NC-SDF: Enhancing Indoor Scene Reconstruction Using Neural SDFs with View-Dependent Normal Compensation

Ziyi Chen[1] Xiaolong Wu[2] Yu Zhang[1] *
[1]Zhejiang University [2]Amap Alibaba

Abstract

State-of-the-art neural implicit surface representations have achieved impressive results in indoor scene reconstruction by incorporating monocular geometric priors as additional supervision. However, we have observed that multi-view inconsistency between such priors poses a challenge for high-quality reconstructions. In response, we present NC-SDF, a neural signed distance field (SDF) 3D reconstruction framework with view-dependent normal compensation (NC). Specifically, we integrate view-dependent biases in monocular normal priors into the neural implicit representation of the scene. By adaptively learning and correcting the biases, our NC-SDF effectively mitigates the adverse impact of inconsistent supervision, enhancing both the global consistency and local details in the reconstructions. To further refine the details, we introduce an informative pixel sampling strategy to pay more attention to intricate geometry with higher information content. Additionally, we design a hybrid geometry modeling approach to improve the neural implicit representation. Experiments on synthetic and real-world datasets demonstrate that NC-SDF outperforms existing approaches in terms of reconstruction quality.

1. Introduction

3D scene reconstruction from multi-view images is a long-term challenge in computer vision, with applications across various domains such as virtual reality, robotics, and autonomous driving. Multi-view stereo (MVS) techniques [4, 6, 19, 23, 36, 48] recover depth maps for each view by matching features between adjacent views, subsequently fusing the depth maps to reconstruct 3D geometry. Their reconstructions tend to be noisy, especially in texture-less areas. Some data-driven methods [7, 13, 28, 39, 41] alleviate this limitation by directly predicting a truncated signed distance field (TSDF) from multi-view images. However, they necessitate expensive 3D supervision and produce over-
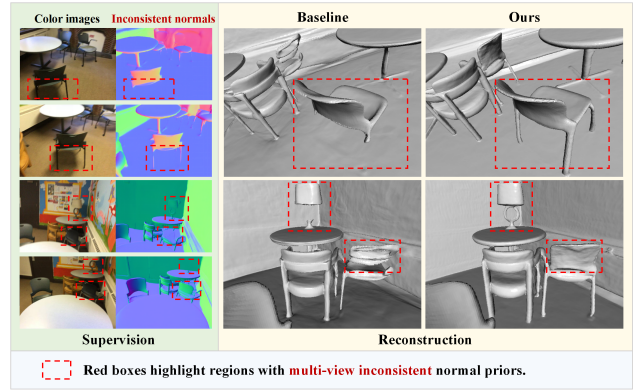

Figure 1. **Comparison between baseline and NC-SDF.** State-of-the-art neural implicit surface representations produce suboptimal reconstructions with noisy or missing surfaces, primarily due to multi-view inconsistency between monocular geometric priors. Our NC-SDF introduces a view-dependent normal compensation model to adaptively learn and correct biases in normal priors. This approach enables the recovery of intricate geometric details while ensuring smoothness in texture-less areas within reconstructions.

smooth results.

Recently, impressive progress has been made in neural implicit surface reconstruction combined with volume rendering techniques. They utilize multi-layer perceptrons (MLPs) to parameterize implicit shape representations, such as occupancy [29] or signed distance fields (SDFs) [45, 50]. Though these methods excel in capturing continuous and smooth surfaces, they face challenges when dealing with indoor scenes containing large texture-less regions. The primary reason is that multi-view photometric consistency fails to provide sufficient constraints in such regions, such as walls and floors.

Recent advancements have mitigated this problem by incorporating additional priors for supervision, including sensor depths [3, 30, 35, 43, 54, 56], semantic priors [16], depth priors from MVS methods [22, 34, 47] and monocular geometric priors [11, 44, 52, 55]. Among these works, [44, 52] have produced state-of-the-art results by utilizing geometric cues from monocular geometry estimation networks [12, 32, 33, 46, 51]. However, their performance is

---

*Corresponding author

heavily dependent on the quality of the geometric predictions. Notably, the estimation networks inevitably introduce biases between the predictions and the ground truths (GTs). Moreover, these biases are related to the viewing direction, since the networks receive input from a single view rather than multiple views. Therefore, such geometric predictions always struggle to satisfy multi-view consistency.

In this work, we present NC-SDF, a neural SDF 3D reconstruction framework with view-dependent normal compensation. The framework is designed to enhance indoor scene reconstruction by addressing multi-view inconsistency between monocular normal priors. To achieve this, We model not only the scene's radiance field and SDF but also the view-dependent biases in normal priors. Through adaptive compensation for the biases at corresponding viewing directions, our NC-SDF enables more consistent supervision and eventually leads to better performance. For more detailed reconstruction, we design an informative pixel sampling strategy to pay more attention to intricate geometry, by prioritizing sampling pixels with higher information content. Recognizing the limited representation power of MLPs, we introduce a hybrid geometry modeling approach based on feature fusion. This approach utilizes the inductive smoothness bias of MLPs to ensure smooth surfaces, and harnesses the high-frequency encodings provided by voxel grids to capture intricate geometry.

In summary, our NC-SDF significantly enhances the reconstruction quality. The combination of our three designs ensures consistent and smooth surfaces while enabling sharp details in the reconstructions. Our contributions can be summarized as follows:

- The view-dependent normal compensation model results in globally consistent and locally detailed reconstructions through adaptive compensation for the view-dependent normal biases.
- The informative pixel sampling strategy and hybrid geometry model further enhance the reconstruction of geometric details.
- Comprehensive experiments on both synthetic and real-world datasets demonstrate that our NC-SDF achieves state-of-the-art indoor scene reconstruction.

2. Related work

2.1. Multi-view surface reconstruction

Traditional MVS methods [4, 6, 23, 36, 37] take multi-view images as inputs, and utilize feature matching and triangulation methods to estimate depth maps for each view. While excelling in reconstructing textured regions, they face difficulties when dealing with texture-less regions. With the development of deep neural networks, several learning-based MVS works [19, 20, 25, 48, 49] utilize convolutional neural networks (CNNs) to learn the mapping from multi-view

images to depth maps. However, the depth maps suffer from scale ambiguity and multi-view inconsistency. Alternatively, other learning-based methods [7, 13, 28, 39, 41] directly predict the TSDF and then extract the mesh from the TSDF volume. These methods produce more consistent reconstructions, but they demand a large amount of ground truth 3D data for training, and the results often lack details due to the limitation of the TSDF resolution.

2.2. Neural implicit surface reconstruction

Neural implicit functions have attracted increasing attention, owing to their advantages of compactness and low memory consumption. DeepSDF [31] proposes to model the SDF of target objects with an MLP, achieving the reconstruction of complex shapes. Neural radiance fields (NeRF) [26] and its variations [5, 14, 27, 40] utilize MLPs to implicitly encode the volume density and view-dependent emitted radiance field of the scene, enabling novel view synthesis. They optimize networks with only color images as constraints, by employing volume rendering techniques. Although iso-surfaces based on volume density can be extracted to recover the 3D geometry of the scene, the resulting mesh often exhibits noticeable noise due to the limited constraints on level sets.

To reconstruct smoother surfaces, several methods [45, 50] suggest using SDF as the output of neural implicit functions and reparameterizing SDF as volume density. This improvement results in superior surface reconstruction while preserving the capacity for novel view synthesis. However, due to the inductive smoothness bias of MLPs, using MLPs alone for scene modeling may result in over-smooth surfaces with limited details. Recent works [27, 43, 52, 56] combine voxel grids with a shallow MLP decoder to improve the representation power of the model. Despite improving the reconstruction of details, this approach introduces noise into the results because of under-constrained voxel grids.

2.3. Prior-guided neural implicit surface reconstruction for indoor scenes

Constraints provided by color images are often insufficient for generating high-quality reconstructions when employing neural implicit representations in indoor scenes. This limitation primarily arises from the presence of large texture-less regions. Recent studies have introduced different kinds of priors as additional supervision to overcome the limitation. GO-Surf [43] incorporates range measurements from depth cameras. Nerfingmvs [47] uses sparse depth information from Structure from Motion (SfM) to mitigate shape blurriness. ManhattanSDF [16] assumes that the normals of walls and floors adhere to the Manhattan-world assumption. HelixSurf [22] utilizes MVS results to improve reconstruction quality and optimization time. NeuRIS [44]
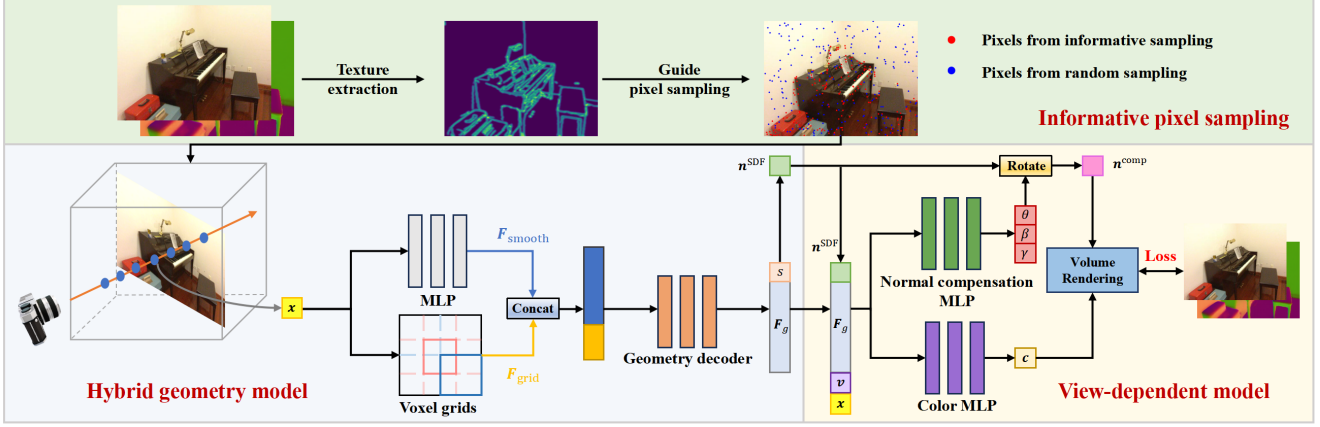
Figure 2. **Pipeline of NC-SDF.** We model the geometry field (SDF), view-dependent radiance field, and view-dependent normal biases with neural implicit functions. Besides, we propose an informative pixel sampling strategy and a hybrid geometry model to further improve the reconstruction of thin geometry.

and MonoSDF [52] explore monocular geometric cues from pretrained networks [12, 32, 33, 46, 51]. While they achieve state-of-the-art reconstruction results, their performance notably degrades in areas with inconsistent and noisy geometric priors. NeuRIS [44] filters out unreliable normal priors by checking multi-view photometric consistency during training, but the handcrafted strategy is vulnerable to noise in real-world datasets.

3. Method

Given multi-view images with known poses, our goal is to produce high-quality 3D reconstructions. We begin with an introduction to our core framework and volume rendering technique (Sec. 3.1). We then delve into our view-dependent normal compensation model (Sec. 3.2), informative pixel sampling strategy (Sec. 3.3), and hybrid geometry model (Sec. 3.4). Finally, we provide details of loss functions (Sec. 3.5). Fig. 2 illustrates the pipeline of our NC-SDF.

3.1. Preliminary

Our neural implicit representations consist of three components: the geometry model f_g, the color model f_c, and the normal compensation model f_n. In brief, the geometry model f_g encodes the SDF, the color model f_c encodes the view-dependent radiance field, and the normal compensation model f_n encodes the view-dependent biases in monocular normal priors. We employ MLPs for the color and normal compensation modeling. And we utilize our proposed hybrid geometry model for the geometry modeling.

We adopt the differentiable volume rendering technique, following NeuS [45]. A ray emitting from the camera center $\boldsymbol{o}$ can be expressed as $\boldsymbol{r} = \boldsymbol{o} + t\boldsymbol{v}$, where $\boldsymbol{v}$ represents the viewing direction of the ray. Along the ray, we sample N

points. For each 3D point $\boldsymbol{x}_i$, the geometry model f_g maps it to a signed distance s_i and geometry feature $\boldsymbol{F}_g$:

$$s_i, \boldsymbol{F}_g = f_g(\boldsymbol{x}_i). \tag{1}$$

The color model f_c outputs the radiance $\boldsymbol{c}_i$ observed from the viewing direction $\boldsymbol{v}$:

$$\boldsymbol{c}_i = f_c(\boldsymbol{x}_i, \boldsymbol{v}, \boldsymbol{n}_i^{\mathrm{SDF}}, \boldsymbol{F}_g), \tag{2}$$

where the normal $\boldsymbol{n}_i^{\mathrm{SDF}}$ is the gradient of the signed distance s_i. In order to avoid confusion with other normals in subsequent discussions, we refer to this normal as the SDF normal. Then the color is accumulated along the ray:

$$\hat{\mathbf{C}} = \sum_{i=1}^{N} T_i \alpha_i \boldsymbol{c}_i, \tag{3}$$

where $T_i = \prod_{j=1}^{i-1}(1 - \alpha_j)$ and α_i is the opaque density which can be further expressed as follows:

$$\alpha_i = \max\left(\frac{\Phi_\tau(s_i) - \Phi_\tau(s_{i+1})}{\Phi_\tau(s_i)}, 0\right), \ \Phi_\tau(x) = (1+e^{-\tau x})^{-1}.$$

The normal compensation model f_n predicts compensation rotation angles γ, β, θ about the x, y, z axes from the viewing direction $\boldsymbol{v}$:

$$\gamma, \beta, \theta = f_n(\boldsymbol{x}_i, \boldsymbol{v}, \boldsymbol{n}_i^{\mathrm{SDF}}, \boldsymbol{F}_g). \tag{4}$$

By applying the compensation rotation to the SDF normal $\boldsymbol{n}_i^{\mathrm{SDF}}$, we obtain the compensated normal $\boldsymbol{n}_i^{\mathrm{comp}}$. Further details are elaborated in Sec. 3.2. We apply the same volume rendering technique to generate the rendered compensated normal map:

$$\mathbf{N}^{\mathrm{comp}} = \sum_{i=1}^{N} T_i \alpha_i \boldsymbol{n}_i^{\mathrm{comp}}. \tag{5}$$

We optimize the neural networks by minimizing the difference between the rendered outputs and reference inputs. The surface can be extracted as the zero level-set of the SDF using the Marching Cubes algorithm [24].

3.2. View-dependent normal compensation model

As mentioned before, multi-view inconsistent priors can significantly impact reconstruction quality. [34] utilizes per-view uncertainty maps from the prediction network to alleviate this problem, but these maps also fall short of satisfying multi-view consistency. Though NeuRIS [44] employs a geometric consistency checking strategy to filter out unreliable priors, the strategy requires manual threshold setting and lacks robustness. Inspired by volume rendering techniques [26], we propose integrating view-dependent biases in monocular normal priors into the implicit representation of the scene. Unlike previous methods, our approach compensates for the view-dependent biases in an adaptive way, without any additional manual operation.

Normal compensation. The compensation process is as follows: we concatenate the spatial position of the point x, its viewing direction v, the SDF normal n^{SDF}, and the geometry feature F_g. This concatenated feature is then fed into our normal compensation model f_n, which outputs compensation rotation angles γ, β, and θ corresponding to the x, y, and z axes, as described in Eq. (4). Following Eq. (6), we rotate the SDF normal n^{SDF} first by an angle γ around the x-axis, then by an angle β around the y-axis, and finally by an angle θ around the z-axis to obtain the compensated normal n^{comp}.

$$n^{\text{comp}} = \mathbf{R}_{\text{ZYX}} n^{\text{SDF}} = \mathbf{R}_{\text{Z}}(\theta)\mathbf{R}_{\text{Y}}(\beta)\mathbf{R}_{\text{X}}(\gamma)n^{\text{SDF}}, \quad (6)$$

where $\mathbf{R}$ represents the corresponding rotation matrix, and the calculation is described in the supplementary material. We use the rendered compensated normal maps $\mathbf{N}^{\text{comp}}$ to align with the noisy normal priors, instead of directly supervising the rendered SDF normal maps $\mathbf{N}^{\text{SDF}}$.

Two-stage training. To stabilize the training, we design a two-stage training strategy. In the first stage, we optimize the color and geometry models to obtain a well-initialized radiance field and SDF. Previous studies [2, 53] have indicated that neural networks tend to memorize clean and easy patterns in the early stages of training. Therefore, our implicit functions can easily learn the distribution of regions with consistent supervision in the first stage. If training continues with inconsistent supervision, the neural network may eventually overfit on the noisy normals, leading to suboptimal reconstructions. To resolve this issue, we introduce the normal compensation model in the second stage and optimize it concurrently with the color and geometry models. By explicitly modeling the biases, our NC-SDF achieves more robust training under noisy supervision signals.

(a) Without our normal compensation model.

(b) With our normal compensation model. ⓒ represents the normal compensation process.
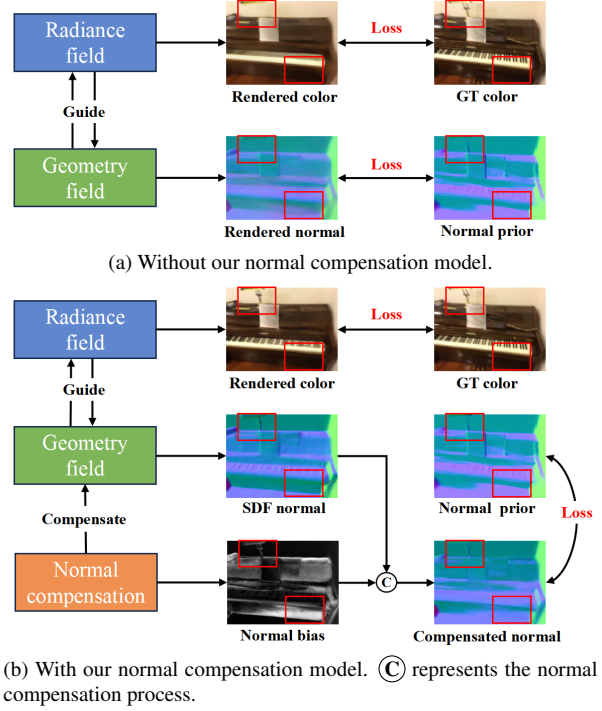
Figure 3. **Visualization of rendered results**, comparing (a) without and (b) with our normal compensation model.

Fig. 3b visually explains the working principle of the normal compensation (NC) model. The NC model's optimization direction is constrained by both RGB images and normal priors, rather than being arbitrary. Despite the noise in the normal priors, RGB images provide reliable supervision. In regions where normal priors exhibit multi-view inconsistency, color constraints play an essential role in facilitating the optimization of the geometry field. In summary, the coupled optimization relationship between the radiance field and the geometry field enables the NC model to disentangle normal priors reasonably.

3.3. Informative pixel sampling strategy

In images of indoor scenes, texture-less regions usually occupy a significant portion of pixels, while geometric details are limited to a small portion. As previous research [18, 21, 42] has pointed out, neural models exhibit biases towards the majority classes when trained on imbalanced datasets. Consequently, these models perform poorly on the minority classes. As is evident in neural implicit scene reconstruction, recovering intricate geometry proves to be more challenging than reconstructing plane surfaces. To address this issue, we propose an informative pixel sampling strategy as an alternative to the random pixel sampling method used in previous works [44, 45, 50, 52].

Texture extraction. Image regions containing fine geometry often possess high information content, characterized by strong contrast and rich textures. We utilize the Canny
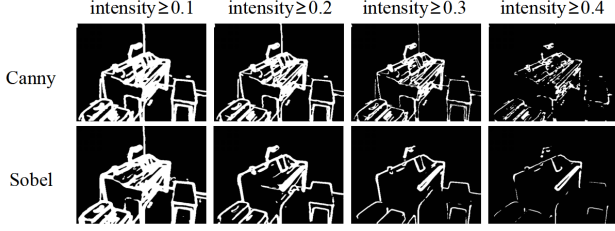
intensity≥0.1 intensity≥0.2 intensity≥0.3 intensity≥0.4

Canny

Sobel

Figure 4. **Visualization of texture intensity maps at different intensity thresholds**. The Canny operator exhibits a more robust performance than the Sobel operator.

edge detection operator [8, 10] to capture these high-texture regions. The Canny operator is well-known for its robustness in extracting structural and textural information from visual objects. Compared to other gradient-based operators such as Sobel and Prewitt [1, 38], the Canny operator provides superior texture localization, noise reduction, and adaptability. In practice, we extract texture intensity maps from each image. The comparison in Fig. 4 demonstrates that the Canny operator outperforms the Sobel operator in terms of robust texture extraction.

Pixel sampling strategy. Our informative pixel sampling strategy evolves during training, adhering to a coarse-to-fine manner. We sample N_{sample} pixels per batch. The pixel sampling is divided into two parts: a proportion r is allocated to informative pixel sampling, while the remaining $1 - r$ is allocated to random sampling. We set an intensity threshold l_i for the extracted texture maps. From the set $\{l \mid l \geq l_i\}$, we randomly sample $r * N_{\text{sample}}$ pixels to create a high-information pixel set $\mathcal{P}_{\text{canny}}$. At the same time, we randomly sample $(1-r) * N_{\text{sample}}$ pixels from all the pixels, forming the set $\mathcal{P}_{\text{random}}$. Finally, the pixel set for each batch is expressed as $\mathcal{P}_{\text{all}} = \mathcal{P}_{\text{canny}} \cup \mathcal{P}_{\text{random}}$.

The proportion r and intensity threshold l_i evolve during training. Initially, we randomly sample pixels to establish a reliable initialization for the entire scene. As training progresses, our sampling strategy gradually transitions towards informative sampling, placing greater emphasis on intricate geometry. This coarse-to-fine approach enables more detailed reconstructions.

3.4. Hybrid geometry model

Using MLPs alone for scene modeling is limited in capturing intricate details. Voxel grids provide improved expressive power, but their high dimensionality can introduce noise into reconstructions. To resolve this, we propose a hybrid geometry model based on feature fusion that combines the strengths of both MLPs and voxel grids models. The hybrid model comprises two feature branches: the MLP branch provides a smooth feature $\boldsymbol{F}_{\text{smooth}}$, encoding low-frequency structures, while the voxel grids branch offers a grid feature $\boldsymbol{F}_{\text{grid}}$, encoding high-frequency structures. We then concatenate $\boldsymbol{F}_{\text{smooth}}$ and $\boldsymbol{F}_{\text{grid}}$ to obtain a feature

that encodes both low-frequency and high-frequency structures. The combined feature is subsequently decoded using a shallow MLP to obtain a geometry feature $\boldsymbol{F}_g$ and signed distance s_i. Our hybrid geometry model overcomes the limitations of using MLPs alone or voxel grids alone for scene modeling, striking a balance in modeling both intricate structures and planar regions.

3.5. Loss functions

We use color images and normal priors for supervision. Furthermore, the gradients of the SDF satisfy the Eikonal equation [15]. We represent a set of camera rays passing through pixels as $\mathcal{R}$, color images as $\mathbf{C}$, and normal priors as $\mathbf{N}$. Our overall loss function is defined as follows:

$$\mathcal{L} = \mathcal{L}_c + \lambda_n \mathcal{L}_n + \lambda_e \mathcal{L}_e, \tag{7}$$

where

$$\mathcal{L}_c = \sum_{\boldsymbol{r} \in \mathcal{R}} \|\hat{\mathbf{C}}(\boldsymbol{r}) - \mathbf{C}(\boldsymbol{r})\|_1, \tag{8}$$

$$\mathcal{L}_n = \sum_{\boldsymbol{r} \in \mathcal{R}} \|\mathbf{N}^{\text{comp}}(\boldsymbol{r}) - \mathbf{N}(\boldsymbol{r})\|_1 + \left\|1 - \mathbf{N}^{\text{comp}}(\boldsymbol{r})^{\mathrm{T}} \mathbf{N}(\boldsymbol{r})\right\|_1, \tag{9}$$

$$\mathcal{L}_e = \frac{1}{N} \sum_{\boldsymbol{r} \in \mathcal{R}} \sum_{i=1}^{N} (\|\nabla s_i\|_2 - 1)^2. \tag{10}$$

4. Experiments

4.1. Experiment Setup

Datasets. We choose two different indoor scene datasets for experiments, including ScanNet [9] and ICL-NUIM [17]. These two datasets provide RGB-D images with camera poses. ScanNet is a real-world dataset and ICL-NUIM is a synthetic dataset. For comparison experiments, We select four scenes from each of the two datasets. For ablation studies, We use the four scenes from ScanNet.

Baselines. We compare our method with the following baselines: (1) Traditional MVS method COLMAP [36]; (2) Neural implicit representation methods without additional supervision, including VolSDF [50] and NeuS [45]; (3) Neural implicit representation methods with additional supervision, including ManhattanSDF [16], HelixSurf [22], NeuRIS [44], and MonoSDF (both MLP and voxel grids version) [52].

Evaluation metrics. Following [28], we evaluate the reconstruction results using accuracy, completeness, precision, recall, and F-score. The F-score is generally considered the most comprehensive indicator.

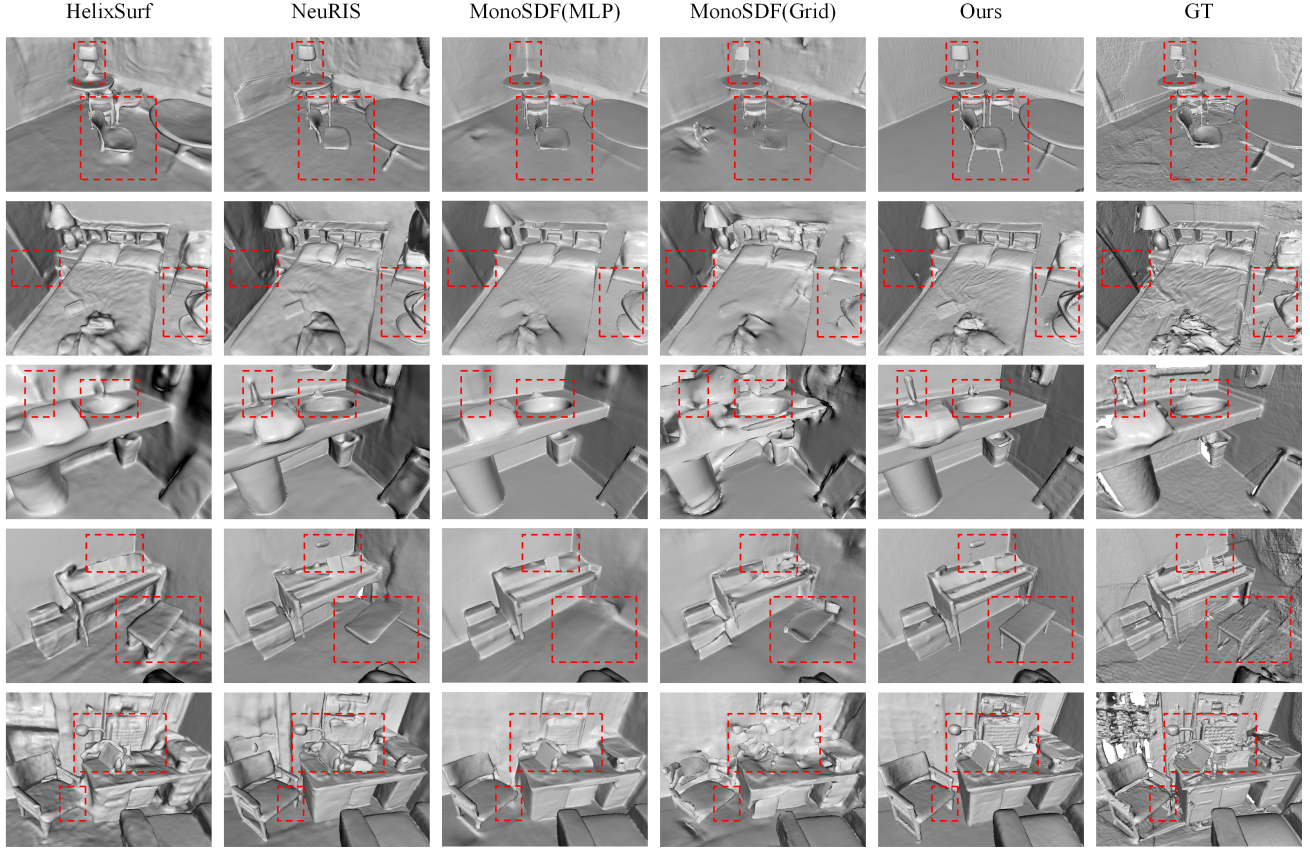**Implementation.** All the experiments are conducted on one NVIDIA RTX 3090 GPU. We employ the OmniData

Figure 5. Qualitative results on ScanNet [9].

Table 1. Quantitative results on ScanNet [9].

Method	Acc↓	Comp↓	Prec↑	Recall↑	F-score↑
COLMAP [36]	0.041	0.231	0.755	0.438	0.548
VolSDF [50]	0.086	0.129	0.470	0.399	0.430
NeuS [45]	0.143	0.208	0.380	0.277	0.320
ManhattanSDF [16]	0.044	0.055	0.749	0.668	0.706
HelixSurf [22]	0.036	0.042	0.791	0.725	0.756
NeuRIS [44]	0.051	0.050	0.709	0.662	0.684
MonoSDF(MLP) [52]	0.036	0.045	0.795	0.708	0.748
MonoSDF(Grid) [52]	0.048	0.050	0.727	0.663	0.693
Ours	**0.033**	**0.039**	**0.811**	**0.754**	**0.781**

Table 2. Quantitative results on ICL-NUIM [17].

Method	Acc↓	Comp↓	Prec↑	Recall↑	F-score↑
NeuRIS	0.043	0.155	0.734	0.556	0.626
MonoSDF(MLP)	0.043	0.143	0.741	0.589	0.650
Ours	**0.021**	**0.105**	**0.955**	**0.780**	**0.852**

grids, with each layer storing 4-channel features. The grids' resolution is adjusted according to the scene's complexity. The decoder of the geometry model, the color model, and the normal compensation model each is represented by an MLP with 4 hidden layers. We sample 1024 pixels per batch. The loss weights are set to $\lambda_n = \lambda_e = 0.1$. We use the Adam optimizer with an initial learning rate of 1×10^{-3}. The first training stage requires around 20,000 iterations. And the second training stage usually requires 60,000 to 80,000 iterations.

4.2. Comparisons

Results in real-world dataset. We visualize the reconstruction results of different methods in Fig. 5 and present the quantitative results in Tab. 1. Both qualitative and quantitative results demonstrate that our method achieves the best reconstruction performance among all methods. COLMAP [36] and neural implicit representation meth-
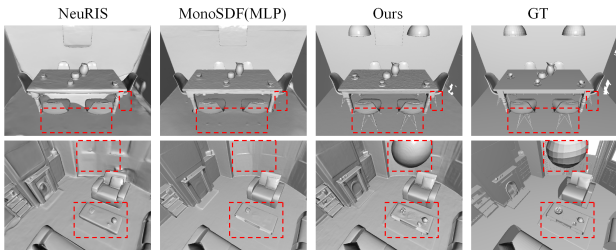

Figure 6. Qualitative results on ICL-NUIM [17].

model [12] to predict monocular normal priors, using images with a resolution of 384×384. The smooth feature branch of the geometry model uses an MLP with 4 hidden layers, while the grid feature branch utilizes 8-layers voxel
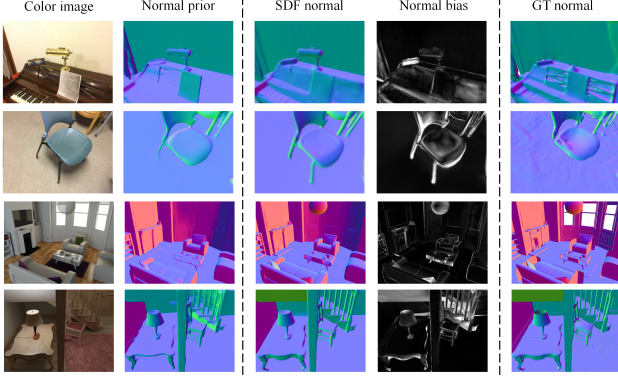
Figure 7. Visualization of normal compensation.

ods without additional supervision including VolSDF [50] and NeuS [45] produce unsatisfactory results in regions with weak textures. The methods with additional priors, such as ManhattanSDF [16], HelixSurf [22], and MonoSDF (MLP) [52] do improve the reconstruction quality; however, their reconstructions exhibit noisy or missing surfaces due to the limitations in the quality of priors and the expressiveness of the model. NeuRIS [44] designs a multi-view consistency checking strategy to filter unreliable normal priors. Despite contributing to more detailed reconstructions, the handcrafted strategy lacks robustness in handling real-world noise and finally leads to non-smooth surfaces. While MonoSDF (Grid) improves the reconstruction of fine geometry by enhancing the expressiveness of the geometry model, it generates noisy surfaces due to the lack of spatial consistency constraints for voxel grids. In contrast, our NC-SDF excels in capturing intricate geometry while producing smooth surfaces in texture-less regions.

Results in synthetic dataset. Both the qualitative results in Fig. 6 and the quantitative results in Tab. 2 validate that our NC-SDF significantly outperforms existing methods.

Additionally, we visualize the rendered outputs related to the normal compensation in Fig. 7. The normal bias in the Fig. 7 is computed as follows:

$$\mathbf{N}^{\text{bias}} = \sum_{j=1}^{3} |\mathbf{N}_j^{\text{SDF}} - \mathbf{N}_j^{\text{comp}}|, \qquad (11)$$

where $\mathbf{N}_j^{\text{SDF}}$ and $\mathbf{N}_j^{\text{comp}}$ represent the rendered SDF normal map and the rendered compensated normal map, respectively, in the j-th channel. The visualization of the bias map proves that our NC model is capable of learning the biases in normal priors. Furthermore, we present the rendered results at different training stages in Fig. 8. The results indicate that the NC model gradually learns the normal biases, thereby resulting in a gradual enhancement in the quality of both view synthesis and geometric reconstruction.

4.3. Ablation studies

For each of our three designs, we conduct the corresponding ablation studies. The quantitative results are shown in Tab. 3
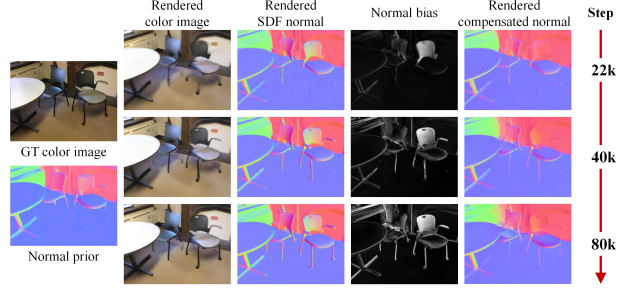
and the qualitative results are shown in Fig. 10. The results indicate the effectiveness of each design, and the combination of these three designs yields the best performance.

We conduct experiments with six configurations: (1) **MLP(baseline)**: MLPs are utilized to model the SDF and radiance field. Random sampling is employed, and supervision is provided by normal priors and color images. (2) **Grid(baseline)**: A modification of (1) where voxel grids are used to model the SDF. (3) **Hybrid**: A modification of (1) where our hybrid geometry model is used to model the SDF. (4) **MLP+IPS**: A variation of (1) that integrates our informative pixel sampling (IPS). (5) **Hybrid+IPS**: This combines the hybrid geometry model with informative pixel sampling. (6) **Ours**: A combination of our three designs, including the informative pixel sampling, the hybrid geometry model, and the normal compensation model.

Effectiveness of the hybrid geometry model. The comparison between **MLP(baseline)**, **Grid(baseline)**, and **Hybrid** illustrates that our hybrid geometry model enhances the reconstruction quality. **MLP(baseline)** tends to produce over-smooth surfaces while **Grid(baseline)** generates noisy surfaces. In contrast, **Hybrid** strikes a balance between the smoothness of surfaces and the sharpness of details.

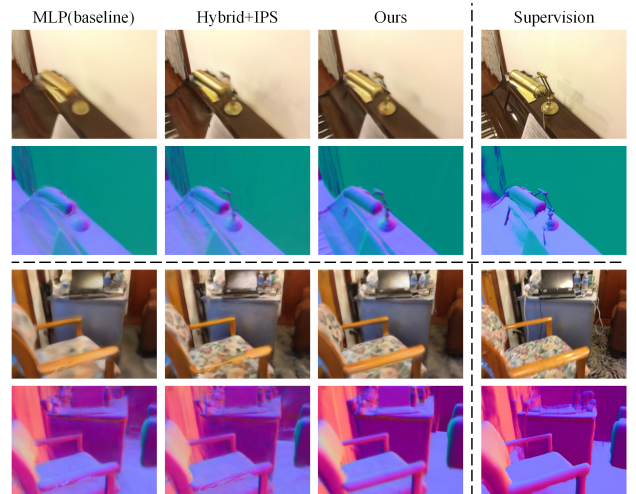Figure 8. Rendered results at different training stages.

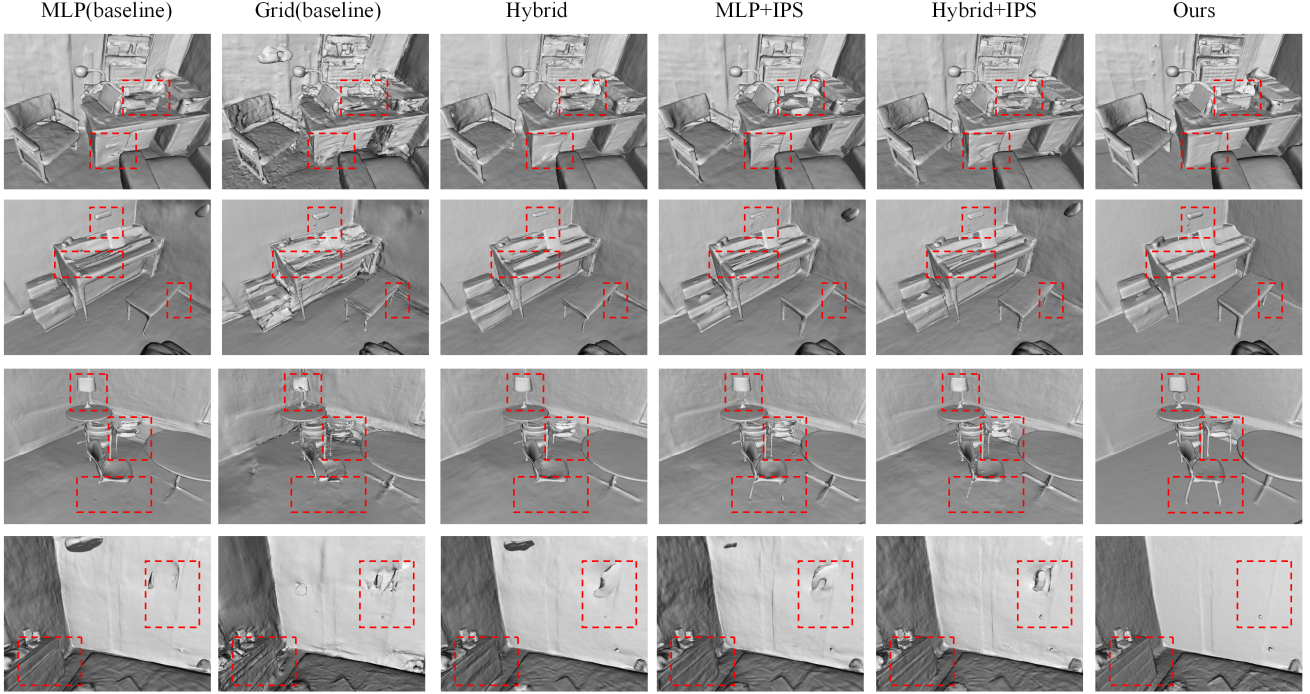Figure 9. Comparison of rendered results from three ablation experiments.

Figure 10. Visualization of ablation studies on ScanNet.

Table 3. Ablation studies on ScanNet.

Method	MLP	Grid	Informative pixel sampling	Normal compensation	Acc↓	Comp↓	Prec↑	Recall↑	F-score↑
MLP(baseline)	✓				0.040	0.042	0.750	0.718	0.733
Grid(baseline)		✓			0.048	0.046	0.716	0.686	0.700
Hybrid	✓	✓			0.039	0.041	0.766	0.727	0.745
MLP+IPS	✓		✓		0.039	0.042	0.764	0.724	0.742
Hybrid+IPS	✓	✓	✓		0.038	0.041	0.771	0.729	0.749
Ours	✓	✓	✓	✓	**0.033**	**0.039**	**0.811**	**0.754**	**0.781**

Effectiveness of the informative pixel sampling. Comparison between **MLP(baseline)** and **MLP+IPS**, as well as **Hybrid** and **Hybrid+IPS**, indicates that our sampling strategy results in a modest improvement in reconstruction quality. And it effectively enhances the reconstruction of geometric details, such as chair legs and doorknobs.

Effectiveness of the normal compensation model. The comparison between **Hybrid+IPS** and **Ours** verifies that our normal compensation model alleviates the problems caused by multi-view inconsistency between monocular normal priors, including the non-smoothness of surfaces and the loss of details. **Hybrid+IPS** improves the reconstruction quality with a 2.18% increase in F-score. The introduction of the normal compensation model further increases the F-score by 4.37%. Furthermore, we visualize the rendered color images and the rendered SDF normal maps from three ablation experiments in Fig. 9. The comparison showcases that the normal compensation model

yields a more accurate radiance field and geometry field.

5. Conclusion

We present NC-SDF, a neural SDF 3D reconstruction framework with view-dependent normal compensation. The framework focuses on enhancing indoor scene reconstruction by addressing multi-view inconsistency between monocular normal priors. Specifically, we integrate view-dependent biases in normal priors into the neural implicit representation of the scene. In addition, we propose an informative pixel sampling strategy and a hybrid geometry modeling approach to further enhance reconstruction details. Experiments on real-world and synthetic datasets demonstrate that NC-SDF achieves state-of-the-art performance in indoor scene reconstruction.

References

[1] Ahmed Shihab Ahmed. Comparative study among sobel, prewitt and canny edge detection operators used in image processing. *J. Theor. Appl. Inf. Technol*, 96(19):6517–6525, 2018. 5

[2] Devansh Arpit, Stanisław Jastrzębski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S Kanwal, Tegan Maharaj, Asja Fischer, Aaron Courville, Yoshua Bengio, et al. A closer look at memorization in deep networks. In *International conference on machine learning*, pages 233–242. PMLR, 2017. 4

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

[4] Connelly Barnes, Eli Shechtman, Adam Finkelstein, and Dan B Goldman. Patchmatch: A randomized correspondence algorithm for structural image editing. *ACM Trans. Graph.*, 28(3):24, 2009. 1, 2

[5] Jonathan T Barron, Ben Mildenhall, Dor Verbin, Pratul P Srinivasan, and Peter Hedman. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5470–5479, 2022. 2

[6] Michael Bleyer, Christoph Rhemann, and Carsten Rother. Patchmatch stereo-stereo matching with slanted support windows. In *Bmvc*, pages 1–11, 2011. 1, 2

[7] Aljaz Bozic, Pablo Palafox, Justus Thies, Angela Dai, and Matthias Nießner. Transformerfusion: Monocular rgb scene reconstruction using transformers. *Advances in Neural Information Processing Systems*, 34:1403–1414, 2021. 1, 2

[8] John Francis Canny. Finding edges and lines in images. 1983. 5

[9] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5828–5839, 2017. 5, 6

[10] Lijun Ding and Ardeshir Goshtasby. On the canny edge detector. *Pattern recognition*, 34(3):721–725, 2001. 5

[11] Wei Dong, Christopher Choy, Charles Loop, Or Litany, Yuke Zhu, and Anima Anandkumar. Fast monocular scene reconstruction with global-sparse local-dense grids. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4263–4272, 2023. 1

[12] Ainaz Eftekhar, Alexander Sax, Jitendra Malik, and Amir Zamir. Omnidata: A scalable pipeline for making multitask mid-level vision datasets from 3d scans. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10786–10796, 2021. 1, 3, 6

[13] Ziyue Feng, Liang Yang, Pengsheng Guo, and Bing Li. Cvrecon: Rethinking 3d geometric feature learning for neural reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17750–17760, 2023. 1, 2

[14] Sara Fridovich-Keil, Alex Yu, Matthew Tancik, Qinhong Chen, Benjamin Recht, and Angjoo Kanazawa. Plenoxels: Radiance fields without neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5501–5510, 2022. 2

[15] Amos Gropp, Lior Yariv, Niv Haim, Matan Atzmon, and Yaron Lipman. Implicit geometric regularization for learning shapes. *arXiv preprint arXiv:2002.10099*, 2020. 5

[16] Haoyu Guo, Sida Peng, Haotong Lin, Qianqian Wang, Guofeng Zhang, Hujun Bao, and Xiaowei Zhou. Neural 3d scene reconstruction with the manhattan-world assumption. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5511–5520, 2022. 1, 2, 5, 6, 7

[17] Ankur Handa, Thomas Whelan, John McDonald, and Andrew J Davison. A benchmark for rgb-d visual odometry, 3d reconstruction and slam. In *2014 IEEE international conference on Robotics and automation (ICRA)*, pages 1524–1531. IEEE, 2014. 5, 6

[18] Haibo He and Edwardo A Garcia. Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering*, 21(9):1263–1284, 2009. 4

[19] Po-Han Huang, Kevin Matzen, Johannes Kopf, Narendra Ahuja, and Jia-Bin Huang. Deepmvs: Learning multi-view stereopsis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2821–2830, 2018. 1, 2

[20] Sunghoon Im, Hae-Gon Jeon, Stephen Lin, and In So Kweon. Dpsnet: End-to-end deep plane sweep stereo. *arXiv preprint arXiv:1905.00538*, 2019. 2

[21] Bartosz Krawczyk. Learning from imbalanced data: open challenges and future directions. *Progress in Artificial Intelligence*, 5(4):221–232, 2016. 4

[22] Zhihao Liang, Zhangjin Huang, Changxing Ding, and Kui 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*, pages 13165–13174, 2023. 1, 2, 5, 6, 7

[23] Philipp Lindenberger, Paul-Edouard Sarlin, Viktor Larsson, and Marc Pollefeys. Pixel-perfect structure-from-motion with featuremetric refinement. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5987–5997, 2021. 1, 2

[24] William E Lorensen and Harvey E Cline. Marching cubes: A high resolution 3d surface construction algorithm. In *Seminal graphics: pioneering efforts that shaped the field*, pages 347–353. 1998. 4

[25] Keyang Luo, Tao Guan, Lili Ju, Haipeng Huang, and Yawei Luo. P-mvsnet: Learning patch-wise matching confidence aggregation for multi-view stereo. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10452–10461, 2019. 2

[26] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1):99–106, 2021. 2, 4

[27] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Transactions on Graphics (ToG)*, 41(4):1–15, 2022. 2

[28] Zak Murez, Tarrence Van As, James Bartolozzi, Ayan Sinha, Vijay Badrinarayanan, and Andrew 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 16*, pages 414–431. Springer, 2020. 1, 2, 5

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

[30] Joseph Ortiz, Alexander Clegg, Jing Dong, Edgar Sucar, David Novotny, Michael Zollhoefer, and Mustafa Mukadam. isdf: Real-time neural signed distance fields for robot perception. *arXiv preprint arXiv:2204.02296*, 2022. 1

[31] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 165–174, 2019. 2

[32] Xiaojuan Qi, Renjie Liao, Zhengzhe Liu, Raquel Urtasun, and Jiaya Jia. Geonet: Geometric neural network for joint depth and surface normal estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 283–291, 2018. 1, 3

[33] Michael Ramamonjisoa and Vincent Lepetit. Sharpnet: Fast and accurate recovery of occluding contours in monocular depth estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019. 1, 3

[34] Barbara Roessle, Jonathan T Barron, Ben Mildenhall, Pratul P Srinivasan, and Matthias Nießner. Dense depth priors for neural radiance fields from sparse input views. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12892–12901, 2022. 1, 4

[35] Jianyuan Ruan, Bo Li, Yibo Wang, and Yuxiang Sun. Slamesh: Real-time lidar simultaneous localization and meshing. *arXiv preprint arXiv:2303.05252*, 2023. 1

[36] Johannes L Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4104–4113, 2016. 1, 2, 5, 6

[37] Johannes L Schönberger, Enliang Zheng, Jan-Michael Frahm, and Marc 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 14*, pages 501–518. Springer, 2016. 2

[38] GT Shrivakshan and Chandramouli Chandrasekar. A comparison of various edge detection techniques used in image processing. *International Journal of Computer Science Issues (IJCSI)*, 9(5):269, 2012. 5

[39] Noah Stier, Alexander Rich, Pradeep Sen, and Tobias Höllerer. Vortx: Volumetric 3d reconstruction with transformers for voxelwise view selection and fusion. In *2021 International Conference on 3D Vision (3DV)*, pages 320–330. IEEE, 2021. 1, 2

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

[41] Jiaming Sun, Yiming Xie, Linghao Chen, Xiaowei Zhou, and Hujun Bao. Neuralrecon: Real-time coherent 3d reconstruction from monocular video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15598–15607, 2021. 1, 2

[42] Yanmin Sun, Andrew KC Wong, and Mohamed S Kamel. Classification of imbalanced data: A review. *International journal of pattern recognition and artificial intelligence*, 23 (04):687–719, 2009. 4

[43] Jingwen Wang, Tymoteusz Bleja, and Lourdes Agapito. Go-surf: Neural feature grid optimization for fast, high-fidelity rgb-d surface reconstruction. In *2022 International Conference on 3D Vision (3DV)*, pages 433–442. IEEE, 2022. 1, 2

[44] Jiepeng Wang, Peng Wang, Xiaoxiao Long, Christian Theobalt, Taku Komura, Lingjie Liu, and Wenping Wang. Neuris: Neural reconstruction of indoor scenes using normal priors. In *European Conference on Computer Vision*, pages 139–155. Springer, 2022. 1, 2, 3, 4, 5, 6, 7

[45] Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. *arXiv preprint arXiv:2106.10689*, 2021. 1, 2, 3, 4, 5, 6, 7

[46] Rui Wang, David Geraghty, Kevin Matzen, Richard Szeliski, and Jan-Michael Frahm. Vplnet: Deep single view normal estimation with vanishing points and lines. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 689–698, 2020. 1, 3

[47] Yi Wei, Shaohui Liu, Yongming Rao, Wang Zhao, Jiwen Lu, and Jie Zhou. Nerfingmvs: Guided optimization of neural radiance fields for indoor multi-view stereo. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5610–5619, 2021. 1, 2

[48] Yao Yao, Zixin Luo, Shiwei Li, Tian Fang, and Long Quan. Mvsnet: Depth inference for unstructured multi-view stereo. In *Proceedings of the European conference on computer vision (ECCV)*, pages 767–783, 2018. 1, 2

[49] Yao Yao, Zixin Luo, Shiwei Li, Tianwei Shen, Tian Fang, and Long Quan. Recurrent mvsnet for high-resolution multi-view stereo depth inference. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5525–5534, 2019. 2

[50] Lior Yariv, Jiatao Gu, Yoni Kasten, and Yaron Lipman. Volume rendering of neural implicit surfaces. *Advances in Neural Information Processing Systems*, 34:4805–4815, 2021. 1, 2, 4, 5, 6, 7

[51] Wei Yin, Yifan Liu, Chunhua Shen, and Youliang Yan. Enforcing geometric constraints of virtual normal for depth pre-

diction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5684–5693, 2019. 1, 3

[52] Zehao Yu, Songyou Peng, Michael Niemeyer, Torsten Sattler, and Andreas Geiger. Monosdf: Exploring monocular geometric cues for neural implicit surface reconstruction. *Advances in neural information processing systems*, 35:25018–25032, 2022. 1, 2, 3, 4, 5, 6, 7

[53] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115, 2021. 4

[54] Youmin Zhang, Fabio Tosi, Stefano Mattoccia, and Matteo Poggi. Go-slam: Global optimization for consistent 3d instant reconstruction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3727–3737, 2023. 1

[55] Jingsen Zhu, Yuchi Huo, Qi Ye, Fujun Luan, Jifan Li, Dianbing Xi, Lisha Wang, Rui Tang, Wei Hua, Hujun Bao, 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*, pages 12489–12498, 2023. 1

[56] Zihan Zhu, Songyou Peng, Viktor Larsson, Weiwei Xu, Hujun Bao, Zhaopeng Cui, Martin R Oswald, and Marc Pollefeys. Nice-slam: Neural implicit scalable encoding for slam. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12786–12796, 2022. 1, 2