# Indoor Scene Reconstruction with Fine-Grained Details Using Hybrid Representation and Normal Prior Enhancement

Sheng Ye, Yubin Hu, Matthieu Lin, Yu-Hui Wen*, Wang Zhao, Yong-Jin Liu*, *Senior Member, IEEE*, and Wenping Wang, *Fellow, IEEE*

**Abstract**—The reconstruction of indoor scenes from multi-view RGB images is challenging due to the coexistence of flat and textureless regions alongside delicate and fine-grained regions. Recent methods leverage neural radiance fields aided by predicted surface normal priors to recover the scene geometry. These methods excel in producing complete and smooth results for floor and wall areas. However, they struggle to capture complex surfaces with high-frequency structures due to the inadequate neural representation and the inaccurately predicted normal priors. This work aims to reconstruct high-fidelity surfaces with fine-grained details by addressing the above limitations. To improve the capacity of the implicit representation, we propose a hybrid architecture to represent low-frequency and high-frequency regions separately. To enhance the normal priors, we introduce a simple yet effective image sharpening and denoising technique, coupled with a network that estimates the pixel-wise uncertainty of the predicted surface normal vectors. Identifying such uncertainty can prevent our model from being misled by unreliable surface normal supervisions that hinder the accurate reconstruction of intricate geometries. Experiments on the benchmark datasets show that our method outperforms existing methods in terms of reconstruction quality. Furthermore, the proposed method also generalizes well to real-world indoor scenarios captured by our hand-held mobile phones. Our code is publicly available at: https://github.com/yec22/Fine-Grained-Indoor-Recon.

**Index Terms**—Surface reconstruction, neural radiance fields, hybrid representation, normal prior enhancement

## 1 Introduction

3D reconstruction of a target scene from a sequence of multi-view RGB images is a fundamental problem in computer vision, which has a wide range of applications in various fields, including robotics, filming, gaming, virtual reality, and so on. Specifically, high-fidelity reconstruction of indoor scenes is extremely challenging because indoor scenes have both large, flat areas (floor, wall, roof, *etc.*) and high-frequency, fine-grained areas (small stuff on the table, delicate furniture, *etc.*).

Encoding scenes with a neural implicit function [1], [2], [3] has attracted increasing attention due to its compactness and good performance. Some recent works have achieved remarkable results in 3D scene reconstruction from 2D supervision, by combining the neural implicit function and differentiable volume rendering. NeuS [4] and VolSDF [5] can reconstruct high-quality watertight surfaces of a single object by representing the geometry as an implicit signed distance field (SDF). Follow-up works try to integrate the implicit SDF representation and additional geometric priors, including semantic segmentation [6], monocular depth [7], or surface normals [8], to address the challenge of indoor scene reconstruction. Compared to traditional MVS-based [9], [10] or TSDF-based [11], [12] approaches, these neural implicit methods can produce complete and promising surfaces, but still struggle with delicate and complex structures.

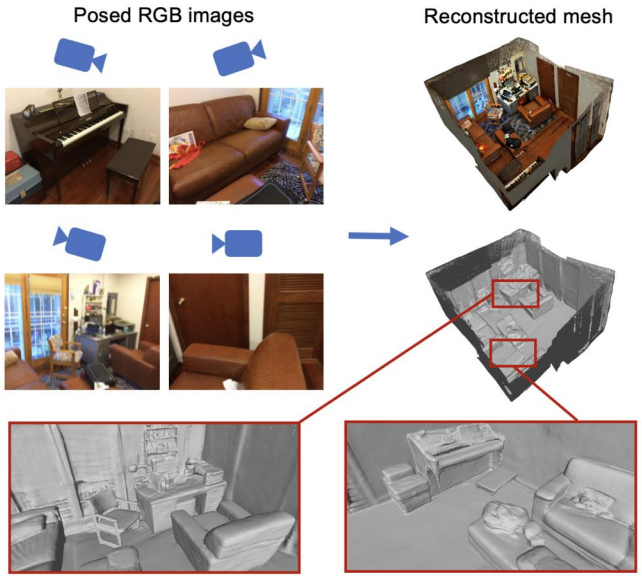


Fig. 1. Our proposed method can reconstruct fine and accurate indoor scenes only from a sequence of posed RGB images.

In this paper, we aim to reconstruct high-fidelity indoor scene surfaces with fine and accurate details utilizing the


- *S. Ye, Y.B. Hu, M. Lin, W. Zhao, and Y.-J. Liu are with BNRist, the Department of Computer Science and Technology, Tsinghua University. E-mail: {yec22, huyb20, yh-lin21, zhao-w19}@mails.tsinghua.edu.cn, liuyongjin@tsinghua.edu.cn.*
- *Y.-H. Wen is with Beijing Key Laboratory of Traffic Data Analysis and Mining, School of Computer and Information Technology, Beijing Jiaotong University. E-mail: yhwen1@bjtu.edu.cn.*
- *W.P. Wang is with the Department of Computer Science & Engineering, Texas A&M University. E-mail: wenping@tamu.edu.*
- ** Corresponding author*

implicit SDF framework [4]. We first analyze and conclude two main limitations of the existing neural implicit methods that lead to the failure of fine-grained reconstruction. One of the limitations is that existing methods typically use a large MLP network to approximate the underlying geometry function of the entire scene. Indeed, they can produce surprisingly good results for object-level surface reconstruction, but are less capable of recovering more complicated indoor scenes. According to SIREN [3] and MonoSDF [7], the deep MLP network possesses an inductive smoothness bias, and is therefore suitable to encode surfaces of flat and low-frequency areas. However, indoor scenes also consist of many high-frequency and fine-grained regions, which are difficult for a single MLP to express. Therefore, a novel architecture with greater expressive capacity is needed. Another limitation is that the surface normal priors used to regularize the implicit SDF may be inaccurate, especially in regions with complex structures. These priors are typically monocular estimated by off-the-shelf neural networks [13], [14], which tend to produce noisy and erroneous results for thin and delicate structures. Models will be misled to generate wrong geometries when directly supervised by these inaccurate priors. Therefore, it is necessary to improve the quality of the predicted normal priors and design a mechanism to measure their reliability.

To improve the capacity of expressing both the flat areas and fine details, we design a novel hybrid representation. We implicitly decouple an indoor scene into different regions and exploit suitable neural representations to encode these regions separately. Specifically, we utilize a branch of MLP network to encode the rough outline of the scene, and another branch of tri-plane features with a shallow decoder to represent the fine-grained details. To recover fine geometric details, some previous studies [15], [16], [17] attempt to utilize the voxel grid features. Nevertheless, they require complex training schemes and regularizations since the voxel representation tends to produce noisy surfaces. Moreover, the voxel grids are also memory-intensive due to their cubically growing consumption. Inspired by the remarkable performance of tri-plane representation in encoding human faces [18], [19], we choose this memory-efficient representation instead of voxel grids to encode the delicate details. In our experiments, we empirically find that the tri-plane branch is suitable for expressing high-frequency details, which complements the MLP branch. In addition, our proposed hybrid architecture does not require complicated strategies for training.

As for the enhancement of predicted normal priors, we observe that artifacts present in the input RGB images can greatly affect the quality of predicted normals. Thus, we devise a sharpening and denoising technique before feeding the images into the prior estimation module, which effectively reduces the errors in predicted normals. Furthermore, we design an uncertainty module, which predicts the pixel-wise uncertainty maps of the estimated normal priors. Besides the RGB images and monocular estimated normals, this module also takes features from the visual foundation model [20] as input to supplement the prediction with high-level structural information. We then treat the predicted uncertainty as the weight of normal prior supervision. This uncertainty module determines the reliability of normal priors and reduces the negative impact of incorrectly predicted priors on reconstruction quality. In conclusion, we summarize our contributions as follows:

- We propose a novel hybrid implicit SDF architecture that incorporates MLP and tri-plane to better represent the low-frequency and high-frequency regions of indoor scenes simultaneously.
- We design an image enhancement technique and an uncertainty module to improve the quality of predicted normal priors and guide our network to effectively leverage more accurate priors.
- Qualitative and quantitative experiments show that results produced by our method are better than state-of-the-art methods. Apart from the commonly used datasets, our method also generalizes well to real-world indoor scenes captured by ourselves, demonstrating the potential for practical applications.

## 2 Related Works

### 2.1 Neural Surface Representation

Representing geometric surfaces by neural implicit functions has recently received increasing attention, because of its compactness and remarkable performance. The seminal work of DeepSDF [1] first proposes to use a neural network to model a signed distance field, which encodes the underlying geometry of the target object. However, DeepSDF requires ground-truth 3D meshes to supervise the learning process. Recent methods incorporate neural implicit functions and differentiable rendering to reconstruct surfaces only from the supervision of multi-view 2D images. DVR [21] first proposes a differentiable rendering formulation for implicit shape and texture representations. IDR [22] further designs an architecture that simultaneously learns the implicit SDF field, camera poses, and a neural renderer. UNISURF [23] utilizes occupancy to represent implicit surfaces and modifies the volume rendering equation correspondingly. Subsequently, NeuS [4] and VolSDF [5] define the density used in volume rendering process as logistic sigmoid function and Laplace's cumulative distribution function applied to an SDF representation. Nevertheless, all those works use a single MLP to approximate the implicit function, which struggles to scale to large and complicated scenes due to the network capacity. There also exist other researches [15], [16], [24], [25] that use multi-scale voxel grids followed by a shallow decoder to represent the scenes, but this leads to non-smoothness. Neuralangelo [26] leverages 3D hash grids to recover dense 3D surface structures. However, hash grids also suffer from localities and hash collisions. In this work, we devise a novel hybrid representation incorporating MLP and tri-plane to enhance the network's ability to express high-frequency and delicate surfaces, while maintaining non-local smoothness.

### 2.2 Indoor Scene Reconstruction

Conventional multi-view stereo (MVS) methods [9], [10] often struggle to reconstruct large texture-less indoor scene areas densely. With the development of large-scale indoor datasets [27], [28], learning-based MVS methods are proposed to alleviate this problem. These methods [29], [30],

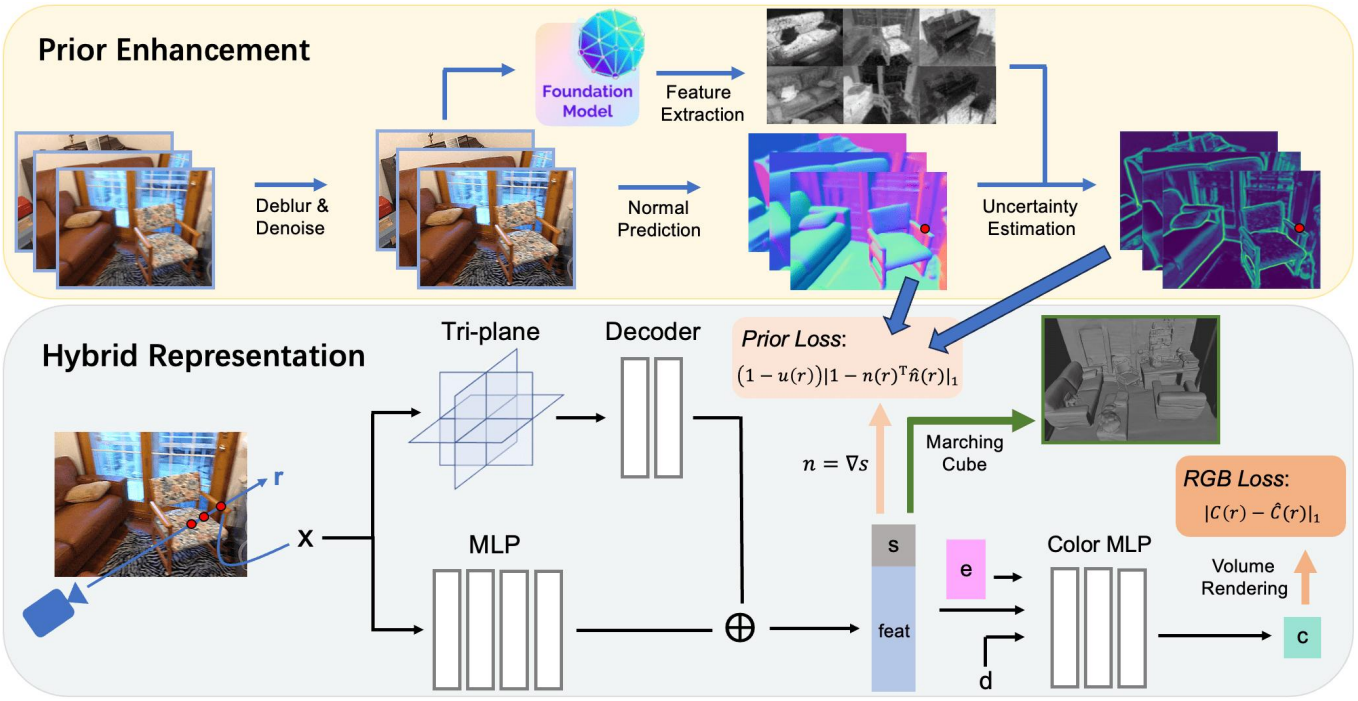


Fig. 2. The overall pipeline of our proposed approach. We tackle high-frequency regions in scene reconstruction from the perspective of both the representation and normal priors. In particular, we propose a hybrid geometry representation to enhance the expressive power, and an image preprocessing technique along with pixel-wise uncertainty to enhance the normal priors.

[31] usually predict a depth map for each frame and fuse these depth maps to build the final scene. However, the resulting surfaces tend to be noisy and incomplete due to the depth inconsistency problem. Another research branch [11], [12], [32], [33] proposes directly regressing input images to truncated signed distance function (TSDF) volume. Then, the meshes can be extracted from TSDF using the marching-cube algorithm. Due to the limitation of TSDF volume resolution, the generated surfaces often lack details.

Inspired by NeRF's [34] excellent performance in novel view synthesis tasks, some recent works attempt to leverage implicit surface representation and differentiable volume rendering to reconstruct indoor scenes. Considering the complexity of indoor scenes, additional priors are also provided to recover a plausible geometry. ManhattanSDF [6] utilizes extra 2D semantic segmentation to detect wall and floor regions, and applies geometry regularization based on Manhattan-world assumption. NeuRIS [8] and MonoSDF [7] exploit monocular estimated normal and depth priors to reconstruct smooth surfaces. NeuRIS also devises a novel cross-view geometric checking technique. HelixSurf [35] attempts to combine PatchMatch-based MVS and neural radiance fields, and proposes a joint optimization pipeline. Although these neural implicit methods significantly improve the reconstruction quality compared to conventional methods, surfaces in complex and fine-grained regions are still unsatisfactory. Furthermore, few studies have explored how to reduce the negative impact of inaccurate priors on the final reconstruction results.

### 2.3 Uncertainty Estimation

Uncertainty estimation, which aims to quantify the reliability of predictions, can help recognize failure scenarios and enable robust applications. Uncertainty is crucial for many computer vision tasks such as optical flow [36], SLAM [37], and multi-view stereo [38]. Predicted uncertainty is used to facilitate the optimization through robust loss [39], guided sampling [40], outlier rejection [41], *etc*. However, the integration of the neural radiance field with uncertainty remains a relatively unexplored terrain. Some current efforts [42], [43] mainly focus on the task of rendering, while we focus on surface reconstruction. In this work, we novelly leverage the estimated uncertainty in a weighted loss to guide our model to utilize more precise priors while avoiding misguidance by less accurate ones.

## 3 Method

In this work, our goal is to reconstruct high-fidelity, room-scale surfaces with fine and accurate details from multiple calibrated RGB images. We represent the scene geometry as an implicit signed distance field encoded by our proposed hybrid architecture. Our architecture can learn 3D geometric information only from 2D image supervision by applying SDF-based volume rendering. As indoor scenes often contain flat and textureless areas, we also utilize estimated surface normal priors as additional geometric constraints. To effectively leverage normal priors, we design an enhancement technique and an uncertainty mechanism for better reconstruction quality. Figure 2 summarizes the overall pipeline of our proposed approach.

In the subsequent sections, we first briefly revisit the preliminary of the SDF-based volume rendering equation. And then, we discuss and describe the details of our proposed hybrid geometric representation, prior enhancement

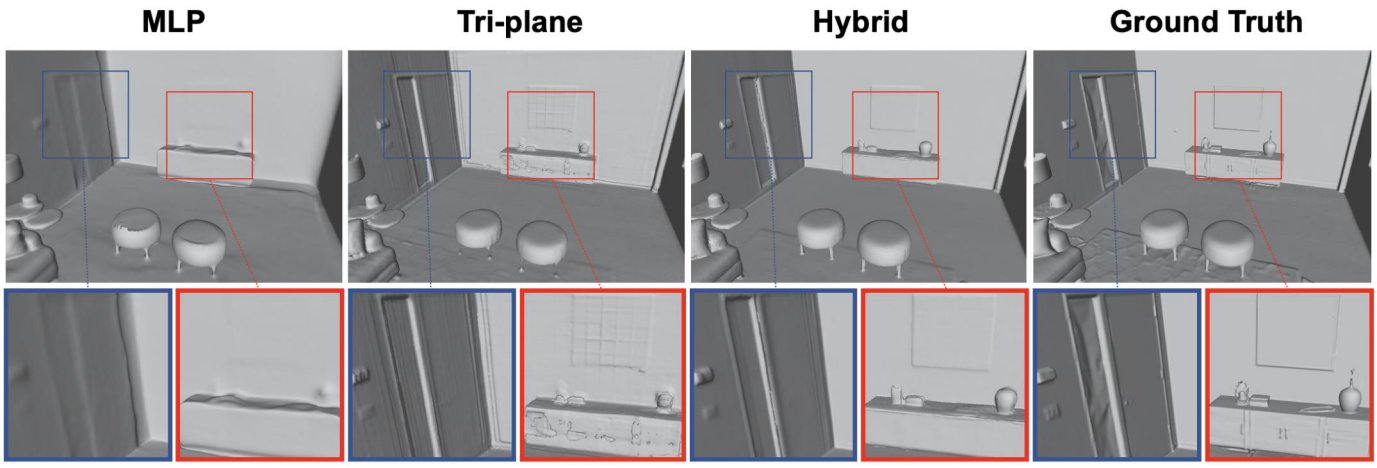


Fig. 3. Comparison of the reconstructed meshes generated by different representations. Zoom in for better visualization.

techniques, and uncertainty estimation module. Finally, we introduce the losses that are used to optimize our model.

### 3.1 Preliminary

A continuous 3D scene can be modeled as a signed distance field $f_g$ and a color field $f_c$ [4], [8]. Both of the two fields are typically represented by neural networks. The differentiable volume rendering technique can learn the signed distance and color fields only from 2D image supervision. We denote a ray emitted from a viewing camera as $\boldsymbol{r}(t) = \boldsymbol{o}+t\boldsymbol{d}$, where $\boldsymbol{o}$ is the camera center and $\boldsymbol{d}$ is the viewing direction. Given a set of sample points $\{\boldsymbol{r}(t_i)|i = 1, ..., N\}$ along the ray, the corresponding pixel color of this ray can be calculated as

$$C(\boldsymbol{r}) = \sum_{i=1}^{N} T_i \alpha_i f_c(\boldsymbol{r}(t_i), \boldsymbol{d}),\ T_i = \prod_{j=1}^{i-1} (1 - \alpha_j), \quad (1)$$

where $T_i$ is the accumulated transmittance, and $\alpha_i$ denotes the opacity of the $i$-th ray segment and is defined as

$$\alpha_i = \max\left(0, \frac{\Phi_\tau(f_g(\boldsymbol{r}(t_i))) - \Phi_\tau(f_g(\boldsymbol{r}(t_{i+1})))}{\Phi_\tau(f_g(\boldsymbol{r}(t_i)))}\right), \quad (2)$$

where $\Phi_\tau$ is the Sigmoid function with a learnable parameter $\tau$. Then, the difference between the rendered pixel color $C(\boldsymbol{r})$ and the ground-truth pixel color $\hat{C}(\boldsymbol{r})$ is minimized to optimize the $f_g$ and $f_c$. After optimization, the surface $\mathcal{S}$ can be extracted as the zero level-set of the signed distance field $\mathcal{S} = \{\boldsymbol{x}|f_g(\boldsymbol{x}) = 0\}$.

### 3.2 Hybrid Representation for Geometry

Previous methods [6], [8], [35] usually utilize a single MLP to represent the geometry of the entire indoor scene. Although equipped with positional encodings that encompass high-frequency bands, these methods still tend to produce low-frequency surfaces and struggle to reconstruct complex and delicate structures. Recent researches [3], [7] point out that the reasons may lie in the smoothness bias of MLP networks. Studies [7], [16] have also tried to replace the MLP with voxel grid features to recover fine details. However, the voxel representation is less compact due to its cubic memory consumption and is prone to generating noisy and inconsistent surfaces, which hinders the reconstruction of flat areas that frequently appear in indoor settings.

To encode the high-frequency features of indoor scenes effectively and efficiently, we attempt to use the tri-plane representation. As optimization updates only propagate to local plane features instead of all parameters (*i.e.*, locality), we find that tri-plane representations are better at expressing fine-grained details than MLP. Besides, tri-planes are also memory-efficient, as their memory consumption only grows quadratically with the resolution. As shown in Figure 3, our preliminary experiments indicate the potential of using the tri-plane for scene reconstruction. The reconstructed 3D meshes of the tri-plane representation are sharper and possess more details than those generated by MLP representation. However, due to the locality nature of the tri-plane, we also observe that using the tri-plane representation alone will cause unwanted artifacts on planar regions.

Empirically, we find that different areas of indoor scenes have different characteristics. The outer contours of a scene are usually smooth and flat, while the interior often contains delicate structures. Thus, in order to accurately reconstruct both the flat areas and the fine-grained details, we propose a hybrid geometry architecture, which contains an MLP branch and a tri-plane branch (see Figure 2). We aim to exploit the smoothness bias of MLP to encode the rough outline of the scene and the high-frequency bias of tri-plane features to encode the delicate structures. To encourage the two branches to focus on different areas of the scene, we find that the initialization plays a vital role. Following SAL [44], we initialize the MLP branch to produce approximate SDF of a unit bounding sphere, and initialize the tri-plane branch to produce a zero SDF. In this way, we implicitly decouple an indoor scene into two distinct regions without the need for extra segmentation annotations. Given a 3D position $\boldsymbol{x} \in \mathbb{R}^3$, our pipeline passes it through the MLP branch to get a coarse SDF $\tilde{\boldsymbol{s}}$ and a coarse hidden feature $\tilde{\boldsymbol{h}}$. Simultaneously, the tri-plane branch outputs the high-frequency complements. Our tri-plane representation $T$ consists of three axis-aligned orthogonal feature planes. Each feature plane has the dimension of $N \times N \times C$, where $N$ is the spatial resolution, and $C$ is the number of channels. We retrieve

the final feature $[T_{xy}, T_{xz}, T_{yz}]$ by projecting $\boldsymbol{x}$ onto these planes to find the tri-plane features and concatenating them together. A shallow, two-layer decoder further regresses the aggregated feature $[T_{xy}, T_{xz}, T_{yz}]$ to the residual SDF $\Delta \boldsymbol{s}$ and residual hidden feature $\Delta \boldsymbol{h}$. The final outputs are the summation of these two branches

$$\boldsymbol{s} = \tilde{\boldsymbol{s}} + \Delta \boldsymbol{s},\ \boldsymbol{h} = \tilde{\boldsymbol{h}} + \Delta \boldsymbol{h}. \tag{3}$$

Subsequently, we feed the hidden feature $\boldsymbol{h}$ and the viewing direction $\boldsymbol{d}$ into the color network to obtain the view-dependent emitted radiance $\boldsymbol{c}$. Inspired by NeRF-W [45], we also attach a learnable appearance embedding $\boldsymbol{e}$ for each view as an extra input to the color network. Appearance embeddings aim to compensate for photometric and environmental variations between images and guarantee a multi-view consistent reconstruction. Our experiments demonstrate that cleaner and sharper surfaces can be obtained by using appearance embeddings.

During the training process, we optimize the hybrid geometry architecture, color network, and appearance embeddings simultaneously. As shown in Figure 3, our hybrid architecture combines the advantages of both MLP and tri-plane, preserving the sharp details and eliminating the artifacts as well. In conclusion, our method leverages characteristics of different regions of indoor scenes and, for the first time, exploits suitable neural representations to encode these regions separately.

### 3.3 Prior Enhancement

#### 3.3.1 Image Sharpening and Denoising Techniques

Surface normal priors can serve as globally consistent geometric constraints to improve the reconstruction quality. Thus, we use the monocular predicted normal priors as extra supervision, which can regularize the SDF especially in textureless regions where color supervision is insufficient. Specifically, we can obtain the surface normal $\boldsymbol{n}$ under certain viewpoints using a similar volume accumulation process as Equation (1):

$$\boldsymbol{n}(\boldsymbol{r}) = \sum_{i=1}^{N} T_i \alpha_i \nabla \boldsymbol{s}_i, \tag{4}$$

where $\nabla \boldsymbol{s}_i$ is the gradient of SDF at different sample points. During training, we minimize the difference between the normal calculated by our model and the pseudo ground-truth normal estimated by off-the-shelf tools [13], [14].

It is worth noting that the estimated normal priors are not always accurate. We observe that artifacts presented in the input RGB images, especially noise and motion blur, can significantly affect the quality of estimated normal priors. These artifacts are prevalent in real-world scenarios. Thus, we propose to use a sharpening and denoising image enhancement before predicting the normal maps. For the sharpening process, we use a $3 \times 3$ kernel to perform convolution on the whole input image. This kernel can be regarded as an edge detector, which amplifies the difference between the center pixel and its surrounding pixels. Formally, this operation is defined as

$$p_i' = 9 \cdot p_i - \sum_{p_j \in \mathcal{N}_i} p_j, \tag{5}$$

where $p_i$ is the center pixel and $\mathcal{N}_i$ is the eight neighboring pixels of $p_i$ inside the $3 \times 3$ kernel. The sharpening process alleviates motion blur and strengthens edges in the images, facilitating the subsequent normal estimation of thin structures. For the denoising process, we employ the median filtering technique that replaces an image pixel with the median value of its neighbors. This operation can effectively remove the isolated noise while maintaining the sharpness of the image. We find that error regions of the estimated normal priors are reduced after applying the denoising operation (more details can be found in our experiments). Although we apply the sharpening and denoising per image, the appearance embeddings introduced in Section 3.2 can handle the slight cross-view inconsistencies caused by image enhancement. Our sharpening and denoising techniques indicate that image processing in 2D can indeed affect the reconstruction quality in 3D, and the bridge between this is surface normal and volume rendering.

#### 3.3.2 Uncertainty Estimation Module

Our sharpening and denoising techniques can reduce the normal estimation errors caused by imperfect image quality. However, it is still difficult to predict accurate surface normal maps of complex and intricate regions, even with high-quality input images. Directly regularizing the implicit SDF field with these inaccurate surface normal predictions can result in degenerated geometries that lack delicate structures. Therefore, we attach a pixel-wise uncertainty map to each estimated normal map, which reflects the reliability of different regions of the normal maps.

For regions with lower uncertainty values, our method depends more on the estimated surface normal priors to achieve fast and accurate geometry reconstruction. For delicate regions with higher uncertainty values, our method assigns a smaller weight to the surface normal loss, such that the optimization of the SDF field depends more on the realistic RGB information and avoids being misled by inaccurate surface normal predictions. Accordingly, we design an uncertainty weighted normal prior loss function to assign larger (smaller) weights to the low (high) uncertainty areas when using these normal priors for supervision, which can be formulated as follows,

$$\mathcal{L}_{prior} = \frac{1}{|\mathcal{R}|} \sum_{\boldsymbol{r} \in \mathcal{R}} (1 - \boldsymbol{u}(\boldsymbol{r})) \left\| 1 - \boldsymbol{n}(\boldsymbol{r})^{\top} \hat{\boldsymbol{n}}(\boldsymbol{r}) \right\|_1, \tag{6}$$

where $\mathcal{R}$ is the set of rays in each batch. $\boldsymbol{n}(\boldsymbol{r})$ and $\hat{\boldsymbol{n}}(\boldsymbol{r})$ are the rendered normal and the pseudo ground-truth normal. Specifically, for the uncertainty value corresponding to a ray $\boldsymbol{r}$, we normalize it to $[0, 1]$, denoted as $\boldsymbol{u}(\boldsymbol{r})$. When $\mathcal{L}_{prior}$ supervises the implicit SDF field together with the typical color loss, our proposed uncertainty values can be viewed as a trade-off between the color information and normal information. For complicated regions with high uncertainty, the proposed loss reduces the weight of normal supervision, thereby strengthening the role of color supervision. By balancing the supervision weights, our model

TABLE 1
Definitions of the metrics to evaluate the 3D reconstruction quality.

| 3D Metric | Definition |
|---|---|
| *Accuracy* | $\text{mean}_{p\in P}(\min_{p^*\in P^*} \|p-p^*\|)$ |
| *Completeness* | $\text{mean}_{p^*\in P^*}(\min_{p\in P} \|p-p^*\|)$ |
| *Precision* | $\text{mean}_{p\in P}(\min_{p^*\in P^*} \|p-p^*\| < \textit{threshold})$ |
| *Recall* | $\text{mean}_{p^*\in P^*}(\min_{p\in P} \|p-p^*\| < \textit{threshold})$ |
| *F-Score* | $\frac{2\times Precision\times Recall}{Precision+Recall}$ |

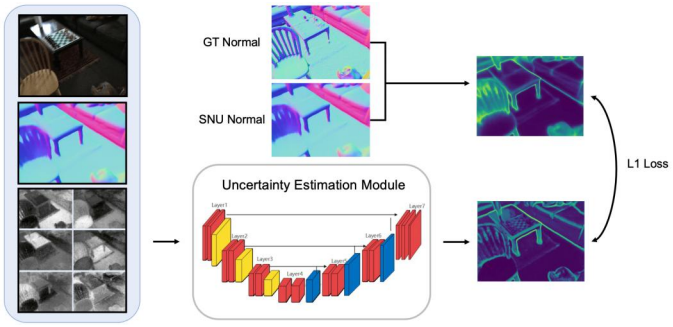


Fig. 4. We propose an uncertainty estimation module to predict the pixel-wise uncertainty maps of the normal priors.

TABLE 2
Quantitative comparison of reconstruction quality on ScanNet dataset (threshold = 0.05). Bold indicates the best.

| Methods | Acc↓ | Comp↓ | Prec↑ | Recall↑ | F-score↑ |
|---|---|---|---|---|---|
| COLMAP | 0.047 | 0.235 | 0.711 | 0.441 | 0.537 |
| NeuS | 0.179 | 0.208 | 0.313 | 0.275 | 0.291 |
| NeuralAngelo | 0.132 | 0.109 | 0.505 | 0.467 | 0.485 |
| ManhattanSDF | 0.072 | 0.068 | 0.621 | 0.586 | 0.602 |
| NeuRIS | 0.051 | 0.048 | 0.720 | 0.674 | 0.696 |
| MonoSDF | 0.035 | 0.048 | 0.799 | 0.681 | 0.733 |
| HelixSurf | 0.038 | 0.044 | 0.786 | 0.727 | 0.755 |
| Ours | **0.033** | **0.041** | **0.814** | **0.737** | **0.773** |

TABLE 3
Quantitative comparison of reconstruction quality on Replica dataset (threshold = 0.05). Bold indicates the best.

| Methods | Acc↓ | Comp↓ | Prec↑ | Recall↑ | F-score↑ |
|---|---|---|---|---|---|
| COLMAP | 0.030 | 0.095 | 0.872 | 0.530 | 0.658 |
| NeuS | 0.187 | 0.290 | 0.201 | 0.133 | 0.161 |
| NeuralAngelo | 0.103 | 0.125 | 0.445 | 0.392 | 0.417 |
| ManhattanSDF | 0.145 | 0.194 | 0.611 | 0.427 | 0.501 |
| NeuRIS | 0.034 | 0.070 | 0.803 | 0.712 | 0.754 |
| MonoSDF | 0.024 | 0.030 | 0.933 | 0.898 | 0.915 |
| Ours | **0.022** | **0.026** | **0.942** | **0.931** | **0.936** |

can resist being misled by unreliable normal priors and effectively utilize more reliable normal priors.

To predict the pixel-wise uncertainty of estimated normal maps, we design a novel module shown in Figure 4. This module adopts a U-Net [46] architecture with RGB images and the estimated normal maps as input. To facilitate the uncertainty prediction, we also input high-level DINO [20] features, since the normal uncertainty is usually structure-relevant. We choose the visual foundation model DINO rather than CLIP [47], as DINO tends to capture pixel-level structure-aware image features during its self-supervised learning process, while CLIP features tend to be global and linguistic-related. Furthermore, a previous work [48] also demonstrates the effectiveness of DINO features in facilitating downstream visual understanding tasks, outperforming other foundation models (*e.g.*, CLIP [47] and DeiT [49]). We extract features from the last attention layer of DINO, and then apply PCA technique to retain the most essential information. We train this uncertainty estimation module on a subset (disjoint from the test set) of the ScanNet [27]. Concretely, we compute the difference between the estimated normals and normals rendered from real 3D meshes to obtain the ground-truth uncertainty maps, and use them to supervise our proposed estimation module.

### 3.4 Loss Functions

The overall loss is the weighted sum of normal prior loss $\mathcal{L}_{prior}$, Eikonal loss $\mathcal{L}_{eik}$, and RGB color loss $\mathcal{L}_{rgb}$:

$$\mathcal{L} = \lambda_p \mathcal{L}_{prior} + \lambda_e \mathcal{L}_{eik} + \lambda_r \mathcal{L}_{rgb}. \tag{7}$$

All reasonable signed distance fields must satisfy the Eikonal equation [50], which constrains the gradient of SDF to be equal to 1. Thus, the Eikonal loss $\mathcal{L}_{eik}$ to regularize the gradient of SDF is denoted by

$$\mathcal{L}_{eik} = \frac{1}{N}\sum_{i=1}^{N} (\|\nabla \boldsymbol{s}_i\|_2 - 1)^2, \tag{8}$$

where $N$ is the total number of sampled points, and $\nabla \boldsymbol{s}_i$ is the gradient of SDF at different sample points. The RGB color loss $\mathcal{L}_{rgb}$ is defined as

$$\mathcal{L}_{rgb} = \frac{1}{|\mathcal{R}|}\sum_{\boldsymbol{r}\in\mathcal{R}} \left\| C(\boldsymbol{r}) - \hat{C}(\boldsymbol{r}) \right\|_1, \tag{9}$$

where $C(\boldsymbol{r})$ and $\hat{C}(\boldsymbol{r})$ are the predicted and ground-truth RGB colors for ray $\boldsymbol{r}$ respectively, and $\mathcal{R}$ is the set of rays in each batch.

## 4 Experiments

### 4.1 Experiment Details

#### 4.1.1 Datasets

We mainly conduct the quantitative and qualitative experiments on two commonly used benchmark datasets: ScanNet [27] and Replica [28]. Furthermore, we also demonstrate the generalization ability of our method on several real-world indoor scenes captured by ourselves.

**ScanNet.** It is a real-world indoor dataset containing various scenes captured with Kinect V1 RGB-D cameras. The BundleFusion [51] is then used to provide camera poses and 3D meshes. For ScanNet, we use the same test split from ManhattanSDF [6].

**Replica.** It is a synthetic dataset which provides high-quality RGB images, dense geometry, and semantic annotations. For Replica, we follow the settings of MonoSDF [7].

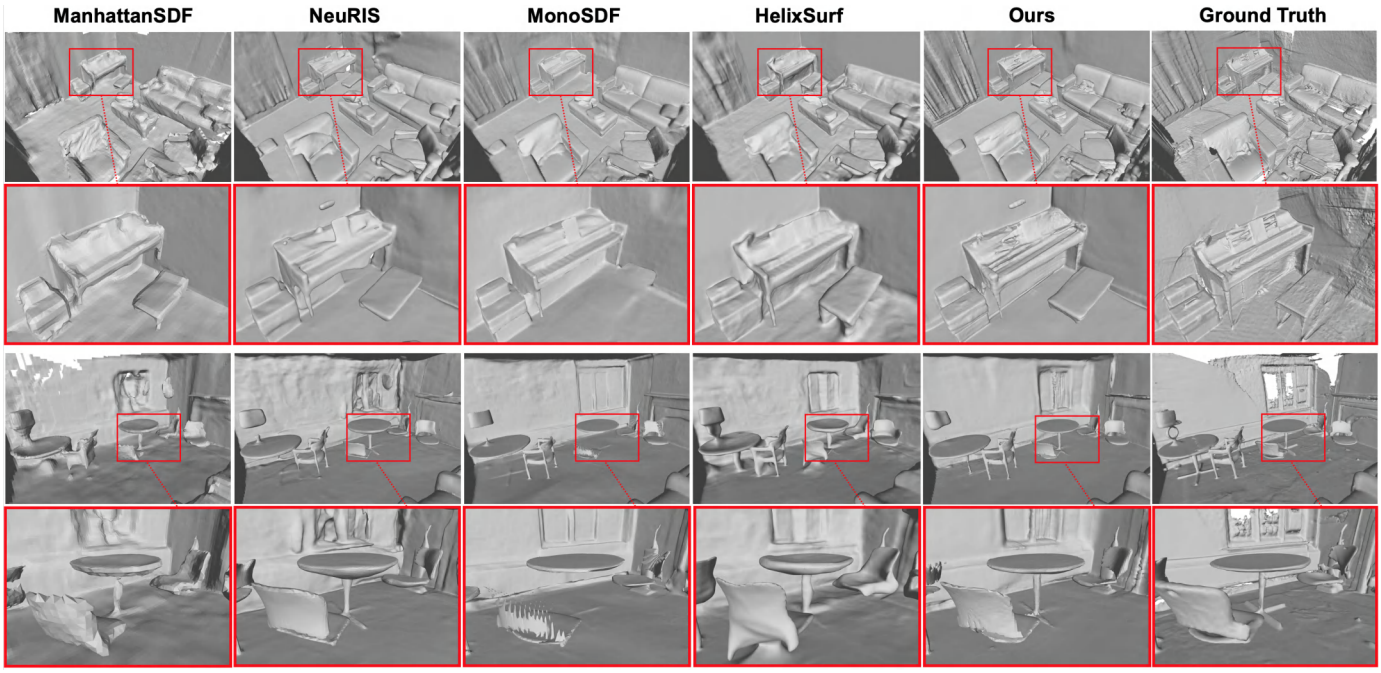


Fig. 5. Qualitative comparison of reconstructed meshes with other baselines on ScanNet dataset.

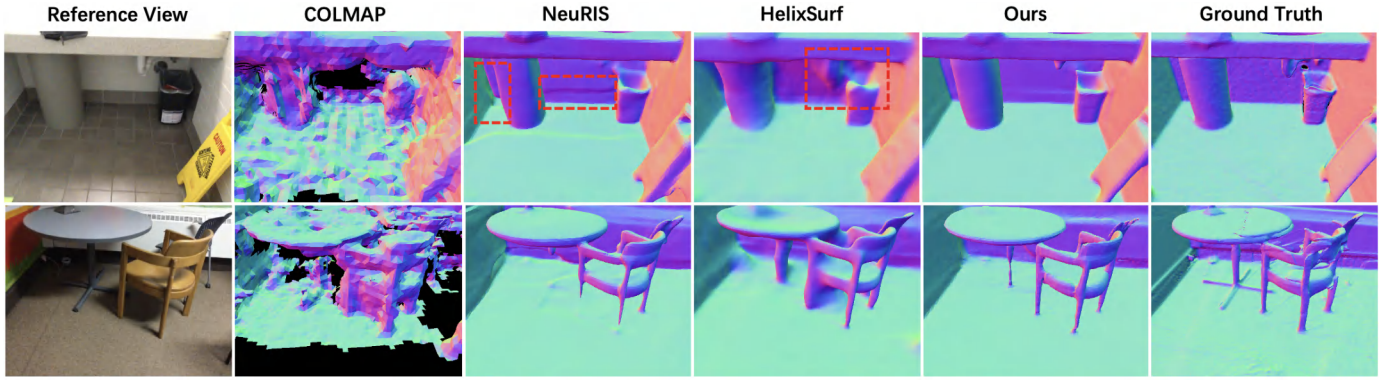


Fig. 6. Qualitative comparison of rendered normal maps with other baselines on ScanNet dataset.

#### 4.1.2 Compared Methods

We compare our method with the following baselines: (1) Traditional MVS method COLMAP [10]; (2) State-of-the-art neural implicit surface reconstruction methods, including NeuS [4] and NeuralAngelo [26]; (3) State-of-the-art neural indoor scene reconstruction methods, including ManhattanSDF [6], NeuRIS [8], MonoSDF [7], and HelixSurf [35]. Since HelixSurf does not fully open source the data preparation code and provides only the preprocessed data on ScanNet, we exclude HelixSurf from experiments on Replica dataset. All baselines are trained from scratch.

#### 4.1.3 Evaluation Metrics

Following previous methods [6], [8], [11], [35], we use five standard metrics to evaluate the reconstruction geometry quality: *Accuracy*, *Completeness*, *Precision*, *Recall*, and *F-score*. These metrics are defined in Table 1, where $P$ and $P^*$ denote the sample points from the predicted and ground-truth mesh.

#### 4.1.4 Implementation

Our proposed model is experimented on an NVIDIA A100 GPU. For the geometry network, we adopt an 8-layer MLP, and $256 \times 256 \times 16$ dimension tri-plane features. For the color network, we adopt a 4-layer MLP. The uncertainty estimation module leverages a U-Net architecture with skip connections. Specifically, this uncertainty module contains four downsample convolution blocks and four upsample blocks. We select 300 scenes (disjoint from the test set) from the ScanNet and construct 12696 data pairs to train this module. We implement our model in Pytorch using Adam optimizer. The loss weights are set to $\lambda_p = \lambda_r = 1.0$, and $\lambda_e = 0.1$. In each batch, $|\mathcal{R}| = 512$. The training process takes 40k - 50k iterations depending on the scene's complexity, and it typically costs 1.5 - 2 hours.

### 4.2 Comparisons

#### 4.2.1 Quantitative Evaluation

Table 2 and Table 3 summarize the quantitative comparison results on ScanNet dataset and Replica dataset. We compare

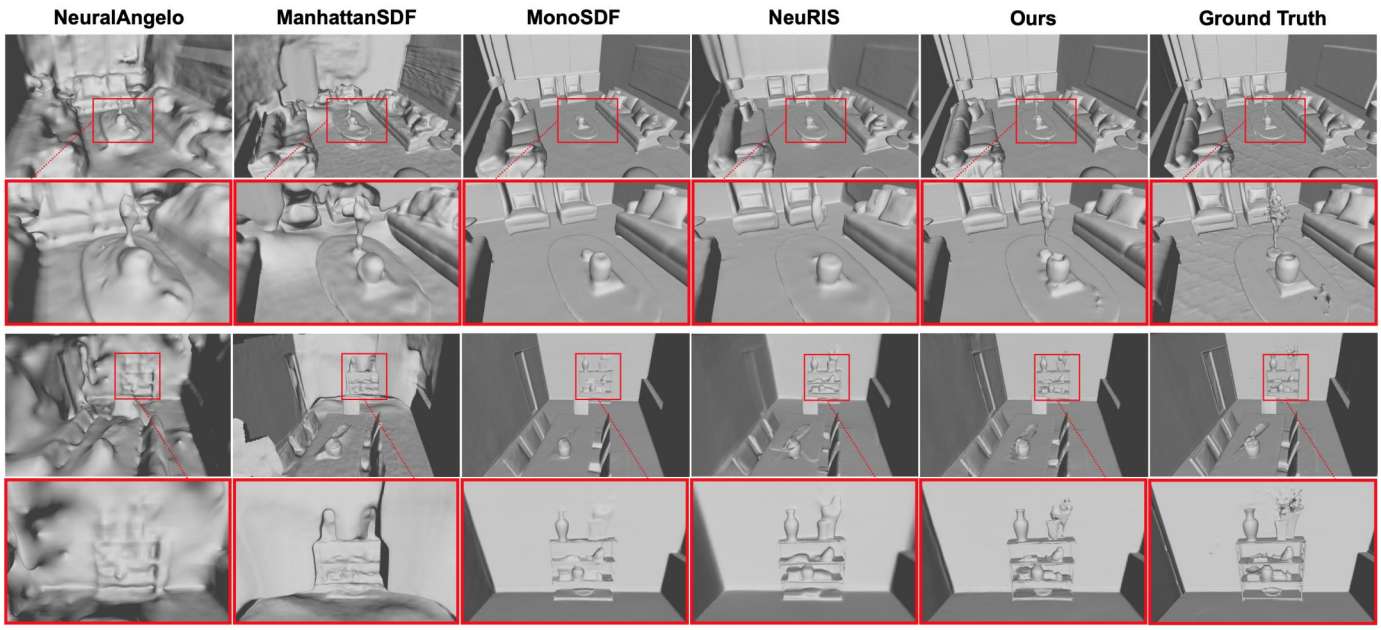


Fig. 7. Qualitative comparison of reconstructed meshes with other baselines on Replica dataset.

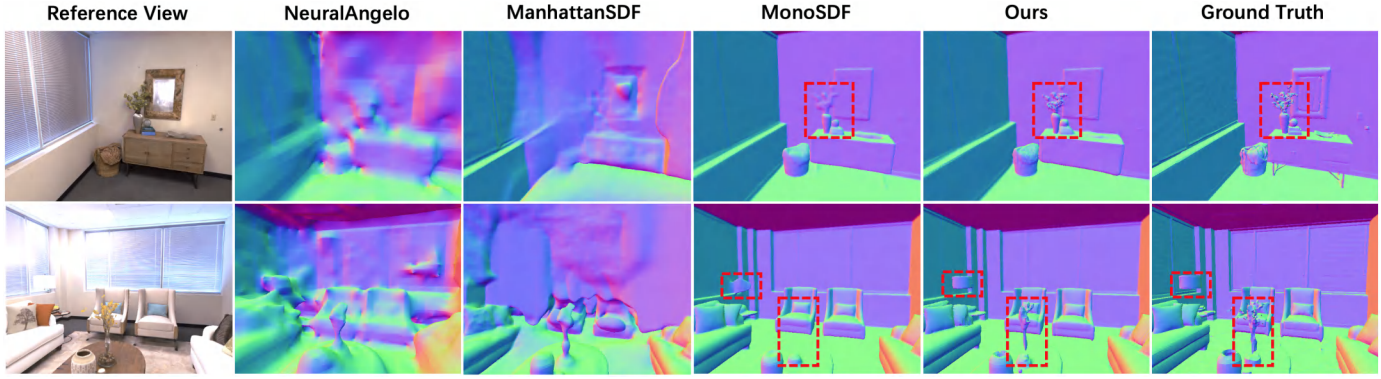


Fig. 8. Qualitative comparison of rendered normal maps with other baselines on Replica dataset.

our method with several baselines and report the averaged metrics. Although NeuS [4] and NeuralAngelo [26] can reconstruct high-fidelity object-level surfaces, their performance drops drastically in indoor scenarios. This is because the images captured from indoor scenes tend to be less idealized (artifacts are common) and overlap less between images, making the model hard to optimize. Aided by extra priors, methods like ManhattanSDF [6], NeuRIS [8], and MonoSDF [7] achieve performance improvement. However, the *Recall* metrics of these methods are usually low, indicating that the reconstruction results may lack fine-grained details. HelixSurf [35] attempts to incorporate MVS point clouds to produce more complete surfaces. Nevertheless, the MVS points can be noisy and degrade the *Precision* metric of the reconstruction meshes.

On both datasets, our proposed method surpasses existing methods on all five metrics. *F-score* is usually regarded as a faithful metric for evaluating geometry quality as it considers both accuracy and completeness. In particular, we achieve remarkable improvement on the *F-score* metric, which indicates that our approach can produce high-quality reconstructed meshes with fine and accurate details.

#### 4.2.2 *Qualitative Evaluation*

We visualize the meshes reconstructed by different methods in Figure 5 and Figure 7. For better comparison, we also render the normal maps of these meshes in Figure 6 and Figure 8. Conventional COLMAP [10], which is based on feature matching, struggles to reconstruct large texture-less indoor areas densely. NeuralAngelo [26] fails to generate smooth and consistent surfaces, possibly due to the discontinuity nature of the hash grid. Despite ManhattanSDF [6] introducing extra constraints based on Manhattan-world assumption, the precision of the reconstructed mesh is still limited. NeuRIS [8] and MonoSDF [7] leverage geometric priors estimated by off-the-shelf networks and adopt large MLPs to encode the entire scene. Because of the limited expression capacity of MLP and the inaccuracy that existed in monocular estimated priors, NeuRIS tends to produce artifacts in complex and intricate regions. Similarly, MonoSDF cannot faithfully reconstruct thin and fine-grained structures. HelixSurf can indeed generate more complete meshes,

TABLE 4
Ablation study of different geometry representations.

| Config | Setting | Prec↑ | Recall↑ | F-score↑ | Mem. | Param. |
|---|---|---|---|---|---|---|
| MLP only | layer=8 hidden=512 | 0.773 | 0.693 | 0.729 | 13GB | **2.3M** |
| Grids only | reso=256 channel=32 | 0.751 | 0.686 | 0.716 | 24GB | 537M |
| Tri-plane only | reso=512 channel=32 | 0.782 | 0.722 | 0.750 | **9GB** | 25.7M |
| MLP + Grids | reso=256 channel=16 hidden=256 | 0.783 | 0.711 | 0.745 | 18GB | 269M |
| MLP + Tri-plane | reso=256 channel=16 hidden=256 | **0.814** | **0.737** | **0.773** | 11GB | 4.5M |

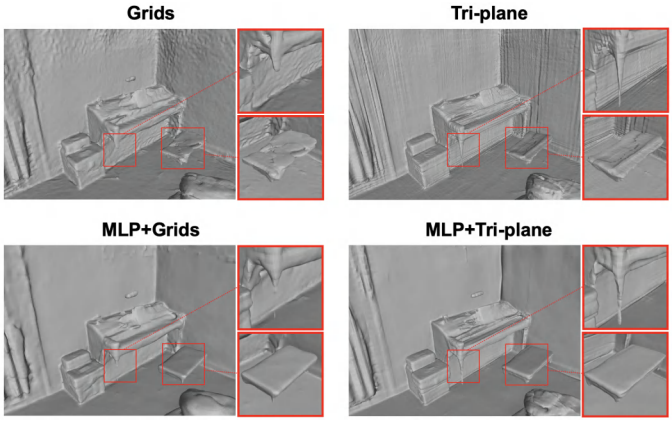


Fig. 9. Comparison of reconstructed meshes produced by different geometry representations.

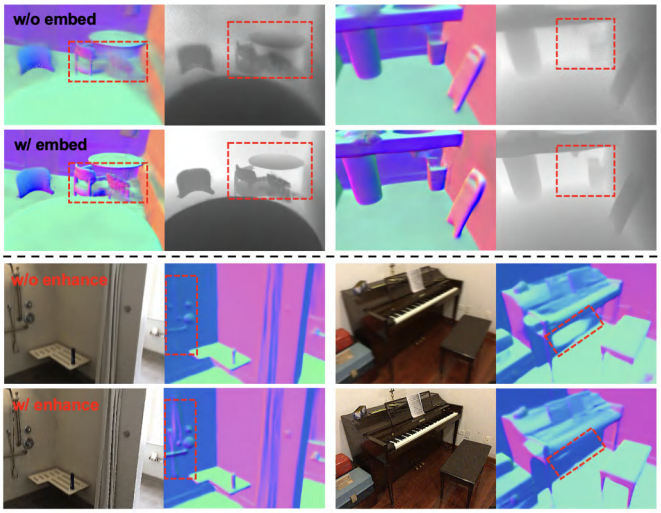


Fig. 10. We visualize the effects of appearance embeddings (top) and prior enhancement (bottom). Zoom in for better comparison.

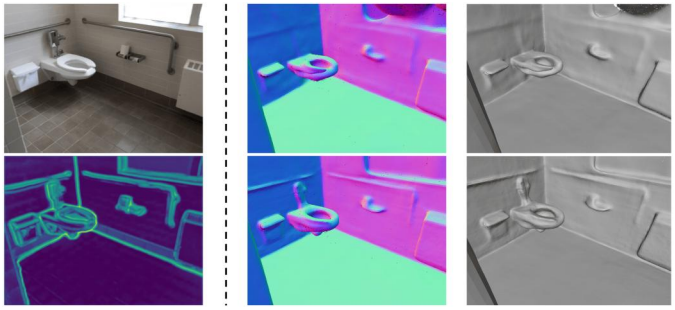

Fig. 11. We visualize the reconstruction results with / without the uncertainty estimation (bottom / top).

but at the expense of accuracy. As shown in Figure 5 and Figure 6, surfaces produced by HelixSurf are rough and blunt in corner and edge areas. Compared with baselines, our method is able to express both high-frequency details and low-frequency smoothness, and produces clean and visually appealing results.

### 4.3 Ablation Study

We conduct ablation experiments on ScanNet by separately changing or removing one of these components in our proposed method: (a) the hybrid geometry representation; (b) the appearance embeddings (w/o embed); (c) the prior enhancement (w/o enhance); (d) the uncertainty estimation module (w/o uncertainty, photometric uncertainty). Table 4 and Table 5 show the results of our ablation study.

#### *4.3.1 Analysis of Geometry Representation*

In Table 4, we compare the effectiveness of several geometry representations and report their parameters and memory consumption. MLP is a parameter compact representation but tends to generate over smooth surfaces, as discussed in Section 3.2. Although voxel grids can encode high-frequency details [7], they produce noisy and bumpy surfaces in planar regions (shown in Figure 9), which lead to the degradation of overall metrics. Also, the memory consumption and parameters of voxel grids increase significantly. In contrast, tri-plane is memory efficient and enables higher resolution to encode intricate structures. Metric improvements also indicate that tri-plane is a powerful geometry representation. Nevertheless, some axis-aligned striped artifacts still exist.

We find that using hybrid architecture (MLP+Grids, MLP+Tri-plane) can consistently achieve better results than using voxel grids or tri-plane alone, because it combines the advantage of both representations. Due to the superior expressive power of tri-plane, the hybrid representation of MLP and tri-plane surpasses the hybrid representation of MLP and voxel grids. Figure 9 also shows that MLP+Triplane is better at encoding thin and delicate structures than MLP+Grids, indicating the geometry representation we adopt is optimal. Note that for hybrid architecture,

TABLE 5
Ablation study of different components of our approach.

| Config | Prec↑ | Recall↑ | F-score↑ |
|---|---|---|---|
| w/o embed | 0.798 | 0.721 | 0.757 |
| w/o enhance | 0.774 | 0.719 | 0.745 |
| w/o uncertainty | 0.808 | 0.728 | 0.766 |
| photometric uncertainty | 0.810 | 0.729 | 0.768 |
| full model | **0.814** | **0.737** | **0.773** |

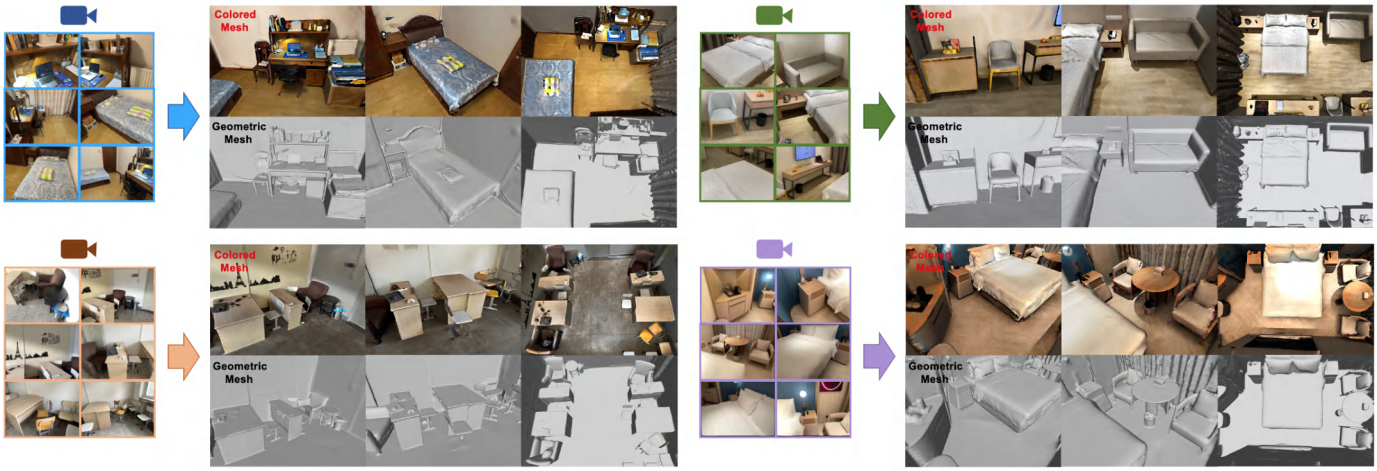


Fig. 12. Our method generalizes well to real-world indoor scenarios captured by hand-held mobile phones.

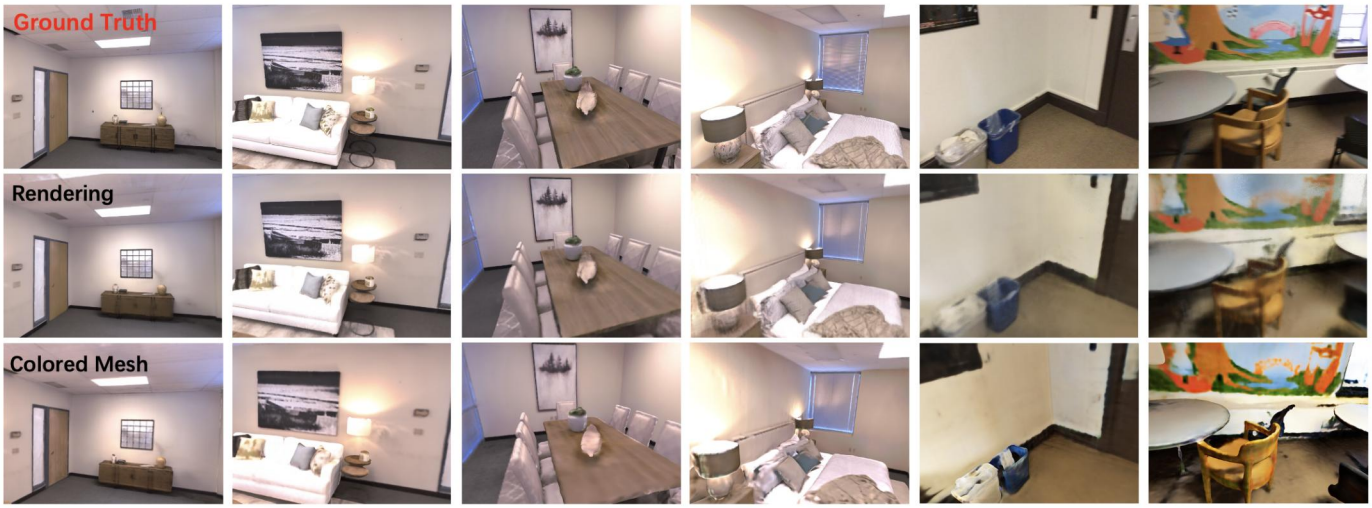


Fig. 13. Results of the novel view synthesis task. The first row is the ground-truth images, the second and third rows are novel views generated by volume rendering and rasterized by our reconstructed colored mesh, respectively.

we halve the hidden units of the MLP branch and feature channels of the grids/tri-plane branch, as we assume the learning burden of each branch is reduced. Therefore, performance improvements mainly come from the architectural design rather than the model size.

#### 4.3.2 Analysis of Appearance Embedding

Since we only supervise our model on 2D image observations, the inconsistency caused by photometric and environmental variations can lead to noisy geometries. Our experiments illustrate that the appearance embeddings can alleviate this problem, as the embeddings can learn to compensate for variations and guarantee a consistent reconstruction. We render the reconstructed normal maps and depth maps in Figure 10. The depth maps and normal maps are sharper and cleaner by adding appearance embeddings.

#### 4.3.3 Analysis of Prior Enhancement

As illustrated in Figure 10, imperfections like motion blur during the video capture process can cause degraded predictions of normal priors, which hinder the subsequent reconstruction of fine and accurate geometries. After applying our proposed enhancement, the image becomes sharper, and the predicted normal maps get better. As a result, Table 5 shows that the accuracy and completeness of reconstructed meshes are improved as well.

#### 4.3.4 Analysis of Uncertainty Estimation

The estimated uncertainty can serve as a measure of reliability and a balance between color supervision and normal supervision. In intricate areas, it guides our network to learn better geometry from rich color information rather than inaccurate normal information, thereby facilitating the correct reconstruction of exquisite structures. Ablation results show that the uncertainty estimation module helps to improve geometry quality. As fine-grained details take up only a small proportion of the scene and cannot be fully reflected by numerical metrics, we also visualize the reconstructed meshes and rendered normal maps in Figure 11. Adding uncertainty as guidance makes the reconstruction results more visually appealing and richer in high-frequency details. In addition, we find that supplementing

TABLE 6
Resource consumption compared to several baseline methods.

| Method | ManhattanSDF | MonoSDF | NeuralAngelo | HelixSurf | Ours |
|---|---|---|---|---|---|
| Param. | 1.1M | 0.8M | 366M | **0.2M** | 4.5M |
| Mem. | 12GB | 16GB | 21GB | 20GB | **11GB** |
| FLOPs | 45.3G | 36.5G | **34.9G** | 47.1G | 44.8G |

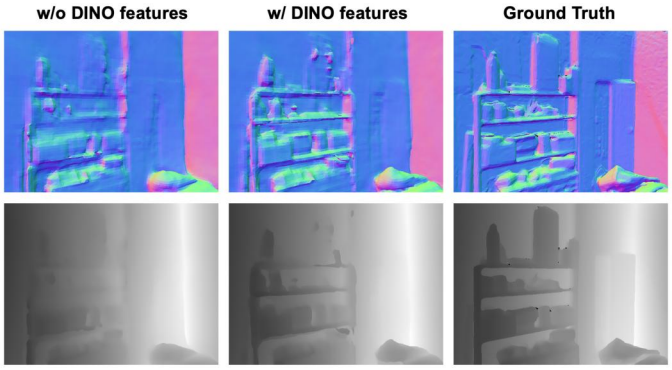


Fig. 14. High-level DINO features can further facilitate the reconstruction of intricate details. We visualize the normal and depth maps.

the uncertainty estimation module with high-level DINO features can further facilitate the reconstruction of complex and delicate structures (shown in Figure 14).

We have also tried to estimate uncertainty by evaluating the cross-view photometric consistency based on PatchMatch technique, whose performance is worse than using our proposed estimation module (see Table 5). We assume this is because the artifacts of captured images (jittering or lighting) can largely affect the cross-view photometric consistency, while our estimation module is trained on many data pairs and obtains a certain degree of robustness.

### 4.4 Real World Scenarios

In order to evaluate the generalization and practicality of our method, we also conduct experiments on real-world indoor scenarios captured by ourselves. Specifically, we walk around each scene and capture a two-minute video with our hand-held mobile phone. We extract 400-500 frames from this video and utilize OpenMVG [52] to estimate camera poses. Then, we use the proposed model to reconstruct 3D surfaces from the calibrated RGB images. Figure 12 shows that our approach can produce high-fidelity meshes under different types of indoor scenes, which shows the potential practical value of our method.

### 4.5 Novel View Synthesis

The novel view is another form of understanding 3D scenes apart from meshes. Figure 13 shows the novel views generated by volume rendering, rasterized by our reconstructed colored mesh, and the corresponding ground-truth images. Note that while our primary goal is to reconstruct high-fidelity, detailed surfaces of an indoor scenario, we can also synthesize realistic images under novel views facilitated by the accurate reconstruction results.

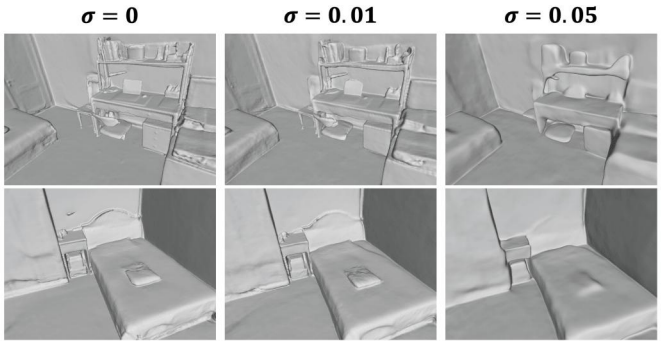


Fig. 15. Reconstructed meshes under noisy camera poses ($\sigma$ indicates the noise intensity). Note that our approach can still produce reasonable surfaces even though the camera poses are inaccurate.

### 4.6 Efficiency and Robustness

In Table 6, we compare the resource consumption of several baseline methods. ManhattanSDF [6] and MonoSDF [7] adopt the parameter compact MLP models but struggle to reconstruct fine geometries. The memory consumption and parameters of NeuralAngelo [26] are large because it utilizes 3D hash grids, which can be regarded as a special version of voxel grids. HelixSurf [35] represents the implicit SDF field with a relatively small MLP accompanied by a PatchMatch-based MVS module. However, the joint optimization of MVS module and neural radiance field in HelixSurf consumes a large memory overhead. In contrast, our method can produce the best quality meshes with comparable parameters and FLOPs, and less memory consumption.

In this work, we assume the camera poses in datasets are precise. However, the camera calibration process may introduce some errors under real circumstances. To evaluate the robustness of our method under extremely noisy poses, we add a random perturbation (between $[0, \sigma]$) to the camera pose of each view and conduct the 3D reconstruction. As illustrated in Figure 15, our method can still generate plausible geometry under noisy camera poses. The robustness is likely due to the uncertainty mechanism and normal priors, which provide the network with valuable cues and regularizations under imperfect poses.

## 5 Conclusion and Limitation

This paper proposes a novel framework to reconstruct indoor scenes with fine-grained details from posed RGB images. To enhance the expressive capability of the network, we design a hybrid geometry representation to encode low-frequency and high-frequency structures separately. This hybrid representation incorporates the advantages of MLP and tri-plane. Besides, we propose a sharpening and denoising enhancement as well as an uncertainty estimation module. The enhancement technique can facilitate the prediction of sharper and clearer normal priors, and the estimated uncertainty maps can prevent our model from being misled by unreliable priors in intricate regions. Quantitative and qualitative experiments show that our approach surpasses existing state-of-the-art methods. Our proposed model also generalizes well to real-world scenarios captured by hand-held mobile phones.

As our method is based on neural radiance fields with volume rendering, the proposed method fails when reconstructing mirrors or glasses. The reason is that the current volume rendering pipeline does not take into account physical reflections and refractions. This can possibly be solved by combining neural radiance fields with Whitted or Monte Carlo ray tracing, which we leave as future work. Furthermore, speeding up the training process for real-time reconstruction and extending our method to very large scenes are also interesting research directions.

## ACKNOWLEDGMENTS

This work was supported by Natural Science Foundation of China (62332019, U2336214, 62202257), and The Talent Fund of Beijing Jiaotong University (2023XKRC045).

## REFERENCES

[1] J. J. Park, P. R. Florence, J. Straub, R. A. Newcombe, and S. Lovegrove, "Deepsdf: Learning continuous signed distance functions for shape representation," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*. Computer Vision Foundation / IEEE, 2019, pp. 165–174.

[2] L. M. Mescheder, M. Oechsle, M. Niemeyer, S. Nowozin, and A. Geiger, "Occupancy networks: Learning 3d reconstruction in function space," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*. Computer Vision Foundation / IEEE, 2019, pp. 4460–4470.

[3] V. Sitzmann, J. N. P. Martel, A. W. Bergman, D. B. Lindell, and G. Wetzstein, "Implicit neural representations with periodic activation functions," in *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., 2020.

[4] P. Wang, L. Liu, Y. Liu, C. Theobalt, T. Komura, and W. Wang, "Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction," in *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, M. Ranzato, A. Beygelzimer, Y. N. Dauphin, P. Liang, and J. W. Vaughan, Eds., 2021, pp. 27 171–27 183.

[5] L. Yariv, J. Gu, Y. Kasten, and Y. Lipman, "Volume rendering of neural implicit surfaces," in *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, M. Ranzato, A. Beygelzimer, Y. N. Dauphin, P. Liang, and J. W. Vaughan, Eds., 2021, pp. 4805–4815.

[6] H. Guo, S. Peng, H. Lin, Q. Wang, G. Zhang, H. Bao, and X. Zhou, "Neural 3d scene reconstruction with the manhattan-world assumption," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*. IEEE, 2022, pp. 5501–5510.

[7] Z. Yu, S. Peng, M. Niemeyer, T. Sattler, and A. Geiger, "Monosdf: Exploring monocular geometric cues for neural implicit surface reconstruction," in *NeurIPS*, 2022.

[8] J. Wang, P. Wang, X. Long, C. Theobalt, T. Komura, L. Liu, and W. Wang, "Neuris: Neural reconstruction of indoor scenes using normal priors," in *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XXXII*, ser. Lecture Notes in Computer Science, S. Avidan, G. J. Brostow, M. Cissé, G. M. Farinella, and T. Hassner, Eds., vol. 13692. Springer, 2022, pp. 139–155.

[9] S. Shen, "Accurate multiple view 3d reconstruction using patch-based stereo for large-scale scenes," *IEEE Trans. Image Process.*, vol. 22, no. 5, pp. 1901–1914, 2013.

[10] J. L. Schönberger, E. Zheng, J. Frahm, and M. Pollefeys, "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*, ser. Lecture Notes in Computer Science, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds., vol. 9907. Springer, 2016, pp. 501–518.

[11] Z. Murez, T. van As, J. Bartolozzi, A. Sinha, V. Badrinarayanan, and A. Rabinovich, "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*, ser. Lecture Notes in Computer Science, A. Vedaldi, H. Bischof, T. Brox, and J. Frahm, Eds., vol. 12352. Springer, 2020, pp. 414–431.

[12] N. Stier, A. Rich, P. Sen, and T. Höllerer, "Vortx: Volumetric 3d reconstruction with transformers for voxelwise view selection and fusion," in *International Conference on 3D Vision, 3DV 2021, London, United Kingdom, December 1-3, 2021*. IEEE, 2021, pp. 320–330.

[13] G. Bae, I. Budvytis, and R. Cipolla, "Estimating and exploiting the aleatoric uncertainty in surface normal estimation," in *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*. IEEE, 2021, pp. 13 117–13 126.

[14] A. Eftekhar, A. Sax, J. Malik, and A. Zamir, "Omnidata: A scalable pipeline for making multi-task mid-level vision datasets from 3d scans," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 10 786–10 796.

[15] J. Wang, T. Bleja, and L. Agapito, "Go-surf: Neural feature grid optimization for fast, high-fidelity RGB-D surface reconstruction," in *International Conference on 3D Vision, 3DV 2022, Prague, Czech Republic, September 12-16, 2022*. IEEE, 2022, pp. 433–442.

[16] Z. Zhu, S. Peng, V. Larsson, W. Xu, H. Bao, Z. Cui, M. R. Oswald, and M. Pollefeys, "NICE-SLAM: neural implicit scalable encoding for SLAM," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*. IEEE, 2022, pp. 12 776–12 786.

[17] T. Wu, J. Wang, X. Pan, X. Xu, C. Theobalt, Z. Liu, and D. Lin, "Voxurf: Voxel-based efficient and accurate neural surface reconstruction," in *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023.

[18] E. R. Chan, C. Z. Lin, M. A. Chan, K. Nagano, B. Pan, S. D. Mello, O. Gallo, L. J. Guibas, J. Tremblay, S. Khamis, T. Karras, and G. Wetzstein, "Efficient geometry-aware 3d generative adversarial networks," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*. IEEE, 2022, pp. 16 102–16 112.

[19] T. Wang, B. Zhang, T. Zhang, S. Gu, J. Bao, T. Baltrusaitis, J. Shen, D. Chen, F. Wen, Q. Chen *et al.*, "Rodin: A generative model for sculpting 3d digital avatars using diffusion," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 4563–4573.

[20] M. Caron, H. Touvron, I. Misra, H. Jégou, J. Mairal, P. Bojanowski, and A. Joulin, "Emerging properties in self-supervised vision transformers," in *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*. IEEE, 2021, pp. 9630–9640.

[21] M. Niemeyer, L. Mescheder, M. Oechsle, and A. Geiger, "Differentiable volumetric rendering: Learning implicit 3d representations without 3d supervision," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 3504–3515.

[22] L. Yariv, Y. Kasten, D. Moran, M. Galun, M. Atzmon, R. Basri, and Y. Lipman, "Multiview neural surface reconstruction by disentangling geometry and appearance," in *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., 2020.

[23] M. Oechsle, S. Peng, and A. Geiger, "UNISURF: unifying neural implicit surfaces and radiance fields for multi-view reconstruction," in *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*. IEEE, 2021, pp. 5569–5579.

[24] C. Sun, M. Sun, and H. Chen, "Direct voxel grid optimization: Super-fast convergence for radiance fields reconstruction," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*. IEEE, 2022, pp. 5449–5459.

[25] T. Müller, A. Evans, C. Schied, and A. Keller, "Instant neural graphics primitives with a multiresolution hash encoding," *ACM Trans. Graph.*, vol. 41, no. 4, pp. 102:1–102:15, 2022.

[26] Z. Li, T. Müller, A. Evans, R. H. Taylor, M. Unberath, M.-Y. Liu, and C.-H. Lin, "Neuralangelo: High-fidelity neural surface recon-

struction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 8456–8465.

[27] A. Dai, A. X. Chang, M. Savva, M. Halber, T. A. Funkhouser, and M. Nießner, "Scannet: Richly-annotated 3d reconstructions of indoor scenes," in *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*. IEEE Computer Society, 2017, pp. 2432–2443.

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

[29] Y. Yao, Z. Luo, S. Li, T. Fang, and L. Quan, "Mvsnet: Depth inference for unstructured multi-view stereo," in *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part VIII*, ser. Lecture Notes in Computer Science, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds., vol. 11212. Springer, 2018, pp. 785–801.

[30] S. Im, H. Jeon, S. Lin, and I. S. Kweon, "Dpsnet: End-to-end deep plane sweep stereo," in *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019.

[31] M. Sayed, J. Gibson, J. Watson, V. Prisacariu, M. Firman, and C. Godard, "Simplerecon: 3d reconstruction without 3d convolutions," in *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XXXIII*, ser. Lecture Notes in Computer Science, S. Avidan, G. J. Brostow, M. Cissé, G. M. Farinella, and T. Hassner, Eds., vol. 13693. Springer, 2022, pp. 1–19.

[32] J. Sun, Y. Xie, L. Chen, X. Zhou, and H. Bao, "Neuralrecon: Real-time coherent 3d reconstruction from monocular video," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*. Computer Vision Foundation / IEEE, 2021, pp. 15 598–15 607.

[33] S.-S. Huang, H. Chen, J. Huang, H. Fu, and S.-M. Hu, "Real-time globally consistent 3d reconstruction with semantic priors," *IEEE Transactions on Visualization and Computer Graphics*, vol. 29, no. 4, pp. 1977–1991, 2021.

[34] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "Nerf: Representing scenes as neural radiance fields for view synthesis," in *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part I*, ser. Lecture Notes in Computer Science, A. Vedaldi, H. Bischof, T. Brox, and J. Frahm, Eds., vol. 12346. Springer, 2020, pp. 405–421.

[35] Z. Liang, Z. Huang, C. Ding, and K. Jia, "Helixsurf: A robust and efficient neural implicit surface learning of indoor scenes with iterative intertwined regularization," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 13 165–13 174.

[36] E. Ilg, O. Cicek, S. Galesso, A. Klein, O. Makansi, F. Hutter, and T. Brox, "Uncertainty estimates and multi-hypotheses networks for optical flow," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 652–667.

[37] C. Kerl, J. Sturm, and D. Cremers, "Dense visual slam for rgb-d cameras," in *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2013, pp. 2100–2106.

[38] F. Wang, S. Galliani, C. Vogel, and M. Pollefeys, "Itermvs: Iterative probability estimation for efficient multi-view stereo," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 8606–8615.

[39] Z. Teed and J. Deng, "Droid-slam: Deep visual slam for monocular, stereo, and rgb-d cameras," *Advances in neural information processing systems*, vol. 34, pp. 16 558–16 569, 2021.

[40] S. Sinha, S. Ebrahimi, and T. Darrell, "Variational adversarial active learning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 5972–5981.

[41] R. Raguram, J.-M. Frahm, and M. Pollefeys, "Exploiting uncertainty in random sample consensus," in *2009 IEEE 12th International Conference on Computer Vision*. IEEE, 2009, pp. 2074–2081.

[42] J. Shen, A. Ruiz, A. Agudo, and F. Moreno-Noguer, "Stochastic neural radiance fields: Quantifying uncertainty in implicit 3d representations," in *International Conference on 3D Vision, 3DV 2021, London, United Kingdom, December 1-3, 2021*. IEEE, 2021, pp. 972–981.

[43] J. Shen, A. Agudo, F. Moreno-Noguer, and A. Ruiz, "Conditional-flow nerf: Accurate 3d modelling with reliable uncertainty quantification," in *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part III*, ser. Lecture Notes in Computer Science, S. Avidan, G. J. Brostow, M. Cissé, G. M. Farinella, and T. Hassner, Eds., vol. 13663. Springer, 2022, pp. 540–557.

[44] M. Atzmon and Y. Lipman, "Sal: Sign agnostic learning of shapes from raw data," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 2565–2574.

[45] R. Martin-Brualla, N. Radwan, M. S. M. Sajjadi, J. T. Barron, A. Dosovitskiy, and D. Duckworth, "Nerf in the wild: Neural radiance fields for unconstrained photo collections," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*. Computer Vision Foundation / IEEE, 2021, pp. 7210–7219.

[46] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*. Springer, 2015, pp. 234–241.

[47] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.

[48] T. Darcet, M. Oquab, J. Mairal, and P. Bojanowski, "Vision transformers need registers," *arXiv preprint arXiv:2309.16588*, 2023.

[49] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," in *International conference on machine learning*. PMLR, 2021, pp. 10 347–10 357.

[50] A. Gropp, L. Yariv, N. Haim, M. Atzmon, and Y. Lipman, "Implicit geometric regularization for learning shapes," in *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, ser. Proceedings of Machine Learning Research, vol. 119. PMLR, 2020, pp. 3789–3799.

[51] A. Dai, M. Nießner, M. Zollhöfer, S. Izadi, and C. Theobalt, "Bundlefusion: Real-time globally consistent 3d reconstruction using on-the-fly surface reintegration," *ACM Transactions on Graphics (ToG)*, vol. 36, no. 4, p. 1, 2017.

[52] P. Moulon, P. Monasse, R. Perrot, and R. Marlet, "Openmvg: Open multiple view geometry," in *Reproducible Research in Pattern Recognition - First International Workshop, RRPR@ICPR 2016, Cancún, Mexico, December 4, 2016, Revised Selected Papers*, ser. Lecture Notes in Computer Science, B. Kerautret, M. Colom, and P. Monasse, Eds., vol. 10214, 2016, pp. 60–74.

.

**Sheng Ye** is a Ph.D. student with the Department of Computer Science and Technology, Tsinghua University, China. He received his BEng degree from Tsinghua University, China in 2022. His research interests include computer vision and 3D reconstruction.

**Yubin Hu** is a Ph.D. student with the Department of Computer Science and Technology, Tsinghua University, China. He received his BEng degree from the Electronic Engineering Department, Tsinghua University, China in 2020. He was a Visiting Undergraduate Research Intern of the Harvard John A. Paulson School of Engineering and Applied Sciences in 2019. His research interests include computer vision and 3D reconstruction.

**Matthieu Lin** is a Ph.D. student with the Department of Computer Science and Technology at Tsinghua University, under the supervision of Professor Yong-Jin Liu. He received his B.S.E degree in Computer Science from ESIEA Paris in 2018 and his M.S. degree in Computer Science from Tsinghua University in 2021. His research interests include reinforcement learning and computer vision.

**Yu-Hui Wen** is an associate professor with the Beijing Key Laboratory of Traffic Data Analysis and Mining, School of Computer and Information Technology, Beijing Jiaotong University, Beijing, China. She received her bachelor's degree from Harbin Institute of Technology (HIT), and the Ph.D. degree in computer science and technology from University of Chinese Academy of Sciences (UCAS), Beijing, China, in 2020. Her research interests include machine vision, computer graphics and human motion analysis.

**Wang Zhao** is a Ph.D. student with the Department of Computer Science and Technology, Tsinghua University, China. He received his BEng degree from the Electronic Engineering Department, Tsinghua University, China in 2019. His research interests include 3D computer vision.

**Yong-Jin Liu** is a Professor with the Department of Computer Science and Technology, Tsinghua University, China. He received the BEng degree from Tianjin University, China, in 1998, and the PhD degree from the Hong Kong University of Science and Technology, Hong Kong, China, in 2004. His research interests include affective computing, computer graphics and computer vision. He is a senior member of IEEE and ACM. For more information, visit https://cg.cs.tsinghua.edu.cn/people/~Yongjin/Yongjin.html

**Wenping Wang** (Fellow, IEEE) received the Ph.D. degree in computer science from the University of Alberta in 1992. He is a Professor of computer science at Texas A&M University. His research interests include computer graphics, computer visualization, computer vision, robotics, medical image processing, and geometric computing. He is or has been an journal associate editor of ACM Transactions on Graphics, IEEE Transactions on Visualization and Computer Graphics, Computer Aided Geometric Design, and Computer Graphics Forum (CGF). He has chaired a number of international conferences, including Pacific Graphics, ACM Symposium on Physical and Solid Modeling (SPM), SIGGRAPH and SIGGRAPH Asia. Prof. Wang received the John Gregory Memorial Award for his contributions to geometric modeling. He is an IEEE Fellow and an ACM Fellow.