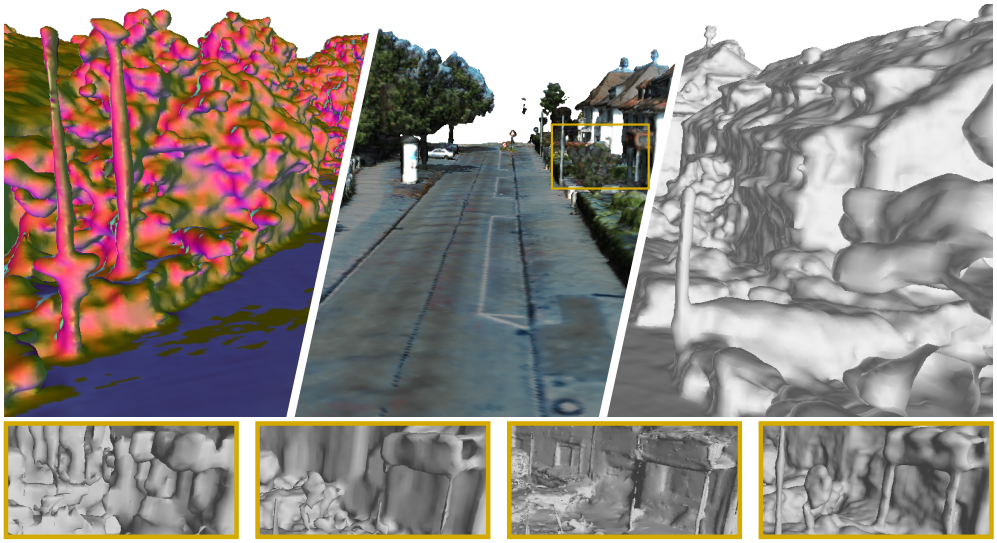
J-NeuS: Joint field optimization for Neural Surface reconstruction in urban scenes with limited image overlap

Fusang Wang[1] Hala Djeghim[2] Fabien Moutarde[1] Désiré Sidibé[2]
[1]CAOR, Mines-Paris PSL, France
[2]IBISC, Evry Paris-Saclay University, France



StreetSurf [12] ViiNeuS [9] GoF [40] J-NeuS (ours)

Figure 1. We introduce **J-NeuS**, a novel hybrid implicit surface reconstruction method specifically designed for large-scale driving sequences with limited camera overlap. Extensive experiments on four major driving datasets demonstrate the superiority of J-NeuS's mesh (top, from left to right: mesh normals, textured mesh and shaded mesh) over previous state-of-the-art methods (bottom).

Abstract

*Reconstructing the surrounding surface geometry from recorded driving sequences poses a significant challenge due to the limited image overlap and complex topology of urban environments. SoTA neural implicit surface reconstruction methods often struggle in such setting, either failing due to small vision overlap or exhibiting suboptimal performance in accurately reconstructing both the surface and fine structures. To address these limitations, we introduce **J-NeuS**, a novel hybrid implicit surface reconstruction method for large driving sequences with outward facing camera poses. J-NeuS leverages cross-representation uncertainty estimation to tackle ambiguous geometry caused by limited observations. Our method performs joint optimization of two radiance fields in addition to guided sampling achieving accurate reconstruction of large areas along with fine structures in complex urban scenarios. Extensive evaluation on major driving datasets demonstrates the superiority of our approach in reconstructing large driving sequences with limited image overlap, outperforming concurrent SoTA methods.*

1. Introduction

Accurate 3D surface reconstruction of large urban scenes is essential for many challenging autonomous driving applications, such as scene relighting [22], sensor simulation [37], and 3D object insertion [34]. However, achieving high-quality reconstructions in driving environments remains a significant challenge, where the difficulty primarily stems from two key factors:

- **Complex outdoor geometry**: Urban scenes often feature arbitrary object arrangements, including large textureless planar surfaces and intricate fine structures.
- **Geometric ambiguity from outward facing sensors**: Limited camera overlap and the linear trajectory of vehicles introduce significant uncertainty in the reconstruction process.

Neural Radiance Fields (NeRF) [18] have emerged as a

powerful approach for 3D scene reconstruction in such settings. Leveraging differentiable volume rendering, NeRF enables accurate reconstruction from images and their associated camera poses, effectively avoiding the error accumulation typically seen in traditional multi-view stereo pipelines. Although NeRF-based methods have demonstrated impressive results in capturing high frequency scene details [1, 20], its under-constrained optimization problem leads to bad scene geometry, especially when vision cues are limited [5, 21]. To adapt to large scenes with complex structures, either 3D priors, such as LiDAR pointcloud, or strong assumptions are introduced to constrain the original optimization problem [8, 23, 28, 29].

Neural implicit surface methods [30, 31, 38], overcome the limitation of NeRF by replacing the volumetric density field with a Signed Distance Function (SDF). The SDF formulation represents the surface at its zero level sets and is regularized using the Eikonal constraint. Current neural SDF methods demonstrate high fidelity reconstruction quality in object-centric scenes with large texture-less surfaces [15, 39]. However prevalent SDF methods like Neuralangelo [15] work better on landmark-centric scenes with large image overlaps but fail to reconstruct urban scenes captured by vehicle-mounted cameras due to limited observation overlap [7, 9]. Moreover, SDF-based methods often struggle to preserve fine structural details due to biased depth estimation and over-regularization of geometric constraints [32, 41], an issue that is critical for downstream autonomous driving applications.

To achieve high-quality surface reconstruction for autonomous driving – *capturing both fine structures and large surfaces* – recent approaches integrate volumetric and SDF representations. These methods partition scenes into distinct regions and apply specialized reconstruction [12] or sampling strategies [27, 35] tailored to each region's unique characteristics. Additionally, other methods [9] demonstrate that SDF representations can benefit from volumetric initialization, enabling faster convergence and improved geometric fidelity. However, these solutions often result in suboptimal geometry quality due to overly strong geometric assumptions about the scene or reliance on coarse initialization that introduce noise from volumetric prediction.

In order to accurately combine the strengths of both volumetric and SDF representation, we propose an uncertainty estimation framework that identifies a noisy predictions arising from geometric ambiguity. Specifically, we estimate two types of uncertainty – *photometric uncertainty and geometric uncertainty* – to jointly train the volumetric and SDF models. These uncertainty measures guide the sampling process, ensuring that each representation is deployed where it performs better while mitigating over-regularization in the SDF to facilitate fine structure learning. Our approach effectively handles the complex geometry of urban environments, enabling efficient rendering and precise surface reconstruction. We evaluate our proposed solution on four public driving datasets: KITTI-360 [16], Pandaset [36], Waymo Open Dataset [26], and nuScenes [2], demonstrating robust reconstructions of intricate urban geometry with limited image overlap.

The main contributions of our method are the following:

- **Joint optimization of NeRF and SDF**: We propose a framework that fuses volumetric (NeRF) and surface (SDF) representations so each excels in regions where it is better.
- **Guided Ray Sampling**: We introduce a novel sampling strategy that leverages cross-representation uncertainty to tackle ambiguous geometric cues, enabling faster and more accurate surface reconstruction.
- **Adaptive Relaxation on geometry regularization**: We dynamically relax the Eikonal constraint and monocular cues in uncertain regions to avoid over-smoothing, ensuring complete fine-structure reconstruction.

2. Related work

Neural implicit surface reconstruction Traditional surface reconstruction techniques, such as Multi-View Stereo (MVS) [19], have long been the cornerstone of 3D reconstruction tasks. However, their multi-step pipelines are prone to error accumulation, often resulting in incomplete or inaccurate reconstructions. In contrast, neural implicit surface reconstruction approaches adapt volume rendering for more accurate surface estimation [30, 38], reducing the need for manual post-processing. Subsequent advances target faster training [31] or to and more complex geometry [15, 32]. Notably, Neuralangelo [15] and Neurodin [32] achieve highly fidelity reconstruction of complex geometries in object-centric scenes where large observation overlaps are present. However, they face significant challenges when applied to autonomous driving scenarios [7, 9], where the camera trajectory is linear and image overlap is limited.

Urban outdoor surface reconstruction. To enhance surface reconstruction in autonomous driving scenes with limited image overlap, recent methods typically rely on 3D supervision [24, 34], strong geometric priors [12, 28], or monocular supervision [9, 12]. While these approaches effectively model large planar areas, they often struggle to capture fine details due to strong geometric assumptions [28], inconsistencies in monocular predictions [12] or over-regularization on intricate features [9].

Recent approaches have also explored 3D Gaussian splatting for surface reconstruction [3, 4, 11, 42]. These methods, however, often rely on heavy post-processing techniques, such as Poisson Surface Reconstruction [14], to extract the final surface, adding computational complexity. GoF [40] and StreetSurfGS [7] offer a direct approach for

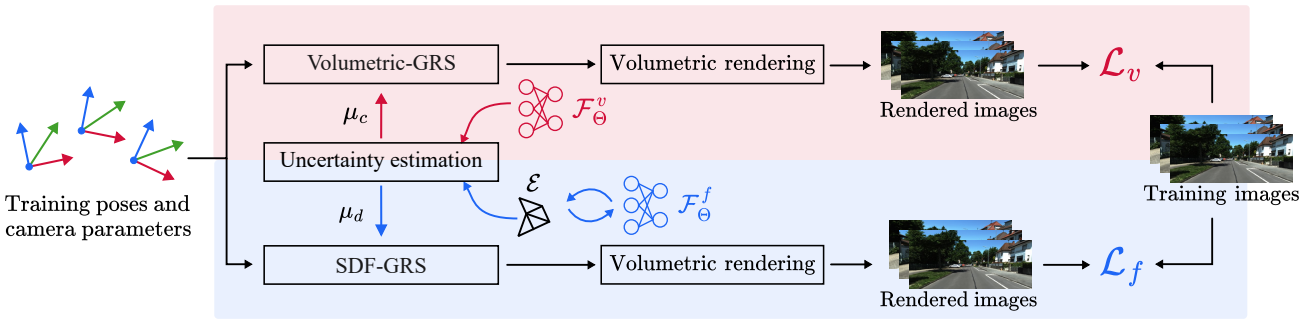


Figure 2. **Overview of J-NeuS**: we jointly train two implicit model: a volumetric representation $\mathcal{F}^v_\Theta$ and a SDF $\mathcal{F}^f_\Theta$ with a mutual guidance provided through our Guided Ray Sampling strategy. A colorized mesh of the scene $\mathcal{E}$ is periodically extracted from the SDF representation.

mesh extraction by explicitly learning level sets. However, they still face memory bottlenecks in large-scale scenes and require TSDF fusion together with additional mesh cleaning to achieve acceptable mesh quality. Unlike HUGS [42], which targets dynamic element decoupling, we focus on static geometric fidelity under limited-overlap conditions. Consequently, we adopt GoF[40], a representative SOTA for direct GS-based extraction, as our primary baseline.

Hybrid scene representation. To overcome the limitations of SDF methods in capturing fine structures, recent approaches integrate volumetric and SDF representations by dividing scenes into distinct regions and applying specialized reconstruction or sampling strategies. Turki et al. [27] and Wang et al. [35] segment scenes based on the scaling factor used to convert SDF into density, and employ distinct sampling techniques for volumetric and surface regions to enable real-time, high-quality rendering. StreetSurf [12] models different regions – *close, far, and sky* – using different hierarchical space partitioning (e.g., 3D/4D hash grids, occupancy grids). This improves performance in urban scenes but imposes strong priors tied to vehicle ego poses, limiting generalizability. Meanwhile, ViiNeuS [9] initializes SDF sampling with volumetric density before gradually transitioning to a pure SDF representation, accelerating convergence and achieving state-of-the-art results on autonomous driving benchmarks. However, this approach risks introducing artifacts from the volumetric representation that can adversely affect the final SDF reconstruction.

Building on these insights, we propose a more adaptive divide-and-conquer technique that dynamically partitions the scene based on uncertainty across both representations. By applying tailored sampling and regularization strategies for different region, our method preserves fine details while effectively handling the large planar areas characteristic of autonomous driving environments.

3. Method

Given a collection of RGB images captured from a moving vehicle in an urban area with limited overlap, our goal is to resolve perspective geometry ambiguities introduced by partial scene observations and to accurately reconstruct surfaces with precise structural details. To achieve this, we propose selectively fusing volumetric and SDF representations, with uncertainty guiding the division of labor between the two representations. Specifically, we use two implicit representations: a volumetric radiance field $\mathcal{F}^v_\Theta$ and an SDF field $\mathcal{F}^f_\Theta$ with trainable weights Θ_v and Θ_f, respectively. We also extract at regular intervals during training a colorized mesh $\mathcal{E}$ from $\mathcal{F}^f_\Theta$ using the marching cubes algorithm. An overview of our method is presented in Fig. 2

To selectively fuse NeRF and SDF representations during training, we first introduce photometric and geometric uncertainty estimation across the two representations $\mathcal{F}^v_\Theta$ and $\mathcal{F}^f_\Theta$ (Section. 3.1). We then jointly optimize both representations with Guided Ray Sampling based on uncertainty estimation to leverage the strengths of both representations (Section. 3.2). Additionally, to ensure the preservation of fine structures, we relax the Eikonal constraint and monocular supervision in uncertain regions to avoid over-regularization of the SDF field (Section. 3.3).

3.1. Cross Representation Uncertainty Estimation

Due to the inherent partial observations in autonomous driving scenarios, both volumetric and SDF representations exhibit epistemic uncertainty in regions where visual cues are sparse. These regions are characterized by high variance in RGB and depth predictions, or deviations from the ground truth. To fully leverage the strengths of both representations, it is crucial to identify areas with high uncertainty and adaptively apply tailored sampling and regularization strategies. In the following sections, we first introduce the key concepts and notation related to our volumetric and SDF representations. We then describe how uncertainty is estimated to selectively fuse both representations effectively.

Background - Implicit volumetric representation. A volumetric radiance field is a continuous function $\mathcal{F}^v_\theta$ mapping a position and direction pair$(\mathbf{x}, \mathbf{u}) \in \mathbb{R}^3 \times \mathbb{S}^2$ to a volume density $\sigma \in \mathbb{R}^+$ and a color $\mathbf{c} \in [0, 1]^3$. Mildenhall

et al. [18] model this function with a multi layer perceptron (MLP) whose weights θ are optimized to reconstruct a 3D scene given a set of posed images during training.

To render an image and depth, volume rendering is applied to alpha-composite the color for each ray, yielding the final pixel color $\hat{C}_v(\mathbf{r}) \in \mathbb{R}^3$ and the depth value $\hat{D}_v(\mathbf{r}) \in \mathbb{R}^+$ to be:

$$\hat{C}_v(\mathbf{r}) = \sum_{i=1}^{N} w_i c_i, \quad w_i = T_i \alpha_i, \tag{1}$$

$$\hat{D}_v(\mathbf{r}) = \sum_{i=1}^{N} w_i z_i, \quad w_i = T_i \alpha_i, \tag{2}$$

where $T_i = \prod_{j=1}^{i-1}(1-\alpha_j)$ is the accumulated transmittance, $\alpha_i \in \mathbb{R}$ the blending factor and $z_i \in \mathbb{R}^+$ is the distance of the sample to the camera center. Here, α_i is computed from the predicted density: $\alpha_i = 1 - \exp(-\sigma_i \delta_i)$, with $\delta_i \in \mathbb{R}^+$ being the distance between samples along the ray.

Background - Implicit surface representation. While the volumetric field predicts a pointwise 3D density $\sigma(x)$ at each location x, the SDF instead predicts a signed distance function $f(x)$ and converts it to density using a logistic function $\phi_s(f)$ with a global scale parameter s, effectively confining the density to a narrow band of width $O(1/s)$ around the surface [35]. To enable volume rendering, the signed distance function f is then transformed into the density σ using sigmoid-shape mapping for alpha compositing. NeuS [30] adopted a new formulation of the blending factor:

$$\alpha_i = \max\left(\frac{\Phi_s(f(p_i)) - \Phi_s(f(p_{i+1}))}{\Phi_s(f(p_i))}, 0\right), \tag{3}$$

where $f(p_i)$ and $f(p_{i+1})$ are signed distance values at section points centered on x_i, $\Phi_s(x)$ the sigmoid function.

Uncertainty estimation. When reconstructing urban scenes, the direct 3D point-based density prediction of the volumetric field enables rapid fitting of high-frequency geometry (e.g., poles), but may also introduce spurious density (floaters) on large planar surfaces [30]. In contrast, the more constrained density prediction of the SDF promotes surface continuity, but may over-smooth fine details when using a uniform or overly small scale parameter s. To harness the complementary strengths of both representations, it is essential to identify and suppress their unreliable predictions. To this end, we introduce two complementary uncertainty estimates: **geometric uncertainty** (μ_d) and **photometric uncertainty** (μ_c).

The first type of uncertainty focuses on identifying regions with ambiguous geometric information: areas where visual cues are insufficient for geometry reconstruction (*e.g.*, textureless surfaces or partially observed complex structures). Ideally, one would measure epistemic uncertainty by comparing predictions with dense 3D ground truth. However, in autonomous driving scenarios, such ground truth is often unavailable. Therefore, we propose to measure uncertainty by evaluating the consistency between the two representations: Given a ray r, we introduce the geometry uncertainty based on the rendered depth from $\mathcal{F}_\Theta^v$, and the distance to the first intersection with the mesh $\mathcal{E}$, $\hat{D}_\mathcal{E}(\mathbf{r})$:

$$\mu_d(\mathbf{r}) = |1 - \frac{\hat{D}_\mathcal{E}(\mathbf{r})}{\hat{D}_v(\mathbf{r})}|. \tag{4}$$

High values of μ_d indicate inconsistency between the volumetric and SDF models, suggesting that at least one representation is uncertain. In such cases, we preferentially rely on the volumetric field for two key reasons. First, during initialization, it captures fine structures more completely and achieves faster convergence [9]. Second, in later stages, probability sampling 3.4 of $\mathcal{F}_\Theta^f$ can effectively filter out noise introduced by the volumetric field $\mathcal{F}_\Theta^v$ while preserving the fine details that $\mathcal{F}_\Theta^v$ initially localized.

Secondly, we introduce a photometric-based uncertainty to evaluate the predictions of the SDF field. We hypothesize that if the learned SDF field is accurate, **a single-sample rendering** at the mesh depth should yield a precise photometric rendering. Casting a ray $\mathbf{r} = \mathbf{o} + t\mathbf{u}$ from the camera center $\mathbf{o}$ through the pixel along direction $\mathbf{u}$, we define the photometric uncertainty indicator as:

$$\mu_c(\mathbf{r}) = |\hat{C}_\mathcal{E}(\mathbf{r}) - \bar{C}(\mathbf{r})|, \tag{5}$$

where $\bar{C}$ indicates rgb ground truth, and $\hat{C}_\mathcal{E}(\mathbf{r})$ represents the rgb value of $\mathcal{F}_\Theta^f$ at the point $\mathbf{o} + \hat{D}_\mathcal{E}(\mathbf{r})\mathbf{u}$, i.e. the point on the Mesh $\mathcal{E}$ of the given ray. $\mu_c(\mathbf{r})$ is used to divide certain and uncertain regions in the SDF representation (see Section. 3.3). We use it to guide the sampling strategy of the $\mathcal{F}_\theta^v$ for certain regions, and adaptively relax geometric regularization in uncertain regions to prevent over-smoothing. While μ_c does not guarantee accurate depth alignment, it is sufficient to quickly populate large planar regions without requiring complete rendering of $\mathcal{F}_\theta^v$. We provide a visualization of μ_c and μ_d in Figure 3.

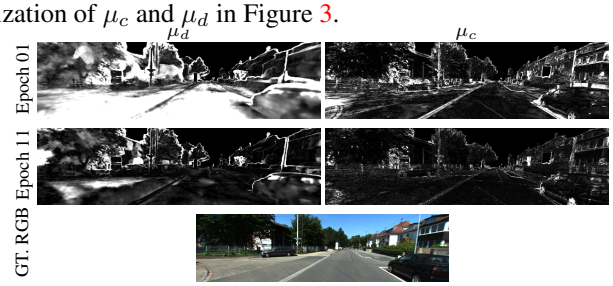


Figure 3. Visualization of μ_d and μ_c, maximum clamped at 0.3.

3.2. Guided Ray Sampling (GRS)

To selectively fuse the volumetric and SDF representation during training, we propose to jointly train $\mathcal{F}_\Theta^v$ and $\mathcal{F}_\Theta^f$ via Guided Ray Sampling (GRS) based on previously introduced uncertainty μ_c and μ_d. Specifically, for the volumetric representation, we use μ_c to guide the sampling toward the surface of the mesh $\mathcal{E}$ where SDF field $\mathcal{F}_\Theta^f$ is confident. For the SDF representation, we focus the sampling to the estimated surface of the best-suited representation, using geometric uncertainty μ_d.

Volumetric-GRS. We consider a batch of rays $r \in \mathcal{R}_v$ with for each ray r the corresponding original sampling interval $[t_n, t_f]$. To improve volumetric reconstruction on planar surfaces, we adjust the sampling bounds to focus more on the estimated depth of the Mesh $D_\mathcal{E}(\mathbf{r})$ if the $\mu_c(\mathbf{r})$ is below a certain threshold τ_c:

$$[t_n, t_f](\mathbf{r}) = \begin{cases} [0, D_\mathcal{E}(\mathbf{r}) + \delta] & \text{if } \mu_c < \tau_c, \\ [0, \infty[& \text{otherwise,} \end{cases} \tag{6}$$

where δ is a hyperparameter (analogous to the shell factor in [35]) periodically updated. This helps avoid noisy planar predictions that may arise from the under-constrained optimization of the volumetric representation, ensuring more accurate surface reconstruction. In practice, τ_c is set as a constant for each dataset (see Section. 2 in supplementary material).

SDF-GRS. After rendering $\mathcal{F}_\Theta^v$ with guided sampling, the predicted depth $\hat{D}_v(\mathbf{r})$ becomes available, allowing us to estimate the geometric uncertainty μ_d to direct the sampling of $\mathcal{F}_\Theta^f$. Since high μ_d values indicates regions where the predictions of the two representations are inconsistent, it is crucial to decide which representation to trust. Here, we propose trusting the volumetric representation, as it typically learns faster and captures complex structures more effectively and expect that noisy predictions from $\mathcal{F}_\Theta^v$ will be handled by the Volumetric-GRS.

Similar to Volumetric-GRS, we adjust the sampling bounds based on the geometric uncertainty $\mu_d(\mathbf{r})$. If $\mu_d(\mathbf{r})$ exceeds a threshold τ_d, we focus the sampling on the estimated depth from the **volumetric field** $\hat{D}_v(\mathbf{r})$. Otherwise, sampling is concentrated around the depth of the mesh $D_\mathcal{E}(\mathbf{r})$, ensuring accurate surface reconstruction in both cases:

$$[t_n, t_f](\mathbf{r}) = \begin{cases} [D_\mathcal{E}(\mathbf{r}) - \delta, D_\mathcal{E}(\mathbf{r}) + \delta] & \text{if } \mu_d < \tau_d, \\ [D_v(\mathbf{r}) - \delta, D_v(\mathbf{r}) + \delta] & \text{otherwise,} \end{cases} \tag{7}$$

with δ being the same shell factor hyperparameter described in the previous paragraph.

Adaptive thresholding. The SDF-GRS strategy employs a threshold τ_d to classify rays as "certain" or "uncertain," but selecting an appropriate value a priori can be difficult across varying scenes and reconstruction tasks. To overcome this, we propose a lightweight, data-driven algorithm that adaptively adjusts τ_d based on the observed ratio of certain to uncertain rays (see Algorithm 1). Empirically, this adaptive scheme generalizes robustly across diverse urban environments (see supplementary material for details).

Algorithm 1: Adaptive Uncertainty Threshold τ

Input: τ, $\{\mu_d(r_i)\}_{i=1}^N$: depth uncertainties of N rays
Data: $\gamma_\uparrow, \gamma_\downarrow$: growth/decay factors; $\rho_{\text{high}}, \rho_{\text{low}}$: ratio thresholds
Output: Updated τ
$u \leftarrow \sum_i [\mu_d(r_i) > \tau]$; $c \leftarrow (N - u)$; $\rho \leftarrow (u/c)$;
if $\rho > \rho_{high}$ **then**
  $\tau \leftarrow \tau \times \gamma_\uparrow$;
else if $\rho < \rho_{low}$ **then**
  $\tau \leftarrow \tau \times \gamma_\downarrow$;
return τ

3.3. Adaptive Relaxation on Geometric Constrain

Another critical factor in capturing fine structures within the SDF field is avoiding over-regularization while the geometry remains under-optimized [32, 33]. Whereas previous urban reconstruction methods enforce the Eikonal constraint across the entire scene [9, 12], we adaptively relax it in regions where the SDF prediction is uncertain. Additionally, we relax normal supervision in the same way based on pseudo ground truth normals $\bar{N}(r)$ predicted by an off-the-shelf network [10], circumventing unreliable supervision. Below, we define our two uncertainty-aware geometric regularization term:

$$\mathcal{L}_N^u = \mathbb{I}_{\mu_c} \cdot \left(\left\| \frac{\nabla f(x_N)}{\|\nabla f(x_N)\|_2} - \bar{N}(r) \right\| + \left\| 1 - \left(\frac{\nabla f(x_N)}{\|\nabla f(x_N)\|_2} \right)^\top \bar{N}(r) \right\| \right), \tag{8}$$

$$\mathcal{L}_{\text{eik}}^u = \mathbb{I}_{\mu_c} \cdot \left(\left\| \nabla f(x_{(i,j)}) \right\|_2 - 1 \right)^2, \tag{9}$$

where x_N denotes the closest sample to the surface, as described in [9]. We define the indicator function $\mathbb{I}_{\mu_c}$ as follows, which can be regarded as a trimmed robust kernel that down-weights uncertain regions during supervised loss computation, in line with RobustNeRF [25]:

$$\mathbb{I}_{\mu_c} = \begin{cases} 1, & \text{if } \mu_c(\mathbf{r}) > \tau_c, \\ 0, & \text{otherwise.} \end{cases}$$

3.4. Optimization details

Probability sampling. While the GRS mechanism adjusts the sampling boundaries of each ray based on uncertainty, we further refine the sampling process using a density estimator. Specifically, we employ a proposal network – *similar to ViiNeuS [9]* – that is self-supervised by the volumetric field $\mathcal{F}_\Theta^v$ using the proposal loss $\mathcal{L}_p$ introduced in MipNeRF360 [1]. At the end of the training, we switch to computing the PDF weights directly from $\mathcal{F}_\Theta^f$ to allow the SDF representation to be freely refined.

Losses. For both representations, $\mathcal{F}_\Theta^v$ and $\mathcal{F}_\Theta^f$, we employ a standard L_1 loss to minimize the pixel-wise color difference between the rendered image $\hat{C}$ and the ground truth image $\bar{C}$, along with a DSSIM [28] loss on color patches. These losses are jointly denoted as $\mathcal{L}_{\text{rgb}}$. Similar to StreetSurf [12], we model the sky color using an auxiliary MLP conditioned on the ray direction and introduce a sky loss $\mathcal{L}_{\text{sky}}$ to enforce zero opacity for sky pixels. The segmentation mask is generated using an off-the-shelf semantic segmentation network [6]. In addition to the geometric regularization described in Section. 3.3, we apply distortion regularization $\mathcal{L}_{\text{d}}$ from MipNeRF [1] to mitigate floaters. Finally, to enhance fine structure learning, we incorporate an additional semantic head into $\mathcal{F}_\Theta^v$ and impose a cross-entropy loss for semantic supervision, denoted as $\mathcal{L}_{\text{sem}}$.

The total losses for the volumetric field ($\mathcal{L}_{\text{v}}$) and the SDF field ($\mathcal{L}_{\text{f}}$) are:

$$\mathcal{L}_{\text{v}} = \mathcal{L}_{\text{rgb}} + \lambda_d \mathcal{L}_d + \lambda_{\text{sky}} \mathcal{L}_{\text{sky}} + \lambda_{\hat{N}} \mathcal{L}_{\hat{N}} + \lambda_{\text{sem}} \mathcal{L}_{\text{sem}}, \quad (10)$$

$$\mathcal{L}_{\text{f}} = \mathcal{L}_{\text{rgb}} + \lambda_d \mathcal{L}_d + \lambda_{\text{sky}} \mathcal{L}_{\text{sky}} + \lambda_{\hat{N}} \mathcal{L}_N^u + \lambda_{\text{eik}} \mathcal{L}_{\text{eik}}^u, \quad (11)$$

where λ_d, λ_{sky}, λ_{normal}, λ_d, λ_{normal}, λ_{eik} are scaling factors. During the early epochs, we reduce the coefficients for both normal supervision and distortion regularization to facilitate a stable initialization.

4. Experiments

Implementation details We use hash encoding to encode the positions [20], and spherical harmonics to encode the viewing directions. We use 2 layers with 64 hidden units for the MLPs $\mathcal{F}_\Theta^h$ and $\mathcal{F}_\Theta^c$. We trained our model on a single consumer GPU with 24GB of VRAM, using Adam optimizer with a cosine learning rate decay from 10^{-2} to 10^{-4}. We use Marching Cubes [17] to generate the final mesh that represents the scene. Further implementation details can be found in the supplementary materials.

Datasets We evaluate our method on four public driving datasets: KITTI-360 [16], nuScenes [2], Waymo Open Dataset [26], and Pandaset [36]. For each dataset, we select four diverse scenes to capture a wide range of urban settings. In the case of Pandaset, we focus on static sequences and mask dynamic vehicles to ensure consistent evaluation.

	KITTI		Pandaset		Waymo		nuScenes	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
StreetSurf	24.04	0.83	22.24	0.66	23.42	0.77	<u>22.28</u>	0.76
ViiNeuS	**24.83**	**0.89**	22.95	0.80	<u>23.74</u>	<u>0.87</u>	21.96	<u>0.83</u>
GOF	23.34	0.86	**25.54**	**0.87**	20.92	0.81	19.13	0.76
J-NeuS (ours)	<u>24.11</u>	<u>0.88</u>	<u>23.1</u>	<u>0.80</u>	**24.90**	**0.88**	**24.10**	**0.86**

Table 2. Mean photometric results for each dataset. We highlight best performing methods in **green** and second one <u>underlined</u>.

Baselines We compare our proposed solution to current state-of-the-art (SoTA) surface reconstruction methods, including the SDF-based approaches StreetSurf [12] and ViiNeuS [9], as well as the Gaussian splatting-based method GoF [40]. StreetSurf [12] models close, far, and sky regions using hierarchical space partitioning, with 3D and 4D hash grids and occupancy grids for efficient ray sampling. ViiNeuS [9] initializes the SDF field with volumetric density predictions, achieving SoTA results in autonomous driving scenario. GoF [40] is a Gaussian splatting method that achieves SoTA performance in object-centric scenes with high image overlap. It leverages 2D Gaussian splatting regularization losses [13] and introduces an enhanced mesh extraction solution tailored for 3D Gaussian splatting (3DGS). In our experiments, we initialize GoF using COLMAP-derived sparse and dense point clouds, denoted as GoF-sparse and GoF-dense, respectively.

Evaluation metrics In addition to conventional photometric metrics (see Table 2), we assess the quality of the reconstructed meshes using two metrics similar to those in ViiNeuS [9].

- **Point to Mesh (P→M):** the mean distances from the ground truth LiDAR points to the predicted SDF-generated mesh.
- **Precision (Prec.):** the percentage of LiDAR points with a distance to the mesh below 0.15m.

4.1. Results

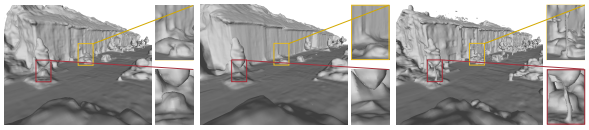
Quantitative analysis. We report quantitative results across four datasets in Tab. 1. Our method consistently outperforms or matches SoTA approaches, achieving the top metrics in most scenes on KITTI-360 and nuScenes and delivers competitive results on Pandaset and Waymo.

We observe higher P→M errors on Waymo and PandaSet due to densely occluded vegetation present in the LiDAR ground truth, not completely visible in the training images (see Figure. 2 in the supplementary). Our GRS strategy focuses sampling around the "visible" surface, forming a thin shell around trees and shrubs, which inflates the P→M error. Despite this, Our method obtains the best photomet-

	KITTI-360 [16]								Pandaset [36]							
	Seq. 30		Seq. 31		Seq. 35		Seq. 36		Seq. 23		Seq. 37		Seq. 42		Seq. 43	
	P→M	Prec.	P→M	Prec.	P→M	Prec.	P→M	Prec.	P→M	Prec.	P→M	Prec.	P→M	Prec.	P→M	Prec.
StreetSurf [12]	0.14	0.50	**0.09**	0.71	0.10	0.67	0.11	0.66	2.52	0.17	0.25	**0.66**	0.36	0.29	0.19	0.29
ViiNeuS [9]	0.13	0.56	0.11	0.71	0.11	0.66	0.13	0.72	**0.17**	0.35	**0.22**	0.44	**0.15**	0.59	**0.17**	0.45
GOF [40] - sparse	–	–	0.29	0.53	0.23	0.63	–	–	0.45	0.32	0.28	0.60	0.28	0.46	0.50	0.35
GOF [40] - dense	0.17	0.71	0.16	0.72	0.20	0.74	0.11	0.80	0.37	0.35	0.27	0.62	0.48	0.33	0.31	0.44
J-NeuS (ours)	**0.10**	**0.78**	**0.09**	**0.85**	**0.09**	**0.84**	**0.09**	**0.85**	0.21	**0.38**	0.23	0.53	0.20	**0.62**	0.20	**0.52**

	Waymo [26]								nuScenes [2]							
	Seq. 10061		Seq. 13196		Seq. 14869		Seq. 102751		Seq. 0034		Seq. 0071		Seq. 0664		Seq. 0916	
	P→M	Prec.	P→M	Prec.	P→M	Prec.	P→M	Prec.	P→M	Prec.	P→M	Prec.	P→M	Prec.	P→M	Prec.
StreetSurf [12]	0.22	0.43	0.35	**0.53**	0.23	0.35	0.25	0.24	0.57	0.29	0.78	0.47	0.67	0.50	0.65	0.28
ViiNeuS [9]	**0.19**	0.44	**0.22**	0.48	**0.14**	0.47	**0.19**	0.30	0.40	0.20	**0.22**	**0.59**	0.40	0.40	0.22	0.54
GOF [40] - sparse	1.87	0.32	2.32	0.20	1.63	0.36	1.54	0.29	1.55	0.07	1.72	0.16	1.49	0.12	1.41	0.18
GOF [40] - dense	1.20	0.38	1.17	0.39	1.55	0.41	2.11	0.34	1.02	0.12	1.55	0.23	1.44	0.12	1.06	0.29
J-NeuS (ours)	0.23	**0.47**	0.27	0.48	0.19	**0.60**	0.22	**0.46**	**0.30**	**0.43**	0.35	0.56	**0.26**	**0.56**	**0.21**	**0.64**

Table 1. Quantitative results on KITTI-360 [16], Pandaset [36], Waymo Open Dataset [26] and nuScenes [2]. We report the mean Point to Mesh (P→M) distance in meters m, and the percentage of points with a distance to mesh below $0.15m$ (Prec.). We highlight best performing methods in **green** and second one underlined. Missing entry (–) designate failure case.



(a) w/o GRS (b) w/o Ada. Relax. (c) J-NeuS (ours)

Figure 4. Ablations study: we deactivate some J-NeuS 's key components. (a) without GRS directed by $\mu_{d,c}$ (b) without adaptive relaxation on Eikonal constrain and normal supervision

ric score on Waymo and the highest average precision on each individual dataset. Averaged over the four datasets, our method achieves a precision score of 0.60, outperforming ViiNeuS (0.49) and other methods. Additionally, we attain the best mean P→M error of 0.20 m, equivalent to ViiNeuS, with StreetSurf at 0.47 m and GoF-dense at 0.82 m. These results highlight our strength in capturing precise geometry, underscoring the effectiveness of our joint optimization design.

Qualitative analysis. To complete our evaluation, we further present qualitative geometric results in Figure 5 across different dataset with complex scenes geometry, including a mixture of large road surfaces and fine structures such as poles, traffic lights, and tree trunks. Both ViiNeuS [9] and StreetSurf [12] are able to reconstruct smooth complete surface but result in incomplete fine structure reconstruction. GoF [40] produce high quality on fine structures but fails large road plane reconstruction (see sequence 0064 of nuScenes dataset). Our results demonstrate that our method outperforms the other baselines in accurately reconstructing both surfaces and fine structures, under conditions of limited image overlap.

Efficiency Table 3 reports the computational performance of each method. J-NeuSachieves the fastest mesh extraction time—under 30s—whereas GoF requires approximately 30

	mesh extraction time (infer. & post-process. min)	train time (min)	params. size (MiB)
GOF	30	$30-60$	$100-300$
StreetSurf	≈ 1	40	92.59
ViiNeuS	< 0.5	20	27
J-NeuS (ours)	< 0.5	35	51

Table 3. Performance was evaluated on the KITTI-360 dataset using the same consumer-grade GPU with 24GB of VRAM.

minutes. Moreover, our approach delivers the highest-quality meshes while maintaining moderate training time and memory usage, a critical combination for large-scale autonomous driving applications.

4.2. Ablation study

	P→M (all)	Prec. (all)	P→M (pole)
w/o GRS	0.11	0.79	0.46
w/o Ada. Relax.	0.13	0.75	0.53
Full model	**0.09**	**0.83**	**0.21**

Table 4. Ablation of our contributions (Guided Ray Sampling and Adaptive Relaxation) on the KITTI-360 dataset.

To have a clear understanding the contribution of each key component in our method, we conducted an ablation study on the KITTI-360 dataset. The results are presented in Table 4 and Figure 4. We observe that both GRS and Adaptive Relaxation improve overall geometry, with Adaptive Relaxation proving essential for complete fine structures reconstruction (reducing the P→M metric from 0.53 to 0.21). These findings align with the motivations behind our method design.

5. Conclusion

In this work, we presented **J-NeuS**, a novel uncertainty estimation framework that effectively combines volumetric SDF representations for robust urban scene reconstruction

Figure 5. Qualitative experiments results on four popular autonomous driving datasets. Complex scene geometries reconstructed by the mesh are color highlighted. We compare our mesh to the ones from StreetSurf [12], ViiNeuS [9] and GoF [40]

under limited view overlap. By estimating both photometric and geometric uncertainty, we introduce Guided Ray Sampling to deploy each representation where it excels. To avoid over regularization we propose a novel robust kernel to adaptively relax geometric regularization for SDF. Extensive quantitative and qualitative evaluations show that J-NeuS outperforms SoTA SDF and Gaussian splatting methods, delivering more accurate reconstructions on both large planar regions and fine structures under autonomous driving sensor setting.

References

[1] Jonathan T. Barron, Ben Mildenhall, Dor Verbin, Pratul P. Srinivasan, and Peter Hedman. Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In *CVPR*, 2022.

[2] Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora,

Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *CVPR*, 2020.

[3] Danpeng Chen, Hai Li, Weicai Ye, Yifan Wang, Weijian Xie, Shangjin Zhai, Nan Wang, Haomin Liu, Hujun Bao, and Guofeng Zhang. Pgsr: Planar-based gaussian splatting for efficient and high-fidelity surface reconstruction, 2024.

[4] Hanlin Chen, Chen Li, and Gim Hee Lee. Neusg: Neural implicit surface reconstruction with 3d gaussian splatting guidance, 2023.

[5] Zheng Chen, Chen Wang, Yuan-Chen Guo, and Song-Hai Zhang. Structnerf: Neural radiance fields for indoor scenes with structural hints, 2022.

[6] Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *CVPR*, 2022.

[7] Xiao Cui, Weicai Ye, Yifan Wang, Guofeng Zhang, Wengang Zhou, and Houqiang Li. Streetsurfgs: Scalable urban street surface reconstruction with planar-based gaussian splatting, 2024.

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

[9] Hala Djeghim, Nathan Piasco, Moussab Bennehar, Luis Roldão, Dzmitry Tsishkou, and Désiré Sidibé. ViiNeuS: Volumetric initialization for implicit neural surface reconstruction of urban scenes with limited image overlap. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

[10] Ainaz Eftekhar, Alexander Sax, Jitendra Malik, and Amir Zamir. Omnidata: A scalable pipeline for making multi-task mid-level vision datasets from 3d scans. In *ICCV*, 2021.

[11] Antoine Guédon and Vincent Lepetit. Sugar: Surface-aligned gaussian splatting for efficient 3d mesh reconstruction and high-quality mesh rendering. In *CVPR*, 2024.

[12] Jianfei Guo, Nianchen Deng, Xinyang Li, Yeqi Bai, Botian Shi, Chiyu Wang, Chenjing Ding, Dongliang Wang, and Yikang Li. Streetsurf: Extending multi-view implicit surface reconstruction to street views, 2023.

[13] Binbin Huang, Zehao Yu, Anpei Chen, Andreas Geiger, and Shenghua Gao. 2d gaussian splatting for geometrically accurate radiance fields. In *ACM SIGGRAPH 2024 Conference Papers*, pages 1–11, 2024.

[14] Michael Kazhdan, Matthew Bolitho, and Hugues Hoppe. Poisson surface reconstruction. In *Eurographics Symp. Geometry Processing*, page 61–70, 2006.

[15] Zhaoshuo Li, Thomas Müller, Alex Evans, Russell H Taylor, Mathias Unberath, Ming-Yu Liu, and Chen-Hsuan Lin. Neuralangelo: High-fidelity neural surface reconstruction. In *CVPR*, 2023.

[16] Yiyi Liao, Jun Xie, and Andreas Geiger. KITTI-360: A novel dataset and benchmarks for urban scene understanding in 2d and 3d. *PAMI*, 2022.

[17] 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.

[18] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020.

[19] Pierre Moulon, Pascal Monasse, Romuald Perrot, and Renaud Marlet. OpenMVG: Open multiple view geometry. In *International Workshop on Reproducible Research in Pattern Recognition*, pages 60–74. Springer, 2016.

[20] Thomas Müller, Alex Evans, Christoph Schied, and Alexander Keller. Instant neural graphics primitives with a multiresolution hash encoding. *ACM Trans. Graph.*, 41(4):102:1–102:15, 2022.

[21] Michael Niemeyer, Jonathan T Barron, Ben Mildenhall, Mehdi SM Sajjadi, Andreas Geiger, and Noha Radwan. Regnerf: Regularizing neural radiance fields for view synthesis from sparse inputs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5480–5490, 2022.

[22] Ava Pun, Gary Sun, Jingkang Wang, Yun Chen, Ze Yang, Sivabalan Manivasagam, Wei-Chiu Ma, and Raquel Urtasun. Lightsim: Neural lighting simulation for urban scenes, 2023.

[23] Konstantinos Rematas, Andrew Liu, Pratul P Srinivasan, Jonathan T Barron, Andrea Tagliasacchi, Thomas Funkhouser, and Vittorio Ferrari. Urban radiance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12932–12942, 2022.

[24] Konstantinos Rematas, Andrew Liu, Pratul P. Srinivasan, Jonathan T. Barron, Andrea Tagliasacchi, Tom Funkhouser, and Vittorio Ferrari. Urban radiance fields. In *CVPR*, 2022.

[25] Sara Sabour, Suhani Vora, Daniel Duckworth, Ivan Krasin, David J Fleet, and Andrea Tagliasacchi. Robustnerf: Ignoring distractors with robust losses. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20626–20636, 2023.

[26] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, Vijay Vasudevan, Wei Han, Jiquan Ngiam, Hang Zhao, Aleksei Timofeev, Scott Ettinger, Maxim Krivokon, Amy Gao, Aditya Joshi, Yu Zhang, Jonathon Shlens, Zhifeng Chen, and Dragomir Anguelov. Scalability in perception for autonomous driving: Waymo open dataset. In *CVPR*, 2020.

[27] Haithem Turki, Vasu Agrawal, Samuel Rota Bulò, Lorenzo Porzi, Peter Kontschieder, Deva Ramanan, Michael Zollhöfer, and Christian Richardt. Hybridnerf: Efficient neural rendering via adaptive volumetric surfaces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19647–19656, 2024.

[28] Fusang Wang, Arnaud Louys, Nathan Piasco, Moussab Bennehar, Luis Roldão, and Dzmitry Tsishkou. Planerf: Svd unsupervised 3d plane regularization for nerf large-scale urban scene reconstruction. In *3DV*, 2024.

[29] 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 *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXXII*, pages 139–155. Springer, 2022.

[30] 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. In *NeurIPS*, 2021.

[31] Yiming Wang, Qin Han, Marc Habermann, Kostas Daniilidis, Christian Theobalt, and Lingjie Liu. Neus2: Fast learning of neural implicit surfaces for multi-view reconstruction. In *ICCV*, 2023.

[32] Yifan Wang, Di Huang, Weicai Ye, Guofeng Zhang, Wanli Ouyang, and Tong He. Neurodin: A two-stage framework for high-fidelity neural surface reconstruction. *arXiv preprint arXiv:2408.10178*, 2024.

[33] Yida Wang, David Joseph Tan, Nassir Navab, and Federico Tombari. Raneus: Ray-adaptive neural surface reconstruction. In *2024 International Conference on 3D Vision (3DV)*, pages 53–63. IEEE, 2024.

[34] Zian Wang, Tianchang Shen, Jun Gao, Shengyu Huang, Jacob Munkberg, Jon Hasselgren, Zan Gojcic, Wenzheng Chen, and Sanja Fidler. Neural fields meet explicit geometric representation for inverse rendering of urban scenes. In *CVPR*, 2023.

[35] Zian Wang, Tianchang Shen, Merlin Nimier-David, Nicholas Sharp, Jun Gao, Alexander Keller, Sanja Fidler, Thomas Müller, and Zan Gojcic. Adaptive shells for efficient neural radiance field rendering. In *SIGGRAPH Asia*, 2023.

[36] Pengchuan Xiao, Zhenlei Shao, Steven Hao, Zishuo Zhang, Xiaolin Chai, Judy Jiao, Zesong Li, Jian Wu, Kai Sun, Kun Jiang, et al. Pandaset: Advanced sensor suite dataset for autonomous driving. In *ITSC*, 2021.

[37] Ze Yang, Yun Chen, Jingkang Wang, Sivabalan Manivasagam, Wei-Chiu Ma, Anqi Joyce Yang, and Raquel Urtasun. Unisim: A neural closed-loop sensor simulator. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1389–1399, 2023.

[38] Lior Yariv, Jiatao Gu, Yoni Kasten, and Yaron Lipman. Volume rendering of neural implicit surfaces. In *NeurIPS*, 2021.

[39] Zehao Yu, Songyou Peng, Michael Niemeyer, Torsten Sattler, and Andreas Geiger. Monosdf: Exploring monocular geometric cues for neural implicit surface reconstruction. *NeurIPS*, 2022.

[40] Zehao Yu, Torsten Sattler, and Andreas Geiger. Gaussian opacity fields: Efficient and compact surface reconstruction in unbounded scenes. *arXiv preprint arXiv:2404.10772*, 2024.

[41] Yongqiang Zhang, Zhipeng Hu, Haoqian Wu, Minda Zhao, Lincheng Li, Zhengxia Zou, and Changjie Fan. Towards unbiased volume rendering of neural implicit surfaces with geometry priors. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4359–4368, 2023.

[42] Hongyu Zhou, Jiahao Shao, Lu Xu, Dongfeng Bai, Weichao Qiu, Bingbing Liu, Yue Wang, Andreas Geiger, and Yiyi Liao. Hugs: Holistic urban 3d scene understanding via gaussian splatting. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21336–21345, 2024.