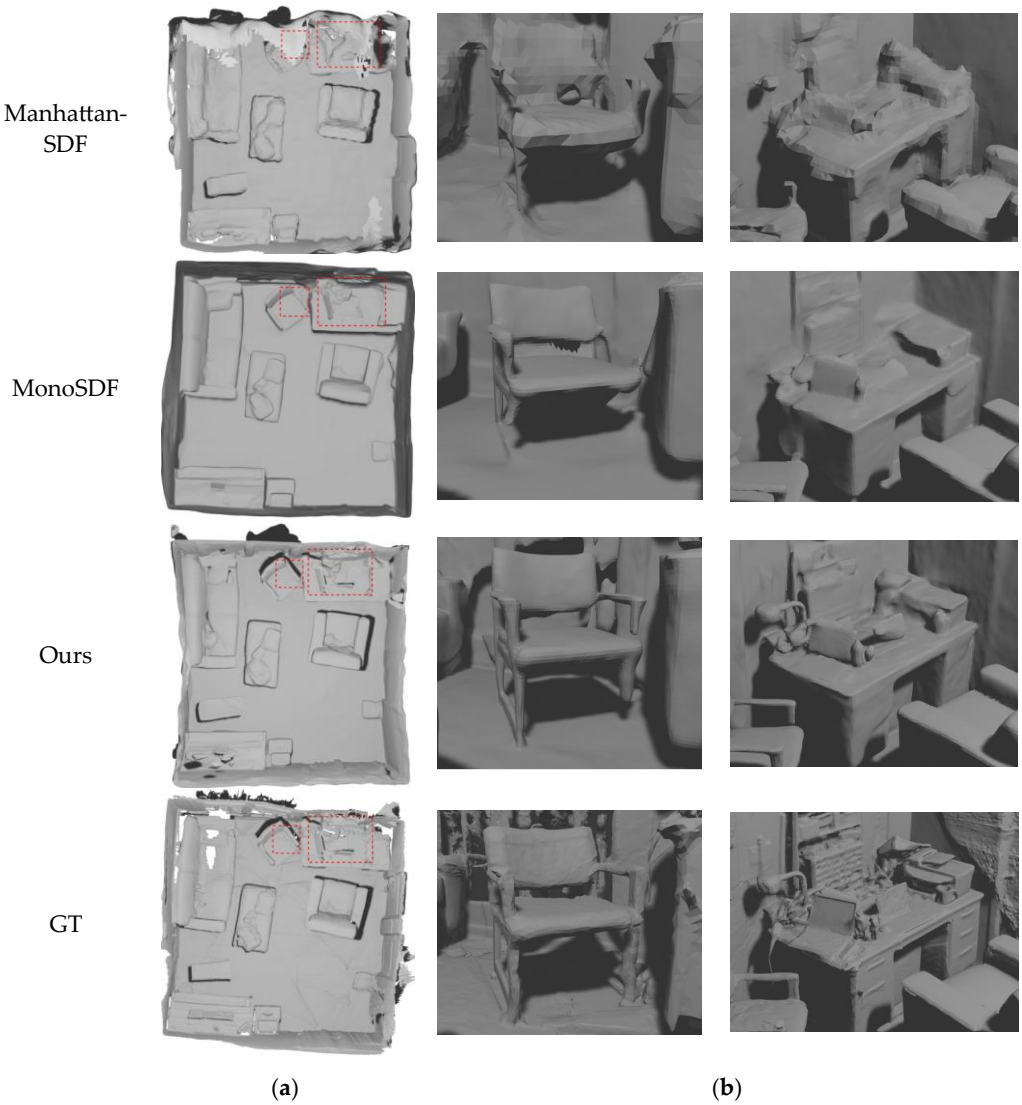


Figure 6. Three-dimensional model reconstructed from scenes in the ScanNet dataset. (**a**) Comparison of 3D models; (**b**) comparison of the specific details in the red dashed box.

4.3. Ablation Experiment

4.3.1. Normal Geometry Prior Ablation Experiment

This section will demonstrate the effectiveness of the adaptive normal prior scheme added in this paper through quantitative and qualitative ablation experimental analysis. There are three settings in this ablation experiment: (1) no normal prior is added, denoted as w/o N; (2) a normal prior is added, but the adaptive scheme is not used, denoted as w/N, w/o A; (3) the adaptive scheme is used with the normal prior, denoted as Ours. All settings are tested on the ScanNet dataset, Hypersim dataset, and Replica dataset. The evaluation indicators of each setting are shown in Table 3. It can be seen that the reconstruction quality can be significantly improved by adding the geometric prior, because it provides additional geometric constraints and alleviates many ambiguity problems caused by the lack of

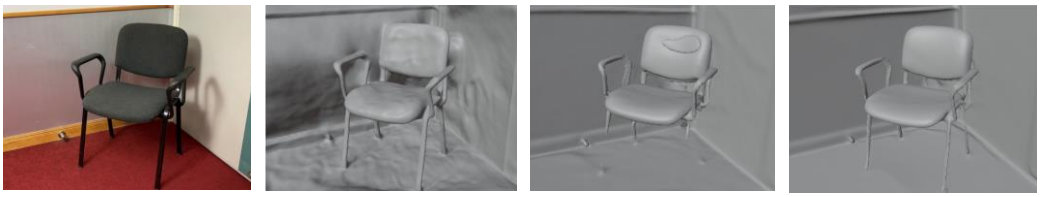
texture information in the original image. On this basis, the adaptive scheme can remove the incorrectly estimated normal map and further improve the reconstruction quality.

Table 3. Normal geometry prior ablation experiment.

Method	Acc ↓	Comp ↓	Prec ↑	Recall ↑	F-Score ↑
w/o N	0.183	0.152	0.286	0.290	0.284
w/N, w/o A	0.050	0.053	0.759	0.699	0.727
Ours	**0.037**	**0.048**	**0.805**	**0.709**	**0.753**

Bold text indicates the best results. ↓ indicates that the smaller the value of this indicator, the better; ↑ indicates that the larger the value of this indicator, the better.

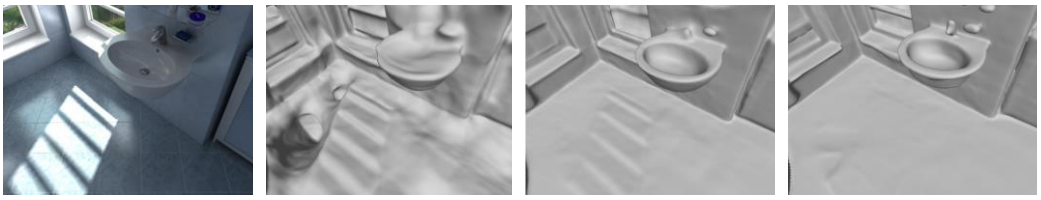
The following will verify the effectiveness of the adaptive normal prior module used in this section through qualitative visualization results. This module provides a certain improvement in the reconstruction of all indoor scenes, but the improvement in thin structure areas and reflective areas is particularly obvious, so these two areas are selected as visualization displays. As shown in Figure 7, a qualitative analysis of the chair leg structure is performed. Figure 7a is the input RGB image, i.e., the reference image, and Figure 7b is the model reconstructed without using the normal prior. The overall reconstruction accuracy is not high, which is reflected in the fact that the reconstruction effect of the surrounding floor and walls is not flat enough, and the surface of the chair is not smooth enough. Figure 7c adds the normal prior, but does not use the adaptive scheme. Although the reconstruction effect of the wall and chair surface is improved, the chair leg part is not reconstructed due to the wrong normal estimation. Figure 7d is the method used in this paper, that is, the adaptive scheme uses the normal prior, which not only greatly improves the accuracy, but also completely reconstructs the thin structure areas such as the chair leg.



(**a**) (**b**) (**c**) (**d**)

Figure 7. Qualitive comparison for thin structure areas using ScanNet dataset: (**a**) reference image; (**b**) model reconstructed without using normal prior; (**c**) model reconstructed with normal prior and without adaptive scheme; (**d**) model reconstructed with normal prior and adaptive scheme.

In addition to having a good visual improvement effect in thin structure areas, the adaptive normal prior scheme is also effective in some areas due to lighting or specular reflection. The image shown in Figure 8a is a reference image. The sunlight projected from the window forms regular white light and shadows on the ground. These light and shadows usually affect the reconstruction of the flatness of the entire ground because the network will understand them as independent geometric structures. Figure 8b is a model reconstructed without using the normal prior. A more obvious step structure is reconstructed in the light and shadow area, which is not the result of correct reconstruction. The reconstruction granularity in other areas is also obviously insufficient. Figure 8c adds the normal prior but does not use the adaptive scheme. Since most of the normal priors have weakened the erroneous impact of this area on the reconstruction, the erroneous outline of this area is already vague and not as obvious as in Figure 8b. However, there are also a few normal maps closer to the light and shadow area that estimate this area as a geometric structure different from the floor. Therefore, after adding the adaptive scheme, these few normal estimates can be removed, and the reconstruction effect is as smooth as the ground area, as shown in Figure 8d.



(a) (b) (c) (d)

Figure 8. Qualitive comparison for reflective areas using Hypersim dataset: (**a**) reference image; (**b**) model reconstructed without using normal prior; (**c**) model reconstructed with normal prior and without adaptive scheme; (**d**) model reconstructed with normal prior and adaptive scheme.

Through the above quantitative and qualitative results, it is not difficult to find that the normal prior can significantly improve the reconstruction effect and geometric details in the indoor scene reconstruction task. The use of the adaptive normal prior can reduce the erroneous reconstruction of many geometric structures and regions, and has a good correction effect on thin structures and partially reflective areas, making the reconstruction of these parts more robust and reasonable and close to the real scene, further proving the effectiveness and usability of the adaptive normal prior module in this section.

4.3.2. Ablation Experiment of Distortion Loss Function

The effectiveness of the new regularization term, the distortion loss function, is proposed in this paper through quantitative and qualitative ablation experimental analysis. Since the distortion loss function L_{dist} is only applicable to scenes with floating debris, the ablation experiment is also conducted on these scene datasets. There are two different settings in this ablation experiment: (1) without adding the distortion loss function, denoted as w/o L_{dist}; (2) after adding the distortion loss function, denoted as Ours. By ablating it in five scenes with floating debris noise in ScanNet, the quantitative indicators shown in Table 4 can be obtained.

Table 4. Distortion loss function ablation experiment.

Method	Acc ↓	Comp ↓	Prec ↑	Recall ↑	F-Score ↑
w/o L_{dist}	0.055	**0.052**	0.742	0.701	0.721
Ours	**0.047**	**0.052**	**0.795**	**0.704**	**0.746**

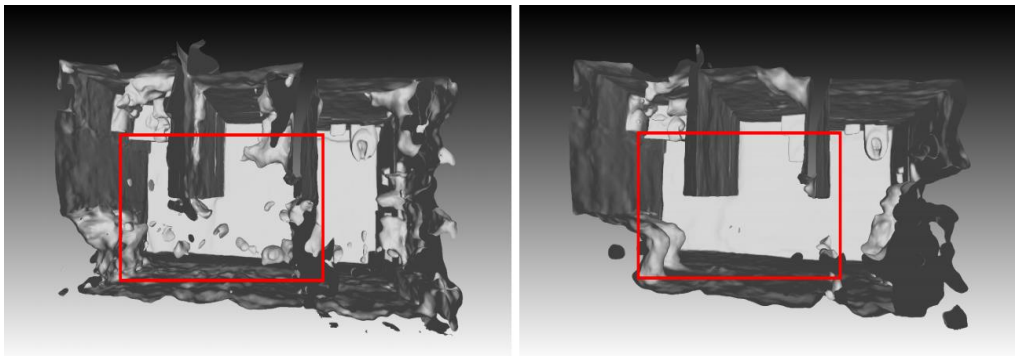
Bold text indicates the best results. ↓ indicates that the smaller the value of this indicator, the better; ↑ indicates that the larger the value of this indicator, the better.

As can be seen from Table 4, Acc and Prec have been significantly improved, which means that the distortion loss function removes some erroneous reconstruction parts and achieves more accurate reconstruction, while Comp and Recall have not changed much, because the method in this paper does not produce too many additional areas for supplementary reconstruction.

Figure 9 shows a visual comparison of a scene with a large amount of floating debris. There are is a large amount of incorrectly reconstructed floating debris in Figure 9a. After adding the distortion loss function, most of this floating debris in Figure 9b is eliminated.

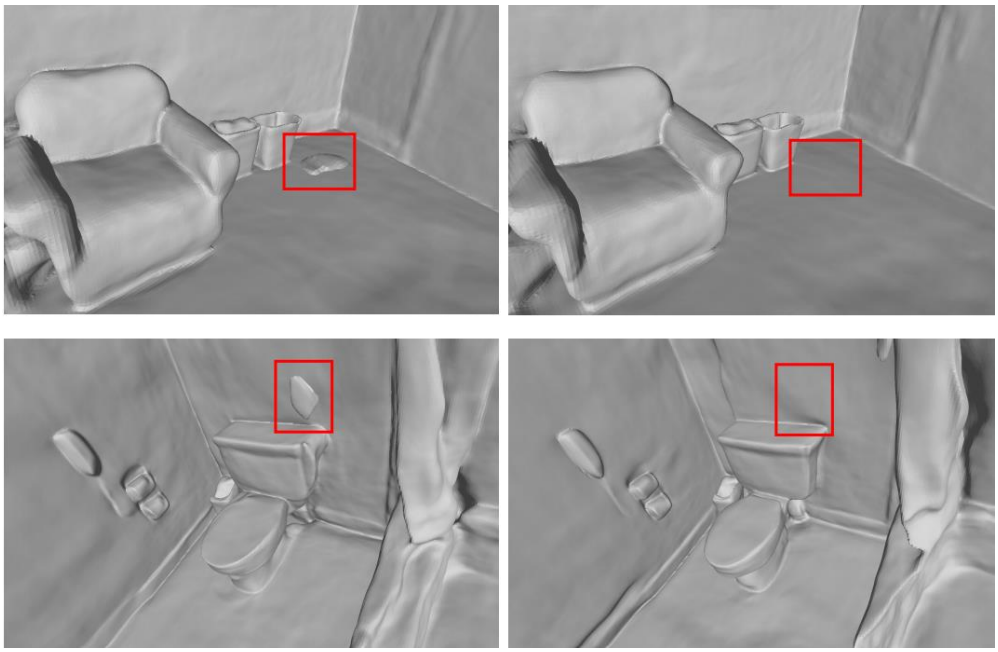
In addition, it also has a good removal effect on single floating debris areas in some scenes. As shown in Figure 10, there are single floating debris areas in both scenes in Figure 10a, which are successfully removed in Figure 10b.

The ablation experiment results on the distortion loss function show that, both qualitatively and quantitatively, the distortion loss function proposed in Chapter 3 of this paper can effectively remove floating object noise in the 3D model and improve the overall quality of the 3D model.



(**a**) W/o L_{dist} (**b**) Ours

Figure 9. Visual comparison for a scene with a large amount of floating debris using the ScanNet dataset; (**a**) reconstruction result without adding a distortion loss function; (**b**) reconstruction result with a distortion loss function.



(**a**) W/o L_{dist} (**b**) Ours

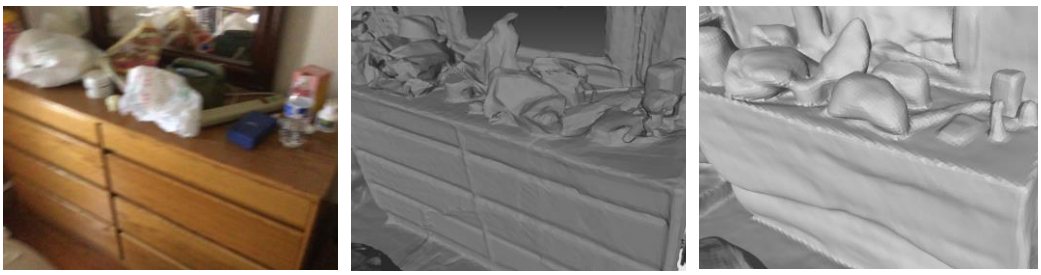
Figure 10. Visual comparison for a scene with single floating debris areas enclosed by red dashed box using the ScanNet dataset; (**a**) reconstruction result without adding distortion loss function; (**b**) reconstruction result with distortion loss function.

4.4. Limitations

The comparative experiment part illustrates that the proposed method has advantages over the benchmark model in both quantitative indicators and visualization results, which fully proves the superiority of the proposed method. The ablation experiment part proves the effectiveness of each module of the proposed method. Nevertheless, the proposed method still has some limitations in some specific scenarios.

For scenes with messy objects, the reconstruction effect of this method is not perfect. This is because in the 3D reconstruction task of indoor scenes, the arrangement and combination of objects will greatly affect the reconstruction process. As shown in Figure 11, there are many irregularly shaped objects on the table in the input image, and they are placed in overlapping and mutually occluding situations. For soft and transparent objects, such as plastic bags, it is difficult to accurately estimate their surface details because their appearance features are not easy to capture. Although this method can reconstruct its general outline, it is difficult to judge that this is a plastic bag based on the outline. The

reconstruction of such soft and non-solid objects in the scene has always been a difficult problem, and even the GT model obtained by scanning has not been able to restore this part well.



Input image GT Ours

Figure 11. The limitations of this method in the 3D reconstruction of scenes with clutter, occlusion, soft non-solid objects, and blurred images, using the ScanNet dataset.

In addition, the fuzziness of the input image is also crucial to the reconstruction result. If the input image has a certain amount of camera blur, as shown in Figures 5–8, many details of the object will be lost in the image, and the loss of these details increases the difficulty of reconstruction. This blur can also lead to inaccurate feature point detection, affecting the feature extraction and feature matching process. This blur can also affect the reconstruction of some areas with insufficient texture features, such as the dividing line and gap between the drawers in Figure 11. Due to the image blur, the area is incorrectly reconstructed.

5. Conclusion and Future Work

Based on the implicit expression of the NeRF and SDF, this paper proposes an indoor scene reconstruction method based on adaptive normal priors, and optimizes the geometric learning process through adaptive normal priors. The method proposed in this paper can significantly improve the reconstruction quality of large plane textures and uneven lighting areas in indoor scenes, and can also remove floating debris in the reconstructed 3D model. However, there are still some problems that need to be further optimized and improved in the future:

1. When there are many objects and elements in the scene and they are irregular, this method can only reconstruct the general outline, and the object category cannot be directly determined by these outlines. At the same time, the reconstructed results at the connections between objects and between objects and backgrounds are discontinuous. One solution is to try to introduce more priors to allow the neural network to obtain more useful information to accurately learn and understand the elements in the scene. Another feasible solution is to find a way to distinguish between object areas and non-object areas, and then learn them separately, which is conducive to further capturing more complex details.
2. This method takes several hours to more than ten hours to train and optimize a single indoor scene, which limits the application of this method in reconstruction over a relatively large range and in real-time reconstruction. One possible solution to improve training efficiency is to use hashed multi-resolution encoding to use a smaller network without sacrificing quality, thereby significantly reducing the number of floating-point and memory access operations, allowing the neural network to be trained at a smaller computational cost while maintaining reconstruction quality, greatly reducing training time.

Author Contributions: Conceptualization, Z.L. and Y.H.; methodology, Z.L. and Y.H.; validation, Z.L. and Y.H.; formal analysis, Z.L., Y.H. and L.Y.; investigation, Z.L., Y.H. and L.Y.; resources, Z.L. and Y.H.; data curation, Z.L. and Y.H.; writing—original draft preparation, Z.L. and Y.H.; writing—review and editing, Z.L. and L.Y.; visualization, Z.L. and Y.H.; supervision, L.Y.; project administration, L.Y. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The dataset used in this study is open access and can be found here: ScanNet dataset: http://www.scan-net.org/, accessed on 4 July 2024; Hypersim dataset: https://github.com/apple/ml-hypersim, accessed on 4 July 2024; Replica dataset: https://github.com/facebookresearch/Replica-Dataset/releases, accessed on 4 July 2024.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Kang, Z.; Yang, J.; Yang, Z.; Cheng, S. A review of techniques for 3d reconstruction of indoor environments. *ISPRS Int. J. Geo-Inf.* **2020**, *9*, 330. [CrossRef]
2. Li, J.; Gao, W.; Wu, Y.; Liu, Y.; Shen, Y. High-quality indoor scene 3d reconstruction with rgb-d cameras: A brief review. *Comput. Vis. Media* **2022**, *8*, 369–393. [CrossRef]
3. Hess, W.; Kohler, D.; Rapp, H.; Andor, D. Real-Time Loop Closure in 2d Lidar Slam. In Proceedings of the2016 IEEE International Conference on Robotics and Automation (ICRA), Stockholm, Sweden, 16–21 May 2016; pp. 1271–1278.
4. Zhang, J.; Singh, S. Loam: Lidar odometry and mapping in real-time. In *Robotics: Science and Systems*; University of California: Berkeley, CA, USA, 2014; Volume 2, pp. 1–9.
5. Shan, T.; Englot, B. Lego-loam: Lightweight and Ground-Optimized Lidar Odometry and Mapping on Variable Terrain. In Proceedings of the 2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Madrid, Spain, 1–5 October 2018; pp. 4758–4765.
6. 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, New Orleans, LA, USA, 4–9 August 1996; pp. 303–312.
7. Newcombe, R.A.; Izadi, S.; Hilliges, O.; Molyneaux, D.; Kim, D.; Davison, A.J.; Kohi, P.; Shotton, J.; Hodges, S.; Fitzgibbon, A. Kinectfusion: Real-Time Dense Surface Mapping and Tracking. In Proceedings of the 2011 10th IEEE International Symposium on Mixed and Augmented Reality, Basel, Switzerland, 26–29 October 2011; pp. 127–136.
8. Murez, Z.; Van As, T.; Bartolozzi, J.; Sinha, A.; Badrinarayanan, V.; Rabinovich, A. Atlas: End-to-end 3d scene reconstruction from posed images. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VII 16*; Springer: Berlin/Heidelberg, Germany, 2020; pp. 414–431.
9. 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, Nashville, TN, USA, 20–25 June 2021; pp. 15598–15607.
10. Schonberger, J.L.; Frahm, J.M. Structure-from-Motion Revisited. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 26 June–1 July 2016; pp. 4104–4113.
11. Schönberger, J.L.; Zheng, E.; Frahm, J.M.; Pollefeys, M. Pixelwise View Selection for Unstructured Multi-View Stereo. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, **2016**, Proceedings, Part III 14*; Springer: Berlin/Heidelberg, Germany, 2016; pp. 501–518.
12. Xu, Q.; Tao, W. Planar prior assisted patchmatch multi-view stereo. *Proc. AAAI Conf. Artif. Intell.* **2020**, *34*, 12516–12523. [CrossRef]
13. Im, S.; Jeon, H.G.; Lin, S.; Kweon, I.S. Dpsnet: End-to-end deep plane sweep stereo. *arXiv* **2019**, arXiv:1905.00538.
14. Wang, F.; Galliani, S.; Vogel, C.; Speciale, P.; Pollefeys, M. Patchmatchnet: Learned Multi-View Patchmatch Stereo. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021; pp. 14194–14203.
15. Xu, Q.; Tao, W. Pvsnet: Pixelwise visibility-aware multi-view stereo network. *arXiv* **2020**, arXiv:2007.07714.
16. 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 (ECCV), Munich, Germany, 8–14 September 2018; pp. 767–783.
17. 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, Seattle, WA, USA, 13–19 June 2020; pp. 1949–1958.
18. Teed, Z.; Deng, J. Deepv2d: Video to depth with differentiable structure from motion. *arXiv* **2018**, arXiv:1812.04605.
19. Huang, P.H.; Matzen, K.; Kopf, J.; Ahuja, N.; Huang, J.B. Deepmvs: Learning Multi-View Stereopsis. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 2821–2830.
20. Cheng, S.; Xu, Z.; Zhu, S.; Li, Z.; Li, L.E.; Ramamoorthi, R.; Su, H. Deep Stereo using Adaptive thin Volume Representation with Uncertainty Awareness. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 2524–2534.
21. Yang, H.; Chen, R.; An, S.P.; Wei, H.; Zhang, H. The growth of image-related three dimensional reconstruction techniquesin deep learning-driven era: A critical summary. *J. Image Graph.* **2023**, *28*, 2396–2409.

22. Liu, S.; Zhang, Y.; Peng, S.; Shi, B.; Pollefeys, M.; Cui, Z. Dist: Rendering deep implicit signed distance function with differentiable sphere tracing. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 2019–2028.
23. Yariv, L.; Kasten, Y.; Moran, D.; Galun, M.; Atzmon, M.; Ronen, B.; Lipman, Y. Multiview neural surface reconstruction by disentangling geometry and appearance. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 2492–2502.
24. 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* **2021**, arXiv:2106.10689.
25. Yariv, L.; Gu, J.; Kasten, Y.; Lipman, Y. Volume rendering of neural implicit surfaces. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 4805–4815.
26. Barron, J.T.; Mildenhall, B.; Verbin, D.; Srinivasan, P.P.; Hedman, P. Mip-nerf 360: Unbounded Anti-Aliased Neural Radiance Fields. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 19–24 June 2022.
27. Liu, L.; Gu, J.; Zaw Lin, K.; Chua, T.S.; Theobalt, C. Neural sparse voxel fields. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 15651–15663.
28. Rebain, D.; Matthews, M.; Yi, K.M.; Lagun, D.; Tagliasacchi, A. Lolnerf: Learn from one look. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 19–24 June 2022.
29. Gafni, G.; Thies, J.; Zollhofer, M.; Nießner, M. Dynamic Neural Radiance Fields for Monocular 4d Facial Avatar Reconstruction. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021.
30. Do, T.; Vuong, K.; Roumeliotis, S.I.; Park, H.S. Surface Normal Estimation of Tilted Images Via Spatial Rectifier. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*; Springer International Publishing: New York, NY, USA, 2020; pp. 265–280.
31. 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 Conference on Computer vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 5828–5839.
32. Roberts, M.; Ramapuram, J.; Ranjan, A.; Kumar, A.; Bautista, M.A.; Paczan, N.; Webb, R.; Susskind, J.M. Hypersim: A Photorealistic Synthetic Dataset for Holistic Indoor Scene Understanding. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 10912–10922.
33. Straub, J.; Whelan, T.; Ma, L.; Chen, Y.; Wijmans, E.; Green, S.; Engel, J.J.; Mur-Artal, R.; Ren, C.; Verma, S.; et al. The replica dataset: A digital replica of indoor spaces. *arXiv* **2019**, arXiv:1906.05797.
34. 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* **2021**, *65*, 99–106. [CrossRef]
35. Guo, H.; Peng, S.; Lin, H.; Wang, Q.; Zhang, G.; Bao, H.; Zhou, X. Neural 3d scene reconstruction with the Manhattan-world assumption. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 19–24 June 2022; pp. 5511–5520.
36. 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.* **2022**, *35*, 25018–25032.
37. Zhu, J.; Huo, Y.; Ye, Q.; Luan, F.; Li, J.; Xi, D.; Wang, L.; Tang, R.; Hua, W.; Bao, H.; et al. I2-SDF: Intrinsic Indoor Scene Reconstruction and Editing via Raytracing in Neural SDFs. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 18–22 June 2023; pp. 12489–12498.