2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

Leveraging Geometric Prior Uncertainty and Complementary Constraints for High-Fidelity Neural Indoor Surface Reconstruction

Qiyu Feng[1*], Jiwei Shan[2*], Shing Shin Cheng[2] and Hesheng Wang[1]

***Abstract*— Neural implicit surface reconstruction with signed distance function has made significant progress, but recovering fine details such as thin structures and complex geometries remains challenging due to unreliable or noisy geometric priors. Existing approaches rely on implicit uncertainty that arises during optimization to filter these priors, which is indirect and inefficient, and masking supervision in high-uncertainty regions further leads to under-constrained optimization. To address these issues, we propose GPU-SDF, a neural implicit framework for indoor surface reconstruction that leverages geometric prior uncertainty and complementary constraints. We introduce a self-supervised module that explicitly estimates prior uncertainty without auxiliary networks. Based on this estimation, we design an uncertainty-guided loss that modulates prior influence rather than discarding it, thereby retaining weak but informative cues. To address regions with high prior uncertainty, GPU-SDF further incorporates two complementary constraints: an edge distance field that strengthens boundary supervision and a multi-view consistency regularization that enforces geometric coherence. Extensive experiments confirm that GPU-SDF improves the reconstruction of fine details and serves as a plug-and-play enhancement for existing frameworks. Source code will be available at https://github.com/IRMVLab/GPU-SDF**

I. INTRODUCTION

Three-dimensional surface reconstruction from multi-view images is a long-standing challenge in computer vision and graphics. Accurate and dense geometry is crucial for applications such as AR/VR systems, robotic navigation and embodied intelligence. Traditional methods have made significant progress but still suffer from incomplete reconstructions and difficulty with textureless surfaces. Recently, differentiable rendering approaches—most notably Neural Radiance Fields [1] and 3D Gaussian Splatting [2]—have enabled photorealistic view synthesis, yet they often struggle to recover explicit surface geometry. To overcome this, researchers have explored neural signed distance functions (Neural SDF) [3], which represent surfaces implicitly as MLP-based signed distance fields optimized via differentiable rendering, enabling higher-fidelity reconstruction. Recent methods [4]–[10] further improve results in textureless regions and complex geometries by incorporating monocular geometric predictions as additional priors. Despite these advances, key challenges remain.

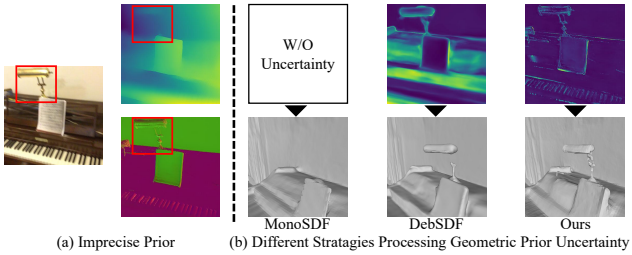

Fig. 1. (a) Monocular geometric priors for 3D reconstruction are often imprecise, especially for thin structures. (b) Comparison of different strategies to process geometric prior uncertainty. MonoSDF [4] uses priors directly, without handling uncertainty. DebSDF [7] relies on an implicit uncertainty that emerges during optimization, discarding supervision in unreliable regions and forcing reliance solely on RGB cues. In contrast, our method explicitly estimates prior uncertainty at the outset. We then employ an uncertainty-guided loss to modulate the influence of priors according to their reliability, rather than discarding them. Together with the additional geometric constraints we introduce, our method reconstructs fine-grained structures more effectively.

As shown in Fig. 1, existing Neural SDF methods can effectively reconstruct the overall indoor structure, but they still struggle to recover fine details such as chair legs and railings. Prior studies (e.g., DebSDF [7]) attribute these failures to several factors, including errors in monocular geometric priors due to domain gaps, the lack of multi-view consistency in independently predicted priors, and the low sampling probability of thin structures. To alleviate these issues, DebSDF proposes a strategy of estimating uncertainty during the SDF optimization process. Specifically, they use the SDF model's own emergent uncertainty about the geometry to filter unreliable supervision or reweight ray sampling. While this strategy improves robustness, it suffers from a critical limitation: **(1) Ignoring Explicit Prior Uncertainty:** This filtering process relies solely on an *implicit* uncertainty that emerges *during* the SDF optimization. This uncertainty reflects the model's own difficulty in fitting the geometric priors, rather than being a direct, upfront assessment of the priors' inherent quality. Consequently, the model must first "learn" to be uncertain by struggling with noisy or inconsistent data—a process that is both indirect and inefficient. This conflation of the model's learning state with the prior's intrinsic quality can lead to suboptimal decisions: confidently incorrect priors might be retained if they are internally consistent, while weakly informative but correct priors might be discarded prematurely. **(2) Under-constrained optimization in high-uncertainty regions:** When geometric supervision is masked out, the learning signal for recovering scene structure becomes primarily reliant on RGB. This is often

*The first two authors contributed equally. This work was supported by National Key R&D Program of China (Grant No.2024YFB4708900). It was also supported in part by the Natural Science Foundation of China under Grant 62225309, U24A20278, 62361166632. It was also supported in part by Innovation and Technology Commission of Hong Kong (ITS/235/22) and in part by Multi-scale Medical Robotics Center, InnoHK. Corresponding Authors: Hesheng Wang, Shing Shin Cheng.

[1] Department of Automation, Shanghai Jiao Tong University, Shanghai 200240, China.
[2] Department of Mechanical and Automation Engineering and T Stone Robotics Institute, The Chinese University of Hong Kong, Hong Kong.

insufficient for recovering accurate geometry in textureless or thin structures where RGB cues are weak or ambiguous.

To overcome these challenges, we propose **GPU-SDF**, a neural implicit reconstruction method with two key innovations for high-quality indoor 3D reconstruction. First, to reduce reliance on indirect, model-derived uncertainty, we introduce a self-supervised approach that explicitly estimates the confidence of geometric priors at the outset, without external networks. This separates prior quality assessment from the learning state of the SDF model. With this explicit confidence, we design a new uncertainty-guided geometric consistency loss. Instead of discarding noisy supervision, this loss adjusts the influence of priors based on their estimated reliability. In this way, the model still learns from weak but useful signals and mitigates degradation caused by missing supervision. Second, to address under-constrained optimization in high-uncertainty regions, we add two complementary geometric constraints. An edge distance field provides robust scene edge information as an auxiliary cue. In parallel, a multi-view consistency regularization enforces coherent geometry across different viewpoints. Together, these constraints supply the guidance needed to reconstruct thin and fine structures where RGB cues alone are ambiguous or unreliable. Moreover, our framework is modular and can be seamlessly integrated as a plug-in into existing neural SDF reconstruction pipelines, thereby enhancing their reconstruction performance. The main contributions are as follows:

- We propose a novel self-supervised uncertainty estimation method combined with an uncertainty-guided geometric supervision strategy, which preserves weak geometric signals in high-uncertainty regions and alleviates degradation caused by discarding priors.
- We design an edge distance field and a multi-view consistency regularization that strengthen supervision in under-constrained regions, improving the reconstruction of thin and fine structures.
- We conduct extensive experiments on challenging benchmarks, demonstrating not only the effectiveness of our method but also its plug-and-play ability to enhance existing SDF-based frameworks.

II. Related Works

A. Neural Implicit and Gaussian Splatting based Indoor Surface Reconstruction

Recently, neural implicit methods have received significant attention for representing scene geometry. NeRF-type approaches [1], [11]–[15] encode coordinate-based density and appearance using multilayer perceptrons (MLPs), represent scene geometry through signed distance functions (SDFs) [3], and extract surface geometry via marching cubes [16]. For indoor scenes, neural implicit reconstruction methods using only monocular RGB images as input often struggle to reconstruct low-texture surfaces and complex geometries. To address this issue, additional priors are incorporated to constrain scene geometry, such as the Manhattan World assumption [17], depth and normal priors [4], [6], or semantic information [18] to regularize textureless regions. Other approaches introduce alternative geometric representations or additional regularization terms [5], [8], [19]–[24] to improve reconstruction quality. In parallel, reconstruction methods based on 3D Gaussian Splatting (3DGS) [2] have shown strong potential for 3D reconstruction. Several works [25]–[29] have explored surface reconstruction within 3DGS framework, but a high-fidelity extraction technique—akin to marching cubes for NeRF's implicit fields—remains underdeveloped, resulting in suboptimal surface quality. Hence, this paper focuses on neural implicit indoor surface reconstruction.

B. Depth and Normal Uncertainty Estimation

Monocular depth and normal estimation is a fundamental task, while it faces inherent challenges such as scale ambiguity, discontinuities, and systematic bias. Estimating the uncertainty of these predictions is therefore crucial for improving accuracy and reliability. Existing approaches include uncertainty estimation with external networks [30], or training a network from scratch [31] to predict probability distributions. However, these methods are computationally expensive. In contrast, self-supervised approaches [32], [33] estimate uncertainty by measuring the variance between outputs generated from perturbed inputs (e.g., image flipping). Owing to their computational efficiency, this work adopts self-supervised methods for assessing the uncertainty of monocular depth and normal predictions.

III. Methods

Given calibrated multi-view images and depth/normal priors obtained from a pre-trained model, our objective is to develop a reconstruction framework based on Neural Signed Distance Fields (SDFs) that produces high-precision dense surfaces while preserving fine geometric details. As illustrated in Fig. 2, the proposed GPU-SDF framework is composed of three main components: *Neural SDF*, *Prior Uncertainty Identification*, and *Loss Functions*. The Neural SDF module follows the general pipeline of mainstream methods. Rather than being tied to a specific baseline, our method can be seamlessly integrated as a plugin to enhance the performance of existing neural SDF frameworks. In this work, we adopt ND-SDF [8], one of the state-of-the-art approaches, as the base network to demonstrate the effectiveness of our design, and provide a detailed review of its key techniques in Sec. III-A. The Prior Uncertainty Identification module (Sec. III-B) introduces an effective self-supervised strategy for estimating the uncertainty of geometric priors. The Loss Function module (Sec. III-C) incorporates both standard losses used in neural SDF frameworks and three additional objectives specifically designed to account for prior uncertainty and enhance reconstruction, especially in regions with fine structures. Finally, the model optimization process is described in Sec. III-D.

A. Preliminary

Neural Signed Distance Fields. Neural SDFs represent scene geometry and appearance with an implicit neural

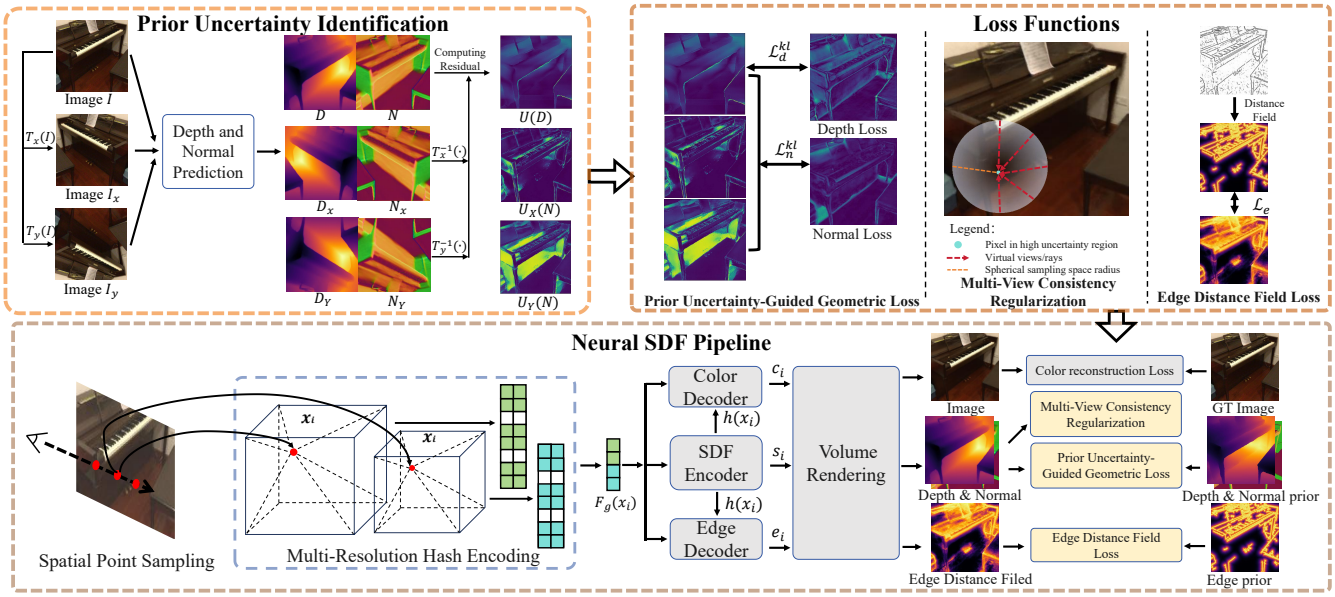

Fig. 2. **Overview of the GPU-SDF framework.** GPU-SDF reconstructs high-fidelity surfaces from multi-view RGB images and initial geometric priors (e.g., depth, normals). It consists of three parts: **(1) Neural SDF pipeline:** builds upon existing frameworks that represent the scene with an SDF and color field, and augments them by jointly learning an edge distance field as a robust geometric cue for fine structure reconstruction. **(2) Prior uncertainty identification:** a self-supervised module explicitly estimates confidence of geometric priors, enabling dynamic adjustment of their influence during optimization. **(3) Loss functions:** an uncertainty-guided consistency loss preserves weak but useful signals, while edge distance field supervision and multi-view regularization facilitate accurate recovery of thin and fine structures.

network, which is optimized through differentiable volume rendering. Specifically, given a ray $r(t) = o+tv$ from camera origin o in direction v, we sample N points $x_i = o+t_i v$ along the ray. The implicit neural network predicts the SDF value s_i and color c_i for each point. To enable volume rendering, SDF values are converted into volume densities as [11]:

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

where β is a learnable parameter.

The rendered color, depth, and normal are then obtained through volume rendering:

$$\hat{C}(r) = \sum_{i=1}^{N} T_i \alpha_i c_i, \ \hat{D}(r) = \sum_{i=1}^{N} T_i \alpha_i t_i, \ \hat{N}(r) = \sum_{i=1}^{N} T_i \alpha_i n_i, \tag{2}$$

where $\alpha_i = 1 - \exp(-\sigma_i \delta_i)$ represents the opacity at x_i, and δ_i denotes distance between adjacent samples. $T_i = \prod_{j=1}^{i-1}(1-\alpha_j)$ denotes the accumulated transmittance of light as it travels to the i point along the ray. n_i is the analytical gradient of the SDF network at point i.

The network parameters are optimized by minimizing reconstruction losses with respect to input RGB images and geometric priors:

$$\mathcal{L}_c = \sum_{r\in\mathcal{R}} \|\hat{C}(r) - C(r)\|_1, \quad \mathcal{L}_d = \sum_{r\in\mathcal{R}} \|w\hat{D}(r) + q - D(r)\|^2,$$

$$\mathcal{L}_n = \sum_{r\in\mathcal{R}} \|\hat{N}(r) - N(r)\|_1 + \|1 - \hat{N}(r)^\top N(r)\|_1, \tag{3}$$

where $\mathcal{R}$ represents the set of sampled rays, and $C(r)$ denotes the GT RGB values. $D(r)$ and $N(r)$ are depth and normal priors [34], and w, q are scale and shift factors estimated by least squares. Finally, an eikonal regularization is imposed to enforce the SDF property:

$$\mathcal{L}_{\text{eik}} = \frac{1}{N} \sum_{i=1}^{N} \left(\|\nabla s(\mathbf{x}_i)\|_2 - 1\right)^2. \tag{4}$$

ND-SDF. To adaptively model the reliability of normal priors across different regions of a 3D scene, ND-SDF [8] extends the standard Neural SDF framework by introducing a learnable deflection field. Specifically, it predicts the deflection angle between the scene normals and the provided normal priors, represented as quaternions q_i. During rendering, the quaternion at the ray–surface intersection is aggregated as $Q(r) = \sum_{i=1}^{N} T_i \alpha_i q_i$ and the rendered normal is corrected via quaternion rotation $\hat{N}^d(r) = Q(r) \otimes \hat{N}(r) \otimes Q^{-1}(r)$, where $\hat{N}(r)$ is the original rendered normal and $\otimes$ denotes quaternion multiplication. The normal reconstruction loss is then built upon the deflected normal $\hat{N}^d(r)$ and the priors:

$$\mathcal{L}_n^d = \sum_{r\in\mathcal{R}} \|\hat{N}^d(r) - N(r)\|_1 + \|1 - \hat{N}^d(r)^\top N(r)\|_1. \tag{5}$$

Furthermore, ND-SDF introduces an angle-aware reweighting mechanism for RGB, depth, and normal losses:

$$\mathcal{L}_p^{ad} = w_p(\Delta\theta)\, \mathcal{L}_p, \quad p \in \{c, d, n\}, \tag{6}$$

where $\Delta\theta = \arccos(\hat{N}(r) \cdot \hat{N}^d(r))$ denotes the deflection angle, and w_p adjusts the confidence of each loss accordingly.

The final reconstruction objective combines the weighted color, depth, and normal terms with eikonal regularization:

$$\mathcal{L}_{recon} = \lambda_c \mathcal{L}_c^{ad} + \lambda_d \mathcal{L}_d^{ad} + \lambda_n \mathcal{L}_n^{ad} + \lambda_{eik} \mathcal{L}_{eik}. \tag{7}$$

B. Prior Uncertainty Identification

To explicitly evaluate the reliability of geometric priors, rather than relying on the implicit uncertainty generated during optimization in existing methods [7], we introduce a self-supervised uncertainty estimation module. This module models uncertainty as a distribution that reflects the variance of prediction errors. Based on this explicit uncertainty, we further design an uncertainty-guided geometric consistency loss (Sec. III-C) that adaptively adjusts the influence of noisy priors, leading to more accurate and robust reconstruction.

Specifically, given an RGB image I and a pretrained depth estimation network D_θ [34], we obtain the monocular depth map as $D = D_\theta(I)$. Existing approaches to estimate depth uncertainty usually require either retraining the model or relying on ground-truth depth [30], [31], [35], both of which are impractical in many real-world scenarios. Therefore, we aim to design a self-supervised, post-hoc method that can be directly applied to pretrained models. Inspired by [32], [33], we introduce a simple yet effective strategy based on flip consistency. Let $T_m(\cdot)$ denote a transformation along axis $m \in \{x, y\}$, corresponding to horizontal or vertical flipping, and $T_m^{-1}(\cdot)$ its inverse operation. Since these transformations preserve pixel-level structures, the depth prediction of the original image I, denoted as D, should remain consistent with the predictions of the flipped images $I_m = T_m(I)$, denoted as D_m. To estimate depth uncertainty, we compute the depth maps for both the original and flipped images, and then realign the flipped predictions:

$$\begin{aligned} D = D_\theta(I), \quad D_x = D_\theta(I_x), \quad D_y = D_\theta(I_y), \\ \tilde{D}_x = T_x^{-1}(D_x), \quad \tilde{D}_y = T_y^{-1}(D_y). \end{aligned} \tag{8}$$

The pixel-wise depth uncertainty $U(D) \in \mathbf{R}^{H \times W}$ is then defined as:

$$\begin{aligned} U_x(D) = |D - \tilde{D}_x|, U_y(D) = |D - \tilde{D}_y|, \\ U(D) = \sqrt{\frac{U_x^2(D) + U_y^2(D)}{2}} \end{aligned} \tag{9}$$

Here, $U_x(D)$ and $U_y(D)$ represent the uncertainty components obtained from horizontal and vertical flips, respectively. Then we adopt their standard deviation as a unified metric for quantifying the uncertainty of depth prior. Following the same principle, we compute normal uncertainty. Given the predicted normal map N, we obtain the aligned flipped predictions $\tilde{N}_x, \tilde{N}_y$ and define:

$$U_x(N) = |N - \tilde{N}_x|, \quad U_y(N) = |N - \tilde{N}_y|. \tag{10}$$

Furthermore, due to the geometric relationship between depth and normals, the uncertainty along the z-axis is directly derived from the depth uncertainty:

$$U_z(N) = U(D). \tag{11}$$

Compared with prior works [32], [33] that rely only on horizontal flips, our approach incorporates both horizontal and vertical consistency, producing a more robust uncertainty map by capturing geometric inconsistencies missed by single-axis tests (see Fig. 3). Furthermore, we extend this principle from depth to normal estimation, providing stronger optimization guidance and yielding higher-quality reconstructions (Sec. IV-B).

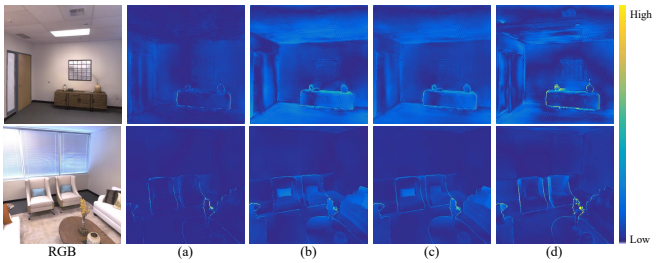

Fig. 3. Visualization of our self-supervised uncertainty estimation. (a) Uncertainty from horizontal flips, $U_x(D)$; (b) uncertainty from vertical flips, $U_y(D)$; (c) the combined uncertainty, $U(D)$, which aligns closely with the ground-truth depth error map in (d).

C. Loss Functions

We optimize and train our model by minimizing the loss function with respect to the implicit network parameters θ of the neural SDF. In addition to the basic losses introduced in Sec. III-A, the estimated prior uncertainty enables two key improvements: 1) adaptively controlling the supervision strength of prior-related losses, leading to our *Prior Uncertainty-Guided Geometric Loss*; 2) introducing additional constraints in regions with high prior uncertainty, including an *Edge Distance Field loss* and a *multi-view consistency regularization* module.

Prior Uncertainty-Guided Geometric Loss. Most existing methods mitigate the effect of uncertain priors by discarding supervision in high-uncertainty regions based on a predefined threshold [7]. However, this simplification leads to under-constrained optimization that relies mainly on RGB supervision, often causing local blurring or structural loss in reconstruction. To overcome this limitation, we design a prior uncertainty-guided geometric loss that adaptively regularizes the contribution of geometric supervision according to its uncertainty. Concretely, let R denote the set of sampled rays. For each ray, the prior normal and its prediction are denoted as N_i^d and $\hat{N}_i^d$ with $i \in \{x, y, z\}$, and the depth prior and prediction are denoted as D and $\hat{D}$, respectively. Inspired by the form of KL divergence, we define the following regularization losses:

$$\mathcal{L}_n^{kl} = \sum_{i \in \{x,y,z\}} \sum_{r \in R} \left| N_i^d - \hat{N}_i^d \right| \cdot \ln \left(\frac{\left| N_i^d - \hat{N}_i^d \right|}{U_i(N)} \right), \tag{12}$$

$$\mathcal{L}_d^{kl} = \sum_{r \in R} \left| D - \hat{D} \right| \cdot \ln \left(\frac{\left| D - \hat{D} \right|}{U(D)} \right). \tag{13}$$

Here, $U_i(N)$ and $U(D)$ represent the prior uncertainties for normals and depth, respectively. This formulation dynamically scales the supervision strength: reliable priors enforce strong constraints, while uncertain priors contribute weaker but still informative regularization. In this way, our method avoids discarding useful information and promotes plausible geometry reconstruction even in high-uncertainty regions.

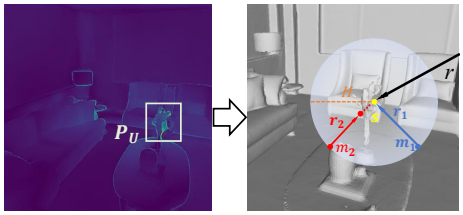

Fig. 4. Illustration of the multi-view consistency regularization. For a pixel within a high-uncertainty region P_U, its corresponding primary ray r intersects the surface at point s. A sphere with radius H is centered at s. Auxiliary rays, such as r_1 and r_2, are then sampled, originating from points (m_1, m_2) on the sphere's surface. If the first surface intersection of an auxiliary ray coincides with s, its visibility flag is set to $v_i = 1$, and it is included in the consistency loss calculation. Otherwise, the flag is set to $v_i = 0$, and the ray is excluded.

Edge Distance Field Loss. Accurately recovering sharp object boundaries and details remains a key challenge in neural surface reconstruction. This difficulty stems from the inherent unreliability of monocular geometry priors in these high-frequency regions. Without explicit guidance, the optimization can be under-constrained, leading to overly smooth or noisy geometry at object edges. To address these issues, we introduce edge information as an additional cue. Edges naturally delineate object boundaries, providing guidance for recovering fine details. They also act as complementary constraints in uncertain regions, helping to stabilize training and maintain structural fidelity. Specifically, we employ TEED [36] to extract edge maps from RGB images and convert them into edge distance fields for supervision. As shown in Fig. 2, similar to [37], we add an edge decoder to the neural SDF framework, which shares the same architecture as the SDF decoder. For each sampled 3D point x_i, the edge decoder predicts an edge value e_i. These predictions are then aggregated through volumetric rendering, as in 2, to compute the rendered edge distance field:

$$\hat{E}(r) = \sum_{i=1}^{N} T_i \alpha_i e_i. \tag{14}$$

For supervision, we pre-compute edge distance fields from edge maps using a distance transform, which serve as pseudo ground-truth. The edge distance field loss is then defined as the L1 difference between the rendered edge distance field $\hat{E}(r)$ and the pre-computed edge distance field $E(r)$:

$$\mathcal{L}_e = \sum_{r \in \mathcal{R}} \|\hat{E}(r) - E(r)\|_1. \tag{15}$$

Multi-View Consistency Regularization. 3D scenes exhibit an inherent property of geometric coherence: the same surface region, when observed from different viewpoints, should yield consistent depth estimates. To leverage this property, we introduce multi-view consistency regularization that activates only in high-uncertainty regions P_U, locally refining weak areas to improve reconstruction stability while avoiding unnecessary overhead.

Specifically, we first identify the high-uncertainty regions P_U based on the geometric prior uncertainty introduced in Sec. III-B. A pixel is included in P_U if its depth uncertainty $U(D)$ exceeds a predefined threshold, which we set empirically to 0.8. As shown in Fig. 4, for each pixel p that belongs to P_U, we compute the intersection point s between its primary ray r and the surface. We then construct a spherical sampling space centered at s with radius H. From this sphere, we uniformly sample M points m_i on its surface. For each sampled point, we generate an auxiliary ray r_i, which originates from m_i and points toward s with direction $f_i = \frac{s - m_i}{\|s - m_i\|}$. For each auxiliary ray r_i, we compute the rendered depth $\hat{D}_i^a$ using the volumetric rendering formula:

$$\hat{D}_i^a = \sum_{p \in r_i} T_p \alpha_p t_p, \tag{16}$$

where p denotes uniformly sampled points along r_i, with starting point m_i and direction f_i. This value corresponds to the estimated distance from m_i to the first surface intersection. The multi-view consistency loss is then defined as:

$$\mathcal{L}_{mv} = \sum_{p \in P_U} \frac{\sum_{i=1}^{M} v_i \left| H - \hat{D}_i^a \right|}{\sum_{i=1}^{M} v_i + \epsilon}, \tag{17}$$

where ϵ is a small constant to avoid division by zero. The visibility indicator v_i is defined as follows: if the first intersection point of the auxiliary ray r_i coincides with that of the primary ray r, we set $v_i = 1$, and the theoretical depth $\hat{D}_i^a$ of r_i should equal the sphere radius H. Otherwise, we set $v_i = 0$. This ensures that only mutually visible rays contribute to the consistency constraint, preventing invalid samples from affecting the loss.

Although our multi-view consistency regularization and RayDF [22] both involve consistency constraints, they are fundamentally distinct in their design philosophy, application scope, and implementation requirements. First, our constraint is specifically designed for neural SDF frameworks and is **selective, local**, activated only in high-uncertainty regions for targeted refinement. In contrast, RayDF is tailored for the ray–surface distance field representation and enforces a **global** consistency constraint uniformly across all rays. Second, our formulation is geometry-driven, deriving supervision from volumetric rendering in a local space, which contrasts with RayDF's classifier-driven approach. Finally, our method is lightweight and requires no auxiliary networks, unlike RayDF, which necessitates a separate classifier and a two-stage optimization, introducing overhead.

Final objective. Finally, the overall loss function for GPU-SDF is as follows:

$$\mathcal{L} = \mathcal{L}_{recon} + \lambda_n^{kl} \mathcal{L}_n^{kl} + \lambda_d^{kl} \mathcal{L}_d^{kl} + \lambda_e \mathcal{L}_e + \lambda_m \mathcal{L}_{mv}. \tag{18}$$

D. Implementation Details

During training, we randomly select four RGB images from the sequence together with their corresponding geometric priors. Using our model, we render the depth and normal maps and compute the normal deflection angle maps, as defined in ND-SDF [8]. Based on the computed normal deflection angles, we apply importance sampling following [7], [8] to select 1,024 pixels from each image. The sampled spatial points are encoded using a layer-activated hash grid with resolutions ranging from 2^5 to 2^{11} across 16 levels, with the initial activation set to 8. The encoded features F_g

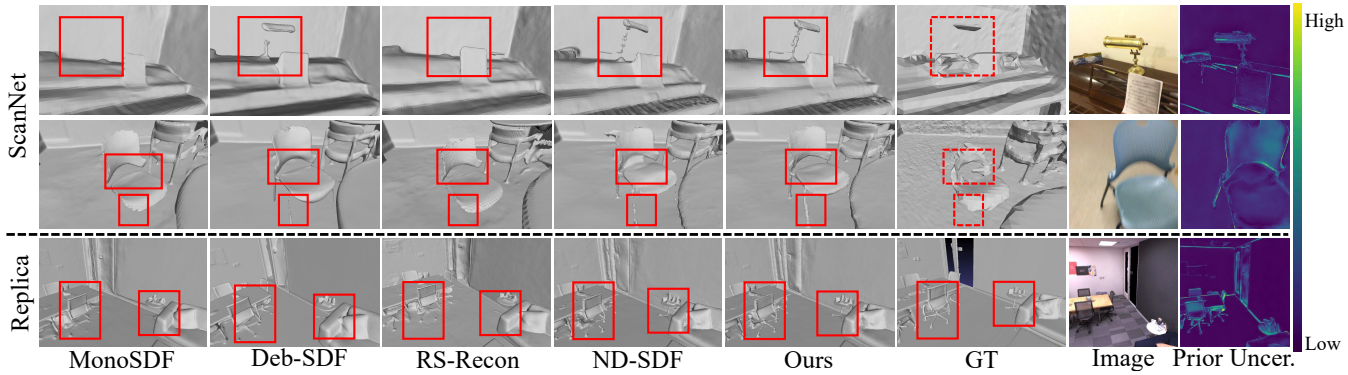

Fig. 5. Visualization results of reconstructed mesh on ScanNet and Replica, with details highlighted in red boxes. The rightmost column shows an example of the prior uncertainty for the corresponding region. Note that the ground-truth mesh of the ScanNet dataset is generated from RGB-D sensors using voxel hashing, and therefore it is not accurate in fine-detail regions. These structures are highlighted with red dashed boxes and can be verified from the RGB images. Our method achieves finer geometric structures compared to previous methods.

are then input into the SDF, color, and edge decoders. Each decoder consists of fully connected layers with a size of 256×256. Rendering is performed as described in Sec. III-A, and the parameters of our neural SDF model are optimized by minimizing 18. We use the Adam optimizer with a learning rate of 1×10^{-3}. The training runs for 128,000 iterations, taking approximately 24 GPU hours on a desktop PC with an NVIDIA RTX 4090. The weights of the loss function are set as: $\lambda_c = 1$, $\lambda_{eik} = 0.05$, $\lambda_d = 0.05$, $\lambda_n = 0.025$, $\lambda_n^{kl} = 0.025$, $\lambda_d^{kl} = 0.05$, $\lambda_e = 0.01$, and $\lambda_m = 512$. For multi-view consistency regularization, the number of points sampled on the sphere's surface is set to $M = 1$, and the sphere radius is set to $H = 0.15$.

IV. Experiments

Datasets. We validate the effectiveness of GPU-SDF on 3 challenging indoor datasets: ScanNet [38], Replica [39], ScanNet++ [40]. ScanNet is a large-scale 3D dataset of real-world indoor scenes. Replica is a high-quality, synthetically created indoor dataset. ScanNet++ is a high-fidelity real-world indoor dataset captured at high resolution. For testing, we selected four representative scenes from ScanNet, eight from Replica, and three from ScanNet++.

Metrics. We use the same evaluation metrics as previous work [4], [7], [8] to assess the meshes extracted from neural surface reconstruction: Accuracy, Completeness, Chamfer Distance, Precision, Recall, F-score and Normal Consistency.

Baselines. We compare GPU-SDF with four categories of methods: (1) Traditional MVS: COLMAP [41]; (2) Neural implicit surface reconstruction from monocular RGB: VolSDF [11], UNISURF [15], Baked-Angelo [42]; (3) Neural implicit methods with additional priors : NeuRIS [6], MonoSDF [4], RS-Recon [43], NeRFPrior [44], Deb-SDF [7], H2O-SDF [45] , ND-SDF [8]; (4) Gaussian Splatting–based methods: GSRec [27]. For fair comparison, we locally evaluate representative methods (MonoSDF, ND-SDF, GSRec) under the same environment and refer to reported results for the others.

A. Experiment Result

The quantitative results on three datasets are reported in Table I and Table II. Our method, GPU-SDF, achieves state-of-the-art performance across these benchmarks. On global metrics, the numerical improvements are modest (e.g., an F-Score increase of 0.3 on ScanNet). This outcome is expected and highlights an inherent limitation of such metrics: they are dominated by large, low-frequency surfaces like walls and floors, which constitute the vast majority of the scene's surface area. Our approach, however, is specifically designed to excel at reconstructing complex, high-frequency details such as chair legs and railings. While these critical local improvements have a limited impact on the global score, their importance is undeniable for high-fidelity reconstruction.

TABLE I

Quantitative performance on ScanNet/Replica. Bold indicates the best performance.

Method	Prior	Acc↓	Comp↓	Chamfer↓	F-score↑
COLMAP	×	4.7/3.0	23.5/9.5	14.1/6.3	53.7/65.8
Unisurf	×	55.4/4.5	16.4/5.3	35.9/4.9	26.7/78.9
VolSDF	×	41.4/4.4	12.0/8.3	26.7/7.2	34.6/69.5
GSRec	D+N	6.7/4.1	6.9/7.7	6.8/5.8	59.9/69.2
NeuRIS	N	5.0/3.4	4.9/7.0	5.0/5.2	69.2/75.4
MonoSDF	D+N	3.6/2.7	3.9/3.1	3.8/2.9	77.1/86.1
RS-Recon	D+N	4.0/2.7	4.0/2.5	3.6/2.6	79.4/91.7
NeRFPrior†	D+N	3.7/–	4.2/–	–/3.8	78.2/81.3
Deb-SDF	D+N	3.6/2.8	4.0/3.0	3.8/2.9	78.5/88.3
H2O-SDF†	D+N	3.2/–	3.7/–	3.5/–	79.9/–
ND-SDF	D+N	3.1/2.7	3.6/2.8	3.4/2.6	82.0/91.6
Ours	D+N	**3.1/2.6**	**3.5/2.5**	**3.3/2.5**	**82.3/92.4**

Note: "–" denotes unavailable metrics. As of our evaluation, the source code for NeRFPrior and H2O-SDF was not publicly available, and their original papers did not report results on the ScanNet/Replica datasets.

Qualitative analysis provides a complement to global metrics. As shown in Fig. 5, our method yields clearer reconstructions of intricate structures. For example, in regions with high geometry uncertainty, details such as chair legs—which are often missing or fragmented in other reconstructions—are recovered with higher clarity and completeness. This improved ability to capture geometric detail directly supports the core motivation of our work and demonstrates the effectiveness of our method in enhancing reconstruction fidelity where conventional approaches often struggle.

B. Ablation Studies

We conducted a series of ablation experiments on the ScanNet++ dataset to evaluate the contributions of each

TABLE II

QUANTITATIVE PERFORMANCE ON SCANNET++. B.A. DENOTES BAKED-ANGELO. BOLD INDICATES THE BEST PERFORMANCE.

Methods	Priors	Metrics	Avg.	0e75f 3c4d9	036bc e3393	7f4d1 73c9c
VolSDF	×	Acc ↓	6.8	5.3	5.3	9.7
		Chamfer ↓	7.6	6.5	7.6	8.6
		Prec ↑	44.2	39.5	47.5	45.7
B.A.	×	Acc ↓	23.6	12.1	16.8	42.0
		Chamfer ↓	13.5	8.7	9.6	22.2
		Prec ↑	50.1	52.5	52.0	45.7
MonoSDF	D+N	Acc ↓	8.0	6.4	3.7	13.8
		Chamfer ↓	5.4	4.6	3.7	8.0
		Prec ↑	62.9	57.4	60.6	70.8
ND-SDF	D+N	Acc ↓	6.8	5.4	3.7	11.2
		Chamfer ↓	4.5	3.5	**3.4**	6.4
		Prec ↑	67.7	66.2	64.8	72.1
Ours	D+N	Acc ↓	**6.1**	**4.2**	**3.7**	**10.3**
		Chamfer ↓	**4.3**	**3.4**	3.6	**6.0**
		Prec ↑	**68.9**	**67.9**	**64.8**	**73.9**

TABLE III

ABLATION STUDIES RESULTS FOR KEY COMPONENTS OF GPU-SDF.BOLD INDICATES THE BEST PERFORMANCE.

Method	Acc. ↓	Chamfer. ↓	Prec.↑
ND-SDF	11.2	6.4	72.1
+D.U.(Horiz.)	10.8	6.3	73.1
+D.U.(Vert.)	10.6	6.2	72.8
+D.U.(Full)	10.7	6.2	73.3
+D.U.+N.U.	10.4	6.0	73.7
w/o EDF	11.0	6.4	72.7
w/o MC	10.8	6.3	72.7
Full Model	**10.3**	**6.0**	**73.9**
MonoSDF	13.8	8.0	70.8
MonoSDF+Ours	**11.4**	**6.9**	**71.6**

key component in our framework. Quantitative results are summarized in Table III.

Effectiveness of Prior Uncertainty. We first analyze the design of the prior uncertainty estimation module, using ND-SDF as the baseline and incrementally adding different strategies. We denote depth uncertainty as *D.U.* and normal uncertainty as *N.U.* The results show that even when using only depth uncertainty derived from a single type of data augmentation (*+D.U.(Horiz.)* or *+D.U.(Vert.)*), reconstruction accuracy improves. Combining both horizontal and vertical flips (*+D.U.(Full)*) produces more reliable uncertainty estimates, as illustrated in Fig. 3, and further enhances reconstruction quality. The largest gain is achieved by combining both depth and normal uncertainty (*+D.U.+N.U.*), which reduces the Chamfer distance from the baseline value of 0.064 to 0.060 and increases precision from 72.1 to 73.7. These results highlight that our uncertainty modeling is an essential component for guiding the reconstruction process.

Impact of Additional Geometric Constraints. We evaluate the necessity of the two geometric constraints introduced for high-uncertainty regions: the edge distance field (EDF) and multi-view consistency (MC). We conduct two controlled comparisons, removing EDF (*w/o EDF*) and removing MC (*w/o MC*), against full model. As shown in Table III, removing either constraint causes a performance drop, confirming that both are essential for suppressing artifacts in under-constrained regions and preserving structural integrity. The visualization results are shown in Fig. 6.

Fig. 6. Visualization results of ablation studies, with details highlighted in red boxes. Full Model achieves the finest detail in thin structures, demonstrating the effectiveness of our method.

Generalization as a Plug-in Module. Finally, we test our framework as a plug-in that integrates seamlessly into other neural SDF pipelines to enhance their performance. To this end, we apply our method to another mainstream baseline, MonoSDF [4]. The results are compelling: after integration (*MonoSDF+Ours*), all metrics show substantial improvements. In particular, the Chamfer distance decreases from 0.080 to 0.069, and Precision increases from 70.8 to 71.6. This experiment demonstrates the strong generalization ability of our approach, indicating that it is not tailored to a specific architecture but serves as a versatile module for improving neural surface reconstruction quality.

V. CONCLUSION AND DISCUSSION

In this paper, we introduce GPU-SDF, a novel neural implicit indoor surface reconstruction method. Our core contributions are twofold. First, we introduce a self-supervised uncertainty estimation with guided supervision, allowing reliable use of geometric priors while preserving information in high-uncertainty regions. Second, to address the resulting under-constrained optimization, we further incorporate an edge distance field and multi-view consistency regularization, which enhance thin structures and fine details. Experiments on challenging benchmarks verified the effectiveness of GPU-SDF and its versatility as a plug-in to existing SDF methods, enabling more robust and reliable 3D reconstruction with imperfect priors. However, the regions corresponding to unseen viewpoints remain unconstrained, and future work will focus on enhancing the reconstruction quality of regions that are occluded or not visible to the camera.

REFERENCES

[1] 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," *CACM*, vol. 65, no. 1, pp. 99–106, 2021.

[2] B. Kerbl, G. Kopanas, T. Leimkühler, and G. Drettakis, "3d gaussian splatting for real-time radiance field rendering," *ACM Transactions on Graphics*, vol. 42, no. 4, pp. 1–14, 2023.

[3] B. Curless and M. Levoy, "A volumetric method for building complex models from range images," in *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques, August 4-9, 1996, New Orleans, LA, USA*, 1996, pp. 303–312.

[4] Z. Yu, S. Peng, M. Niemeyer, T. Sattler, and A. Geiger, "Monosdf: Exploring monocular geometric cues for neural implicit surface reconstruction," *NeurIPS*, vol. 35, pp. 25 018–25 032, 2022.

[5] X. Lyu, P. Dai, Z. Li, D. Yan, Y. Lin, Y. Peng, and X. Qi, "Learning a room with the occ-sdf hybrid: Signed distance function mingled with occupancy aids scene representation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision, October 1-6, 2023, Paris, France*, 2023, pp. 8940–8950.

[6] 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 *European Conference on Computer Vision, October 23-27, 2022, Tel Aviv,Israel*. Springer, 2022, pp. 139–155.

[7] Y. Xiao, J. Xu, Z. Yu, and S. Gao, "Debsdf: Delving into the details and bias of neural indoor scene reconstruction," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[8] Z. Tang, W. Ye, Y. Wang, D. Huang, H. Bao, T. He, and G. Zhang, "Nd-sdf: Learning normal deflection fields for high-fidelity indoor reconstruction," *arXiv preprint arXiv:2408.12598*, 2024.

[9] T. Deng, G. Shen, X. Chen, S. Yuan, H. Shen, G. Peng, Z. Wu, J. Wang, L. Xie, D. Wang, H. Wang, and W. Chen, "Mcn-slam: Multi-agent collaborative neural slam with hybrid implicit neural scene representation," *arXiv preprint arXiv:2506.18678*, 2025.

[10] T. Deng, G. Shen, C. Xun, S. Yuan, T. Jin, H. Shen, Y. Wang, J. Wang, H. Wang, D. Wang *et al.*, "Mne-slam: Multi-agent neural slam for mobile robots," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 1485–1494.

[11] L. Yariv, J. Gu, Y. Kasten, and Y. Lipman, "Volume rendering of neural implicit surfaces," *Advances in Neural Information Processing Systems*, vol. 34, pp. 4805–4815, 2021.

[12] 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," *Advances in Neural Information Processing Systems*, vol. 34, pp. 27 171–27 183, 2021.

[13] Z. Li, T. Müller, A. Evans, R. H. Taylor, M. Unberath, M.-Y. Liu, and C.-H. Lin, "Neuralangelo: High-fidelity neural surface reconstruction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 17-24, 2023, Vancouver, BC, Canada*, 2023, pp. 8456–8465.

[14] T. Wu, J. Wang, X. Pan, X. Xudong, C. Theobalt, Z. Liu, and D. Lin, "Voxurf: Voxel-based efficient and accurate neural surface reconstruction, may 1-5, 2023, kigali, rwanda," in *The Eleventh International Conference on Learning Representations*, 2022, pp. 13 572–13 575.

[15] M. Oechsle, S. Peng, and A. Geiger, "Unisurf: Unifying neural implicit surfaces and radiance fields for multi-view reconstruction," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 5589–5599.

[16] W. E. Lorensen and H. E. Cline, "Marching cubes: A high resolution 3d surface construction algorithm," in *Seminal graphics: pioneering efforts that shaped the field*, 1998, pp. 347–353.

[17] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 18-24, 2022, New Orleans, LA, USA*, 2022, pp. 5511–5520.

[18] S. Zhi, T. Laidlow, S. Leutenegger, and A. J. Davison, "In-place scene labelling and understanding with implicit scene representation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision, July 21-26, 2017, Honolulu, HI, USA*, 2021, pp. 15 838–15 847.

[19] P. Zins, Y. Xu, E. Boyer, S. Wuhrer, and T. Tung, "Multi-view reconstruction using signed ray distance functions (srdf)," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 17-24, 2023, Vancouver, BC, Canada*, 2023, pp. 16 696–16 706.

[20] X. Long, C. Lin, L. Liu, Y. Liu, P. Wang, C. Theobalt, T. Komura, and W. Wang, "Neuraludf: Learning unsigned distance fields for multi-view reconstruction of surfaces with arbitrary topologies," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 17-24, 2023, Vancouver, BC, Canm*, 2023, pp. 20 834–20 843.

[21] Y. Wang, D. Huang, W. Ye, G. Zhang, W. Ouyang, and T. He, "Neurodin: A two-stage framework for high-fidelity neural surface reconstruction," *Advances in Neural Information Processing Systems*, vol. 37, pp. 103 168–103 197, 2025.

[22] Z. Liu, B. Yang, Y. Luximon, A. Kumar, and J. Li, "Raydf: Neural ray-surface distance fields with multi-view consistency," *Advances in Neural Information Processing Systems*, vol. 36, pp. 14 689–14 691, 2024.

[23] D. Azinović, R. Martin-Brualla, D. B. Goldman, M. Nießner, and J. Thies, "Neural rgb-d surface reconstruction," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 18-24, 2022, New Orleans, LA, USA*, 2022, pp. 6290–6301.

[24] Z. Zhu, S. Peng, V. Larsson, Z. Cui, M. R. Oswald, A. Geiger, and M. Pollefeys, "Nicer-slam: Neural implicit scene encoding for rgb slam," in *International Conference on 3D Vision (3DV)*, March 2024.

[25] B. Huang, Z. Yu, A. Chen, A. Geiger, and S. Gao, "2d gaussian splatting for geometrically accurate radiance fields," *SIGGRAPH*, 2024.

[26] D. Chen, H. Li, W. Ye, Y. Wang, W. Xie, S. Zhai, N. Wang, H. Liu, H. Bao, and G. Zhang, "Pgsr: Planar-based gaussian splatting for efficient and high-fidelity surface reconstruction," *IEEE Transactions on Visualization and Computer Graphics*, 2024.

[27] Q. Wu, J. Zheng, and J. Cai, "Surface reconstruction from 3d gaussian splatting via local structural hints," in *European Conference on Computer Vision*. Springer, 2024, pp. 441–458.

[28] H. Xiang, X. Li, K. Cheng, X. Lai, W. Zhang, Z. Liao, L. Zeng, and X. Liu, "Gaussianroom: Improving 3d gaussian splatting with sdf guidance and monocular cues for indoor scene reconstruction," in *2025 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2025, pp. 2686–2693.

[29] A. Guédon and V. Lepetit, "Sugar: Surface-aligned gaussian splatting for efficient 3d mesh reconstruction and high-quality mesh rendering," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 5354–5363.

[30] G. Bae, I. Budvytis, and R. Cipolla, "Irondepth: Iterative refinement of single-view depth using surface normal and its uncertainty," *arXiv preprint arXiv:2210.03676*, 2022.

[31] R. Marsal, F. Chabot, A. Loesch, W. Grolleau, and H. Sahbi, "Monoprob: self-supervised monocular depth estimation with interpretable uncertainty," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024, pp. 3637–3646.

[32] M. Poggi, F. Aleotti, F. Tosi, and S. Mattoccia, "On the uncertainty of self-supervised monocular depth estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 3227–3237.

[33] J. Hornauer and V. Belagiannis, "Gradient-based uncertainty for monocular depth estimation," in *European Conference on Computer Vision*. Springer, 2022, pp. 613–630.

[34] 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, October 10-17, 2021, Montreal, QC, Canada*, 2021, pp. 10 786–10 796.

[35] G. Bae, I. Budvytis, and R. Cipolla, "Estimating and exploiting the aleatoric uncertainty in surface normal estimation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 13 137–13 146.

[36] X. Soria, Y. Li, M. Rouhani, and A. D. Sappa, "Tiny and efficient model for the edge detection generalization," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 1364–1373.

[37] J. Shan, Y. Li, L. Yang, Q. Feng, L. Han, and H. Wang, "Dds-slam: Dense semantic neural slam for deformable endoscopic scenes," in *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2024, pp. 10 837–10 842.

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

[39] 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," (2019-06-13)[2024-05-20]. [Online]. Available: https://arxiv.org/abs/1906.05797

[40] C. Yeshwanth, Y.-C. Liu, M. Nießner, and A. Dai, "Scannet++: A high-fidelity dataset of 3d indoor scenes," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 12–22.

[41] J. L. Schonberger and J.-M. Frahm, "Structure-from-motion revisited," in *CVPR*, 2016, pp. 4104–4113.

[42] L. Yariv, P. Hedman, C. Reiser, D. Verbin, P. P. Srinivasan, R. Szeliski, J. T. Barron, and B. Mildenhall, "Bakedsdf: Meshing neural sdfs for real-time view synthesis," in *ACM SIGGRAPH 2023 conference proceedings*, 2023, pp. 1–9.

[43] R. Yin, Y. Chen, S. Karaoglu, and T. Gevers, "Ray-distance volume rendering for neural scene reconstruction," in *European Conference on Computer Vision*. Springer, 2024, pp. 377–394.

[44] W. Zhang, E. Y.-t. Jia, J. Zhou, B. Ma, K. Shi, Y.-S. Liu, and Z. Han, "Nerfprior: Learning neural radiance field as a prior for indoor scene reconstruction," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 11 317–11 327.

[45] M. Park, M. Do, Y. J. Shin, J. Yoo, J. Hong, J. Kim, and C. Lee, "H2o-sdf: Two-phase learning for 3d indoor reconstruction using object surface fields," in *The Twelfth International Conference on Learning Representations*.