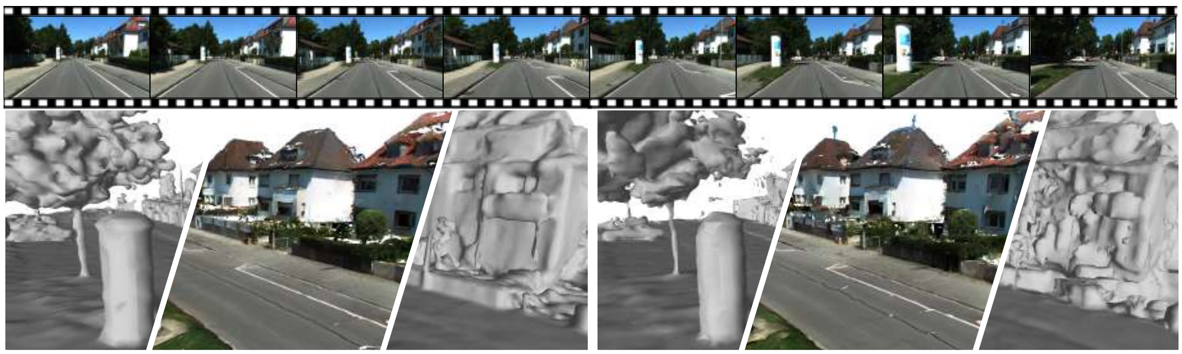
ViiNeuS: Volumetric Initialization for Implicit Neural Surface reconstruction of urban scenes with limited image overlap

Hala Djeghim[1,2] Nathan Piasco[1] Moussab Bennehar[1] Luis Roldão[1]
Dzmitry Tsishkou[1] Désiré Sidibé[2]
[1]Noah's Ark, Huawei Paris Research Center, France
[2]IBISC, Univ. Evry Paris-Saclay , Evry, France.



ViiNeuS StreetSurf [11]

Figure 1. We introduce **ViiNeuS**, a novel SDF initialization method tailored to accurately reconstruct large-scale driving scenes from RGB images with limited overlaps. Extensive experiments on popular driving datasets show the superiority of ViiNeuS's mesh (left) over previous state-of-the-art methods such as StreetSurf[11] (right). The figure presents both the initial mesh and the textured one as derived from both techniques.

Abstract

Neural implicit surface representation methods have recently shown impressive 3D reconstruction results. However, existing solutions struggle to reconstruct driving scenes due to their large size, highly complex nature and their limited visual observation overlap. Hence, to achieve accurate reconstructions, additional supervision data such as LiDAR, strong geometric priors, and long training times are required. To tackle such limitations, we present ***ViiNeuS****, a new hybrid implicit surface learning method that efficiently initializes the signed distance field to reconstruct large driving scenes from 2D street view images. ViiNeuS's hybrid architecture models two separate implicit fields: one representing the volumetric density of the scene, and another one representing the signed distance to the surface. To accurately reconstruct urban outdoor driving scenarios, we introduce a novel volume-rendering strategy that relies on self-supervised probabilistic density estimation to sample points near the surface and transition progressively from volumetric to surface representation. Our solution permits a proper and fast initialization of the signed distance field without relying on any geometric prior on the scene, compared to concurrent methods. By conducting extensive experiments on four outdoor driving datasets, we show that ViiNeuS can learn an accurate and detailed 3D surface representation of various urban scene while being two times faster to train compared to previous state-of-the-art solutions.*

1. Introduction

Achieving accurate 3D reconstruction from multi-view-consistent images has been an important topic in computer vision and computer graphics communities for decades. A 3D representation of the perceived environment is fundamental for many tasks, such as scene re-lighting [47], scene editing [48] and 3D object insertion for artificial data augmentation [40].

Recently, neural implicit surface representation methods [38, 44] have shown impressive 3D surface reconstruction performance from monocular images, achieving highly detailed and precise mesh representations. Such methods are originally inspired by Neural Radiance Fields

(NeRF) [23], and re-framed for multi-view surface reconstruction. NeRF introduced volume rendering for Novel View Synthesis (NVS) by sampling points along camera rays. These points, along with their viewing directions, are fed into a Multi-Layer Perceptron (MLP) to predict their respective colors and densities. The final image is rendered by blending these predicted values using alpha compositing. Conversely, neural surface representation methods aim to approximate the surface using a learnable Signed Distance Function (SDF) [38, 44]. Unlike NeRF, the outputs of the MLP are colors and SDF values instead of colors and volumetric densities, which are later blended using a modified alpha-compositing equation adapted to SDF. An accurately reconstructed surface is represented as the zero-level set of the SDF. Assuming objects are centered in the scene, the SDF is initially set to a spherical shape to bound the region of interest. While current methods achieve impressive reconstruction results, they are mainly designed for bounded, high-overlap observations of object-centric or landmark scenes. Applying them to urban scenes with limited image overlap leads to sub-optimal reconstructions due to their significantly different nature. Challenges include partial scene observability (e.g., occlusions from parked cars) and the need to handle both close- and far-range objects. Moreover, the geometry of driving scenes is highly complex, with fine structures and large untextured areas like roads and buildings. To overcome such limitations, implicit surface reconstruction of large outdoor driving scenes commonly relies on LiDAR supervision [28, 40] or strong geometric priors [11]. These approaches require longer training time to produce qualitative outdoor scene reconstruction meshes [19].

To address those challenges, we propose ViiNeuS, a hybrid approach to accurately initialize a signed distance field of a large area and represent complex surfaces from urban outdoor driving scenarios with high accuracy and geometric precision while requiring half of the training time compared to other state-of-the-art solutions [11]. We propose a hybrid architecture that models two separate implicit fields: one representing the volumetric density of the scene and another one representing the signed distance to the surface. ViiNeuS's novel volume rendering approach relies on a self-supervised density estimator which permits to sample points near the surface along with a progressive sampling strategy that smoothly initializes and transitions from a volumetric to a surface representation, drastically reducing the model convergence time. Inspired by previous state-of-the-art works [11, 45], we supervise surface normals by relying on normal cues predicted from a pre-trained monocular estimator [7]. We extensively validate our solution on KITTI-360 [20], Pandaset [42], Waymo Open Dataset [33], and nuScenes [3]. By comparing the quality of the extracted meshes from the signed distance field, we show that our method can efficiently reconstruct urban outdoor scenarios showing SoTA accuracy.

To summarize, the main contributions of our work are:

- a novel learning method that progressively initialize a signed distance field from street views with limited images overlap,
- a hybrid architecture, suited to the complexity of outdoor urban scene, that models two distinct implicit fields to represent both the volumetric density and the signed distance field,
- a new volume rendering technique that relies on self-supervised probabilistic density estimation to efficiently sample points near the surface, significantly reducing overall training times to obtain a scene mesh,
- an extensive comparison of SoTA implicit SDF and explicit MVS reconstruction solutions on 4 major driving datasets, showing the adaptability and superiority of ViiNeuS for urban scenes surface reconstruction.

2. Related work

Multi-view surface reconstruction. Traditional surface reconstruction techniques represent a scene either by occupancy grids [6, 27, 35, 36] or relying on point clouds [8, 29, 30, 32]. Occupancy grids-based methods are often limited by the voxel resolution and suffer from high memory consumption. On the other hand, Multi-View Stereo (MVS) techniques [24] generate, from posed images, dense point clouds associated with surface normal that can be utilized in a subsequent surface reconstruction step [17]. These methods heavily depend on the quality of the generated point cloud. Nevertheless, they can accumulate errors due to their complex multi-step pipelines, and are mostly designed for object-centric scenes with high overlaps between the observations. We refer the reader to dedicated surveys for in-depth analysis of such techniques [2, 22, 49].

Neural implicit surface representations. NeRF-based methods achieve impressive NVS results but fail to accurately reconstruct scene surfaces given the lack of constraints or priors on the surface when modeling the scene as a density volume. Neural surface reconstruction approaches, explicitly represent the surface as an SDF and propose various adaptations to train such methods through differentiable volume rendering [38, 44]. NeuS [38] proposes a novel SDF-based rendering equation that relies on SDF values to compute the alpha blending factor used in volume rendering instead of densities as in NeRF [23]. However, NeuS's geometric initialization is designed for object-centric datasets and requires long training times. Recent works aim to reduce the training time [39] or reconstruct large-scale highly detailed landmark scenes [19]. Neuralangelo [19], a SoTA method for outdoor landmarks reconstruction, requires many overlapping images assuming the structure to be reconstructed is centered within the scene. Indeed, similar to NeuS, they initialize the SDF with a spherical shape. Such approaches are not tailored for large unbounded driving scenarios and still require hours of training. In contrast, our method is built upon NeuS's formulation but is designed to reliably represent urban outdoor driving scenes.

Driving scenes surface reconstruction. Few surface reconstruction methods consider acquisition from a moving camera rig for driving sequences surface reconstruction. Existing solution are mainly guided by LiDAR such as FEGR [40] and Urban Radianc Fields [28], limiting their large-scale deployment. StreetSurf [11], the most related method to our proposal, divides the scene into separate close-range and far-range views and uses ego pose-based road surface initialization. This method assumes the road and sidewalk are at the same height across the entire scene, which limits its applicability to long, narrow driving scenarios, as seen in the Waymo Open Dataset [33]. Instead, ViiNeuS does not rely on such topological assumptions, making it more adaptable to a variety of driving scenes, including those that are short and curvy. Additionally, ViiNeuS is more compact and faster to train, as it represents the entire scene with a single hybrid implicit field.

Hybrid scene representation. Some state-of-the-art methods have explored hybrid rendering solutions combining volumetric and surface representations. Wang et al. [41] introduced a hybrid method based on NeuS [38] using a spatially varying kernel that adapts based on surface fuzziness. This kernel allows the extraction of a narrow mesh envelope around the surface, which enables ray casting only inside this envelope, accelerating both rendering time and quality. Other methods [12, 15, 34] combine explicit mesh representations with SDF representations. The hybrid volume-mesh representation reduces storage and resource demands while outputting implicit geometry. Additionally, methods like NeuSG [4] and other 3D Gaussian Splatting-based [18] solutions [10, 14] aim to optimize Gaussians for surface reconstruction, with some approaches directly identifying level sets for surface extraction, such as in Gaussian Opacity Fields [46]. The aforementioned solutions mainly focus on improving the rendering speed and are better fitted for scenes with high image overlap. In our proposal, we jointly train a hybrid model to benefit from the fast training of a density-based approach along with an accurate geometric representation obtained thanks to the signed distance field learning.

3. Method

ViiNeuS models the volumetric density and signed distance field in a hybrid architecture. The details of this architecture are presented in Sec. 3.2, while the novel underlying volume rendering strategy is explained in Sec. 3.3. Our hybrid strategy relies on a self-supervised density distribution estimation to sample points near the surface and progressively transition from a volumetric to an implicit surface representation from limited overlapping street views. To facilitate the convergence of our method, we further introduce regularization techniques for the signed distance field. The overall architecture of our proposed method is illustrated in Fig. 2.

3.1. Preliminaries

NeRF's Implicit Volumetric Representation. NeRF [23] takes as input the 3D positions $x_i \in \mathbb{R}^3$ and the viewing direction $d \in \mathbb{R}^2$ of sampled points along each camera ray r and predicts their corresponding densities $\sigma_i \in \mathbb{R}$ and colors $c_i \in \mathbb{R}^3$. This representation of the scene is approximated using a Multi-Layer Perceptron (MLP). Finally, volume rendering is used to alpha-composite the colors of the samples along each ray and render the final pixel color $\hat{C}(r) \in \mathbb{R}^3$:

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

where $T_i = \prod_{j=1}^{i-1}(1-\alpha_j)$ is the accumulated transmittance along the ray and $\alpha_i \in \mathbb{R}$ a blending factor. This factor (referred to in the literature as the alpha-value) is used for alpha-composing, and is given by:

$$\alpha_i = 1 - \exp(-\sigma_i \delta_i), \tag{2}$$

with $\delta_i \in \mathbb{R}$ being the distance between the samples along the ray. Note that in ViiNeuS's hybrid formulation, we refer to this value as α_i^v to disambiguate from alpha values derived from the estimated SDF which is introduced in Eq. 3.

SDF representation with NeuS. For an accurate surface-based representation of a 3D scene, NeuS [38] presented a different formulation of the NeRF's volume rendering equation. They represent the scene as a signed distance function, where the surface is the zero-level set of such function: $S = \{x \in \mathbb{R}^3 | f(x) = 0\}$, where $f(x)$ is the signed distance value at a given position x. NeuS adopts a volume rendering formulation that introduces a density distribution function $\phi_s(x)$, which is formulated as the derivative of the sigmoid function $\Phi_s(x) = (1 + e^{-sx})^{-1}$, that is, $\phi_s(x) = se^{-sx}/(1 + e^{-sx})^2$. The standard deviation of ϕ_s is $1/s$ and, as the network converges to an accurate surface representation, this parameter moves toward zero. Following NeuS's formulation, the volumetric alpha used in Eq. 1 is replaced by the following equation to approximate the color $\hat{C}(r)$:

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

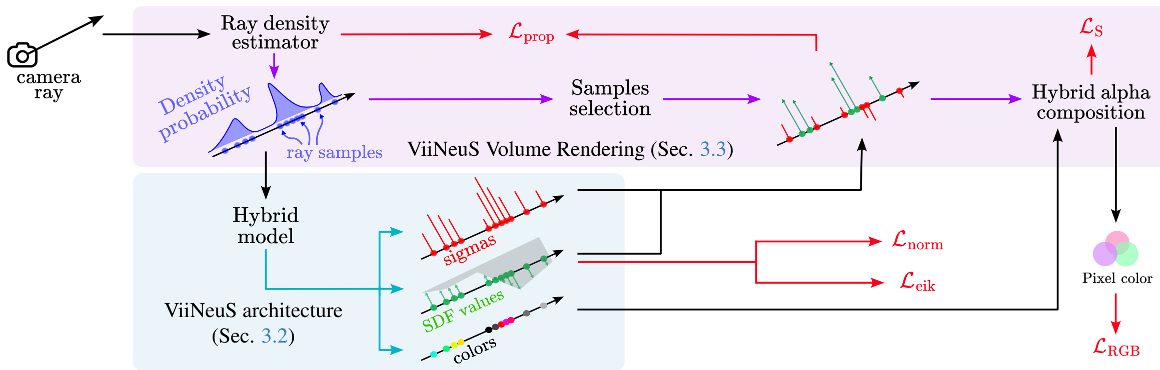


Figure 2. **ViiNeuS overview –** Our solution can be divided into two key components: an hybrid scene representation (Sec. 3.2) and a ray-based volumetric rendering that progressively transitions from density to SDF sample alpha composition (Sec. 3.3).

where $f(p_i)$ and $f(p_{i+1})$ are signed distance values at section points centered on x_i. In our formulation, we denote alpha values derived from the estimated SDF as α_i^f.

SDF field initialization In [43], the authors show that geometric initialization of the signed distance field (SDF) to a bounded sphere is key for learning an SDF function. However, such initialization assumes full scene observability from the training views and a centered region of interest, which does not apply to driving scenes where the camera mounted on the ego-vehicle remains on the road and follows a linear trajectory. StreetSurf [11] initializes the SDF based on road-surface. Instead, we propose a progressive initialization of the SDF, tailored for the complexities of urban scenes, ensuring effective SDF learning across both close and far-range elements in the entire driving sequence.

3.2. ViiNeuS's hybrid architecture

Our hybrid method ViiNeuS learns the two different fields simultaneously, as illustrated in Fig 3. ViiNeuS expresses the scene using two functions, $\mathcal{F}^h_\Theta$ and $\mathcal{F}^c_\Theta$, encoded, respectively, by two MLPs. The input 3D position x and the viewing direction d are embedded, respectively, through a positional and directional embedding into $\text{emb}_x(x)$ and $\text{emb}_d(d)$.
$\mathcal{F}^h_\Theta$ takes as input $\text{emb}_x(x)$ and outputs a density σ, a signed distance value $f(x)$, and a latent vector h, that is $(\sigma, f(x), h) = \mathcal{F}^h_\Theta(\text{emb}_x(x))$.
$\mathcal{F}^c_\Theta$ on the other hand, takes as input the latent vector h that is outputted by $\mathcal{F}^h_\Theta$ in addition to the embedded viewing direction $\text{emb}_d(d)$ and the normal vector $\overrightarrow{n}$ obtained from the SDF gradient, to predict the color $c = \mathcal{F}^c_\Theta(\text{emb}_d(d), h, \overrightarrow{n})$.

3.3. ViiNeuS's volume rendering

Probabilistic density estimation. Probabilistic density estimation along the ray is employed for a more targeted sampling strategy that samples fewer points and accelerates the overall method's training. Guided by the estimated density, N points are sampled along the ray, where higher density indicates the presence of the surface. Using ViiNeuS's architecture, alpha values are computed for these samples, and volume rendering is performed to obtain the final color associated with the ray:

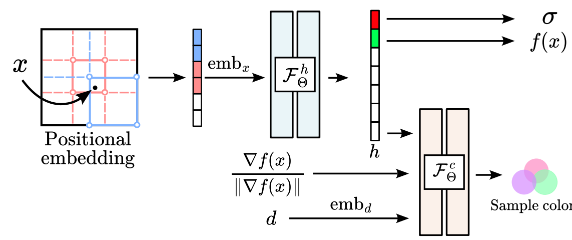


Figure 3. **ViiNeuS architecture –** Our method uses two MLP functions $\mathcal{F}^h_\Theta$ and $\mathcal{F}^c_\Theta$ to output SDF values and density (σ and $f(x)$, respectively) along with color values given an input sample. We inspire from NeRF and design $\mathcal{F}^c_\Theta$ to output the color given the lattent vector h outputted by $\mathcal{F}^h_\Theta$, the viewing direction (d) and the normal vector $\overrightarrow{n}$ obtained from the gradient of the SDF.

$$\hat{C}(r) = \sum_{i=1}^{N} T_i \alpha_i c_i,$$
$$\text{with } \alpha_i = \begin{cases} \alpha_i^v & \text{if computed from } \sigma_i \text{ (Eq. 2)}, \\ \alpha_i^f & \text{if computed from } f(x_i) \text{ (Eq. 3)}. \end{cases} \tag{4}$$

This resulting color is then compared to the ground truth value to train our model. The strategy of associating a sample to either a volumetric or surface-based representation is discussed in the next section.

The weights w_i associated with each sample (Eq. 1) serve as a supervision signal to our probabilistic density estimator. Hence, we can refine in a self-supervised manner the prediction of the expected density along the ray. We rely on Mip-NeRF 360's [1] proposal networks as our volumetric density estimator and use their introduced proposal loss,

$\mathcal{L}_{\text{prop}}$, to refine density estimation from the weights w_i of our main model.

SDF progressive initialization: Volumetric to surface representation. Starting from a pure volumetric density training ($\alpha_i = \alpha_i^v$), we introduce a progressive initialization of the signed distance field that transitions from a volumetric to an SDF representation. Our approach is based on the assumption that estimating the density σ_i at a specific location requires less complex geometric understanding, in contrast to estimating the signed distance to the closest point to a complex urban outdoor surface (i.e. $f(x_i)$). The 3D geometry represented as a density field contains less information than the SDF field; the former contains occupancy information (which can be considered as binary) whereas the latter represents, additionally, continuous distances to the surface, including close and far-range points as well as the empty spaces between driving scene objects, such as cars and trees. Therefore, in this specific scenario where observations have reduced overlap introducing ambiguity in the geometry estimation, learning a density field is simpler compared to the SDF field (see figure 5 of our ablation section). Furthermore, by initially leveraging a pure volumetric density representation during training, we shorten the convergence time of our density estimator.

Our progressive initialization strategy can be outlined as follows:

- Volumetric stage: All samples along the ray express volumetric densities, and their alpha values are computed with σ_i.
- Hybrid stage: we gradually assign samples along the ray to train our signed distance field. Within the same ray, a portion of the sample values is represented by volumetric alpha α_i^v, while the remainder is computed with α_i^f.
- Surface stage: we increase the number of samples for which alpha values are computed using $f(x)$ until we transition to a complete surface representation.

Hybrid stage regularization. Composing alpha values obtained from our two different geometric representations is not straightforward as the learning dynamic of the density and the SDF are not the same. As mentioned earlier and due to the partial observability of the reconstructed environment, learning a volumetric representation of the scene based only on spatial density is easier and faster compared to learning an arbitrary signed distance field of a complex urban scene. To facilitate the convergence of our representation, we introduce two regularization methods: **probability-based sampling attribution** and **SDF-gradient normalization**.

Probability-based samples attribution. To assign the alpha value based on density or SDF to a sample, we initially start with a random selection for each ray. We noticed that this association strategy is not suitable: the signed distance field could not be properly initialized during the hybrid stage. To solve this problem, we introduce a probability-based sample association. We assign α_i^f to samples with the highest predicted density to initialize the signed distance field. It facilitates the learning of the signed distance field by limiting the range of predicted values, as all the points are sampled close to a surface. By sampling points near the surface, we handle the wide range of distances of driving scenarios and complex geometries.

SDF-gradient normalization. During the hybrid stage, where α_i^v and α_i^f are jointly composed, we observed that the surface representation often predicts large SDF gradient values to align α_i^f with α_i^v (for more details, see our supplementary materials). To address this, we normalize the SDF gradient in the α_i^f computation to prevent the gradient from compensating the alpha distribution difference.

3.4. Optimization

Photometric loss. We use the standard L_1 loss to minimize the pixel-wise color difference between the rendered image $\hat{C}$, and the ground truth image C, as well as DSSIM [37] loss on color patches.

Monocular cues. As we target to reconstruct outdoor datasets, we add an additional sky loss to our model. Similar to StreetSurf [11], we model the sky color with an auxiliary MLP conditioned on the ray direction and use a loss to constrain the opacity of sky pixels to zero. The segmentation mask is obtained with an off-the-shelf semantic segmentation network [5].

Inspired by MonoSDF[45] and StreetSurf [11], we use off-the-shelf monocular depth estimation models to supervise the rendered normals $\hat{N}(r)$. In standard solutions [11, 45], the monocular normal loss is minimized according to the predicted normal accumulated along the ray:

$$\hat{N}(r) = \sum_{i=1}^{N} T_i \alpha_i \frac{\nabla f(x_i)}{\|\nabla f(x_i)\|_2}, \tag{5}$$

$$\mathcal{L}_{\hat{N}} = \left\| \hat{N}(r) - \bar{N}(r) \right\| + \left\| 1 - \hat{N}(r)^\top \bar{N}(r) \right\|. \tag{6}$$

In our proposal, however, for more accurate and targeted supervision, we supervise only the normal associated with the closet sample x_N to the surface estimated by transmittance saturation:

$$\mathcal{L}_{\hat{N}} = \left\| \frac{\nabla f(x_N)}{\|\nabla f(x_N)\|_2} - \bar{N}(r) \right\| + \left\| 1 - \left(\frac{\nabla f(x_N)}{\|\nabla f(x_N)\|_2} \right)^\top \bar{N}(r) \right\|. \tag{7}$$

	KITTI-360 [20]							
	Seq. 30		Seq. 31		Seq. 35		Seq. 36	
	P→M	Prec.	P→M	Prec.	P→M	Prec.	P→M	Prec.
StreetSurf [11]	0.14	0.50	**0.09**	**0.71**	**0.10**	**0.67**	**0.11**	0.66
ViiNeuS (ours)	**0.13**	**0.56**	0.11	**0.71**	0.11	0.66	0.13	**0.72**
COLMAP [31]	**0.11**	0.70	**0.09**	**0.81**	**0.08**	0.78	**0.08**	0.74
OpenMVS [25]	**0.11**	**0.77**	**0.08**	**0.80**	**0.08**	**0.82**	0.10	**0.80**
GOF [46] - sparse	–	–	0.29	0.53	0.23	0.63	–	–
GOF [46] - dense	0.17	0.71	0.16	0.72	0.20	0.74	0.11	**0.80**

	Pandaset [42]							
	Seq. 23		Seq. 37		Seq. 42		Seq. 43	
	P→M	Prec.	P→M	Prec.	P→M	Prec.	P→M	Prec.
StreetSurf [11]	2.33	0.10	0.47	0.40	0.36	0.30	**0.27**	0.31
ViiNeuS (ours)	**0.49**	**0.23**	**0.28**	**0.46**	**0.33**	**0.29**	0.35	**0.34**
COLMAP [31]	0.53	0.41	**0.85**	0.52	0.65	0.53	**0.26**	0.63
OpenMVS [25]	0.62	**0.56**	1.0	0.53	0.73	0.58	0.31	**0.64**
GOF [46] - sparse	**0.45**	0.51	0.90	0.54	0.42	0.59	0.51	0.45
GOF [46] - dense	0.51	0.48	0.86	**0.62**	**0.41**	**0.67**	0.30	0.55

	Waymo [33]							
	Seq. 10061		Seq. 13196		Seq. 14869		Seq. 102751	
	P→M	Prec.	P→M	Prec.	P→M	Prec.	P→M	Prec.
StreetSurf [11]	0.22	0.43	0.35	**0.53**	0.23	0.35	0.25	0.24
ViiNeuS (ours)	**0.19**	**0.44**	**0.22**	0.48	**0.14**	**0.47**	**0.19**	**0.30**
COLMAP [31]	**0.18**	0.57	**0.71**	**0.53**	**0.14**	**0.67**	**0.19**	0.59
OpenMVS [25]	0.25	**0.58**	0.76	**0.53**	0.20	**0.67**	0.22	**0.64**
GOF [46] - sparse	1.87	0.32	2.32	0.20	1.63	0.36	1.54	0.29
GOF [46] - dense	1.20	0.38	1.17	0.39	1.55	0.41	2.11	0.34

	nuScenes [3]							
	Seq. 0034		Seq. 0071		Seq. 0664		Seq. 0916	
	P→M	Prec.	P→M	Prec.	P→M	Prec.	P→M	Prec.
StreetSurf [11]	0.57	**0.29**	0.78	0.47	0.67	**0.50**	0.65	0.28
ViiNeuS (ours)	**0.40**	0.20	**0.22**	**0.59**	**0.40**	0.40	**0.22**	**0.54**
COLMAP [31]	**0.25**	0.51	**0.77**	0.67	**0.65**	0.62	**0.70**	0.68
OpenMVS [25]	0.28	**0.54**	0.85	**0.69**	0.75	**0.70**	0.74	**0.69**
GOF [46] - sparse	1.55	0.07	1.72	0.16	1.49	0.12	1.41	0.18
GOF [46] - dense	1.02	0.12	1.55	0.23	1.44	0.12	1.06	0.29

Table 1. Quantitative results on KITTI-360 [20], Pandaset [42], Waymo Open Dataset [33] and nuScenes [3]. 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 implicit SDF methods in green and explicit MVS methods in blue. Missing entry (–) designate failure case. GOF - sparse refers to GOF [46] gaussians initialized with sparse point cloud. GOF - dense refers to GOF [46] gaussians initialized with dense point cloud.

Signed distance field regularization. Similar to [38], to constrain our model to respect the eikonal equation [9] we use the eikonal regularization defined as:

$$\mathcal{L}_{\text{eik}} = \frac{1}{N}\sum_{i=1}^{N}\left(\|\nabla f(x_i)\|_2 - 1\right)^2. \tag{8}$$

According to NeuS formulation of the alpha compositing equation (Eq. 3), the network converges properly when the standard deviation $1/s$ of ϕ_s approaches zero. Due to the complexity of driving scenarios and the large scale of urban scenes, the parameter s doesn't always reach a satisfactory level of convergence. Unlike StreetSurf which manually fixes the parameter s to increment linearly, we introduce a regularization term on s during training: $\mathcal{L}_{\text{s}} = 1/(s+\epsilon)$.

4. Experiments

We evaluate the performance of ViiNeuS on multiple driving scenes and provide qualitative and quantitative comparisons compared to relevant methods.

4.1. Implementation details.

We use hash encoding to encode the positions [26] (with the same default parameters as in InstantNGP [26]), and spherical harmonics to encode the viewing directions. We use 2 layers with 64 hidden units for the MLPs $\mathcal{F}^h_\Theta$ and $\mathcal{F}^c_\Theta$. Pure volumetric stage last for the 100 first steps and the hybrid stage ends at 35% of the total training time. We perform our experiments on a single high-tier GPU using Adam optimizer with a cosine learning rate decay from 10^{-2} to 10^{-4}. For the parameter s, we apply a slightly slower weight decay (10^{-3} to 10^{-5}) to stabilize convergence. We train our model for a total of 14k iterations. Following the literature, we use Marching Cubes [21] to generate the final mesh that represents the scene. Further details on implementation can be found in the supplementary materials.

4.2. Datasets.

We evaluate our method on four public autonomous driving datasets: KITTI-360 [20], nuScenes [3], Waymo Open Dataset [33] and Pandaset [42]. We conduct our experiments on four static scenes from each one of these datasets. Since our method is designed for reconstructing the static background of urban areas with limited image overlap, we mask dynamic vehicles and pedestrians in the Pandaset sequences during training. Further details on the selected sequences are provided in the supplementary materials. We use poses provided by the datasets except for Waymo where we recompute the vehicle trajectory and sensor calibration with MOISST [13] as the ones provided are not accurate.

4.3. Baseline.

We evaluate the state-of-the-art (SoTA) method StreetSurf [11] for driving scenes surface reconstruction, using its official implementation with monocular depth supervision and trajectory optimization disabled for a fair comparaison. All experiments were conducted on the same computational setup. GOF [46], a Gaussian splitting method for landmark surface reconstruction, was also evaluated using the official implementation. We modified it to initialize the Gaussians with a dense point cloud from COLMAP and adjusted the rasterizer for cameras with an off-center focal point. We evaluate both sparse and dense point cloud initialization. We extensively tested Neuralangelo [19] on the selected driving sequences, with various training strategies and hyperparameters tuning. However we did not obtain satisfactory results: Neuralangelo requires a bounded scene to initialize the SDF in a sphere shape, making it unsuitable for the open, complex nature of driving sequences. For more details on these experiments, we refer readers to our supplementary materials. To provide a more comprehensive comparison we benchmark ViiNeuS mesh reconstruction with meshes obtained by two methods based on classic MVS: Delaunay triangulation implemented in COLMAP [31] and mesh reconstruction [16] from OpenMVS [25]. For both of the MVS meshing baselines, we used the dense point cloud generated by COLMAP as it provides denser results compared to the one obtained with OpenMVS library.

4.4. Evaluation metrics.

To evaluate the quality of the reconstructed surfaces, we report two metrics:

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.

4.5. Results

Quantitative analysis. We report quantitative results on the 4 datasets in the Tab. 1. Results show that ViiNeuS achieves on par results with StreetSurf on KITTI-360 dataset, with a mean P→M error of 12 centimeters. However, for the remaining datasets, ViiNeuS consistently outperforms StreetSurf by achieving lower P→M error and higher precision among most of the scenes. The similar performance on KITTI-360 is due to its topology with mainly straight roads and open scenes which are more suitable to the design choices done by StreetSurf. Conversely, the other datasets contain more challenging scenarios with wide, non-straight roads, and a high level of detail (we refer to the qualitative analysis). ViiNeuS demonstrates comparable P→M error with explicit MVS methods on KITTI-360, which can be attributed to the fact that KITTI-360 is an ideal dataset for multi-view reconstruction thanks to the good quality of the images and the large FOV of the cameras mounted on the vehicle. However, in more complex environments with smaller overlaps between observations, ViiNeuS consistently outperforms explicit MVS methods by achieving lower P→M error. Since they use a dense point clouds as initialization, explicit MVS methods can achieve higher precision by reconstructing finer structures (see Qualitative analysis).

The results averaged for the 4 datasets give a mean P→M of **0.24m** for ViiNeuS , two times smaller compared to 0.47m for StreetSurf and a mean precision of **0.46%** for ViiNeuS and 0.42% for StreetSurf (see supplementary materials for additional quantitative evaluation).

	Total train time (min)	GPU memory (Gb)	Params size (MiB)
GOF training† [46]	30-60	8-19	100-300
GOF mesh extraction [46]	30	8	-
StreetSurf [11]	40	19	92.59
ViiNeuS (ours)	**20**	**8**	**27**
COLMAP [31]	30	<1	–
OpenMVS [25]	30	<1	–

Table 2. **Training analysis:** reported using the same GPU, equivalent to an RTX4090. † Training time excludes COLMAP point cloud initialization.

Efficiency. We report in Tab. 2 the computational performance of each method. ViiNeuS is more efficient than StreetSurf [11], using a single 3D hierarchical hash grid instead of two grids (3D and 4D), with a third of the parameters and half the GPU memory. GOF, on the other hand, has variable GPU memory and parameters depending on the number of Gaussians per scene. For a single scene, ViiNeuS requires half the training time compared to StreetSurf (20 min vs 40 min), while GOF takes at least 1.5 hours, including 30 minutes each for point cloud initialization, training, and mesh extraction.

Qualitative analysis. The quantitative evaluation presented in the previous section (Sec. 4.5) is computed from the accumulated ground-truth LiDAR which are sparse and do not cover some regions of the scene. To complete our evaluation, we present qualitative results in Figs. 4. Results demonstrate that ViiNeuS reconstructs higher-quality surfaces compared to StreetSurf and explicit MVS methods. Even if finer details can be recovered on some parts of the scene, meshes obtained by explicit MVS methods have noisy normals and are incomplete, which limits their applicability to some tasks such as mesh texturing (see supplementary materials). We observe that our method recovers more accurate and realistic scene details at many regions such as buildings, cars, poles, and trees. Our analysis supports the hypothesis that SteertSurf's disentanglement of the close and far range based on the assumption that driving scenes are long and narrow is not adapted to the majority of driving sequences. In addition, StreetSurf's road-surface initialization fails in sequences where the road and sidewalks are not at the same height or at non-flat roads such as downhills (see Fig. 4 – Seq. 30).

4.6. Ablation study

Convergence of density field compared to SDF field. We report in Fig. 5 the Chamfer Distance between LiDAR and predicted point clouds, for Seq. 31 from KITTI-360 and Seq. 916 from nuScenes, comparing two different NeRF: one solely based on density volumetric field and another one solely based on SDF at various training steps (without normal supervision, as we cannot supervise normal in case

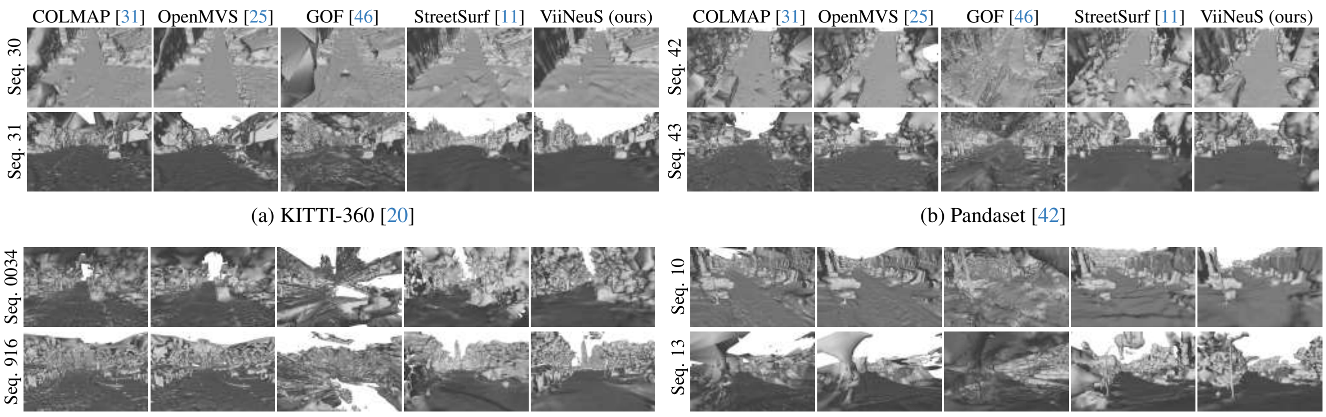


Figure 4. Qualitative experiments results on (a) KITTI-360, (b) Pandaset, (c) nuScenes and (d) Waymo Open Dataset. We compare our mesh extracted from our SDF to GOF, COLMAP, OpenMVS and StreetSurf meshes.

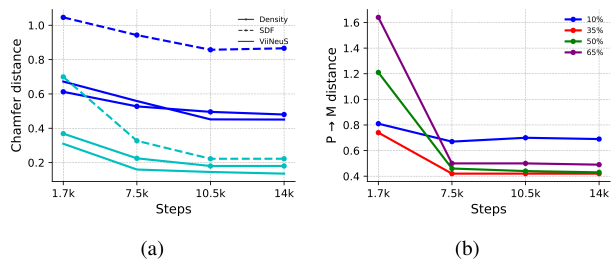


Figure 5. Chamfer distance between LiDAR and predicted point cloud of Density field and SDF field at different training steps (left) for Seq. 31 KITTI-360 and Seq. 0916 nuscenes. P→M for different hybrid stage duration at different training steps for Seq. 23 from Pandaset (right).

of pure density volumetric field). The results support the claim that density field converges faster compared to the SDF field.

Duration of hybrid stage. To analyze the effectiveness of our progressive initialization approach, we report the P→M distance at various iterations with different transition times in Fig. 5. The results demonstrate that a short hybrid stage cannot be sufficient to initialize the SDF field, while a long hybrid stage contributes to losing the initialization of the SDF field and results in sub-optimal reconstructions.

ViiNeuS's components. We ablate the effect of ViiNeuS's SDF progressive initialization strategy and show quantitative results in Tab. 3. The results demonstrate that our initialization that transitions from the volumetric to the surface representation is crucial to accurately initialize the SDF field and efficiently reconstruct driving sequences. In addition, we report our method with standard accumulated normal supervision, and without the network regularization loss.

Table 3. Ablations study: we deactivate some of the key components of ViiNeuS in order to measure the impact on the final reconstruction. (a) without progressive initialization (b) with the accumulated normal for $\mathcal{L}_{\hat{N}}$ (c) without $\mathcal{L}_s$.

	w/o SDF init		w/o $\mathcal{L}_{\hat{N}}$		w/o $\mathcal{L}_s$		ViiNeuS	
Seq.	P→M	Prec.	P→M	Prec.	P→M	Prec.	P→M	Prec.
KITTI-360. 31	0.35	0.18	0.19	0.27	0.24	0.15	**0.15**	**0.71**
Pandaset. 23	0.85	0.18	0.80	0.18	0.54	0.17	**0.51**	**0.23**
nuScenes. 0916	0.35	0.4	0.4	0.21	0.35	0.15	**0.25**	**0.54**
Waymo. 10061	0.3	0.36	0.22	0.32	0.34	0.17	**0.23**	**0.43**

5. Discussion

Applications. Thanks to the quality and completeness of ViiNeuS's reconstructed meshes, we can use modern MVS tools such as OpenMVS [25] to obtain a detailed and colorized representation of large urban scenes, as shown in Fig. 1.

Limitations. Although our method can achieve highly detailed surface reconstruction in driving scenarios, it fails to reconstruct fine details in very wide and open sequences. We recommend consulting the supplementary materials for the failure cases of ViiNeuS.

6. Conclusion

In this work we presented **ViiNeuS**, an implicit hybrid approach to efficiently reconstruct large urban scenes. Our method's hybrid architecture models the volumetric density and signed distance in two separate fields. We introduced a novel hybrid volume rendering strategy to progressively transition from volumetric representation to a signed distance representation. The quantitative and qualitative evaluations that we performed show that ViiNeuS achieves more precise reconstruction than state-of-the-art methods while being significantly faster and is better suited for a large variation of driving scenarios.

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. 4

[2] Matthew Berger, Andrea Tagliasacchi, Lee M. Seversky, Pierre Alliez, Gaël Guennebaud, Joshua A. Levine, Andrei Sharf, and Cláudio T. Silva. A survey of surface reconstruction from point clouds. *Computer Graphics Forum*, 36, 2017. 2

[3] 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. 2, 6, 8

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

[5] Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *CVPR*, 2022. 5

[6] Jeremy S De Bonet and Paul Viola. Poxels: Probabilistic voxelized volume reconstruction. In *ICCV*, 1999. 2

[7] 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. 2

[8] Silvano Galliani, Katrin Lasinger, and Konrad Schindler. Gipuma: Massively parallel multi-view stereo reconstruction. In *ICCV*, 2016. 2

[9] Amos Gropp, Lior Yariv, Niv Haim, Matan Atzmon, and Yaron Lipman. Implicit geometric regularization for learning shapes. In *MLSYS*. 2020. 6

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

[11] 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. 1, 2, 3, 4, 5, 6, 7, 8

[12] Yuan-Chen Guo, Yan-Pei Cao, Chen Wang, Yu He, Ying Shan, Xiaohu Qie, and Song-Hai Zhang. Vmesh: Hybrid volume-mesh representation for efficient view synthesis. In *SIGGRAPH Asia*. ACM, 2023. 3

[13] Quentin Herau, Nathan Piasco, Moussab Bennehar, Luis Roldão, Dzmitry Tsishkou, Cyrille Migniot, Pascal Vasseur, and Cédric Demonceaux. Moisst: Multimodal optimization of implicit scene for spatiotemporal calibration. In *IROS*, 2023. 6

[14] Binbin Huang, Zehao Yu, Anpei Chen, Andreas Geiger, and Shenghua Gao. 2d gaussian splatting for geometrically accurate radiance fields. In *SIGGRAPH 2024 Conference Papers*. Association for Computing Machinery, 2024. 3

[15] Zhangjin Huang, Zhihao Liang, Haojie Zhang, Yangkai Lin, and Kui Jia. Sur2f: A hybrid representation for high-quality and efficient surface reconstruction from multi-view images, 2024. 3

[16] Michal Jancosek and Tomás Pajdla. Exploiting visibility information in surface reconstruction to preserve weakly supported surfaces. *International Scholarly Research Notices*, 2014, 2014. 7

[17] Michael Kazhdan, Matthew Bolitho, and Hugues Hoppe. Poisson surface reconstruction. In *Eurographics Symp. Geometry Processing*, page 61–70, 2006. 2

[18] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42 (4), 2023. 3

[19] 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. 2, 6

[20] Yiyi Liao, Jun Xie, and Andreas Geiger. KITTI-360: A novel dataset and benchmarks for urban scene understanding in 2d and 3d. *PAMI*, 2022. 2, 6, 8

[21] William E. Lorensen and Harvey E. Cline. Marching cubes: A high resolution 3d surface construction algorithm. *ACM Transactions on Graphics*, 21(4):163–169, 1987. 6

[22] Charles Malleson, Jean-Yves Guillemaut, and Adrian Hilton. 3d reconstruction from rgb-d data. *RGB-D Image Analysis and Processing*. 2

[23] 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. 2, 3

[24] 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. 2

[25] Pierre Moulon, Pascal Monasse, Romuald Perrot, and Renaud Marlet. OpenMVG: Open multiple view geometry. In *International Workshop on Reproducible Research in Pattern Recognition*, 2016. 6, 7, 8

[26] 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. 6

[27] Despoina Paschalidou, Ali Osman Ulusoy, Carolin Schmitt, Luc van Gool, and Andreas Geiger. Raynet: Learning volumetric 3d reconstruction with ray potentials. In *CVPR*, 2019. 2

[28] Konstantinos Rematas, Andrew Liu, Pratul P. Srinivasan, Jonathan T. Barron, Andrea Tagliasacchi, Thomas Funkhouser, and Vittorio Ferrari. Urban radiance fields. In *CVPR*, 2022. 2, 3

[29] Luis Roldão, Raoul de Charette, and Anne Verroust-Blondet. 3d surface reconstruction from voxel-based lidar data. *(ITSC)*, pages 2681–2686, 2019. 2

[30] Johannes L Schönberger, Enliang Zheng, Jan-Michael Frahm, and Marc Pollefeys. Pixelwise view selection for unstructured multi-view stereo. In *ECCV*, 2016. 2

[31] Johannes Lutz Schönberger, Enliang Zheng, Marc Pollefeys, and Jan-Michael Frahm. Pixelwise view selection for unstructured multi-view stereo. In *ECCV*, 2016. 6, 7, 8

[32] Robust Multiview Stereopsis. Accurate, dense, and robust multiview stereopsis. *TPAMI*, 32(8):1362–1376, 2010. 2

[33] 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. 2, 3, 6, 8

[34] Jiaxiang Tang, Hang Zhou, Xiaokang Chen, Tianshu Hu, Errui Ding, Jingdong Wang, and Gang Zeng. Delicate textured mesh recovery from nerf via adaptive surface refinement. In *ICCV*, 2023. 3

[35] Shubham Tulsiani, Tinghui Zhou, Alexei A. Efros, and Jitendra Malik. Multi-view supervision for single-view reconstruction via differentiable ray consistency. In *CVPR*, 2017. 2

[36] Ali Osman Ulusoy, Andreas Geiger, and Michael J. Black. Towards probabilistic volumetric reconstruction using ray potentials. In *3DV*, 2015. 2

[37] Fusang Wang, Arnaud Louys, Nathan Piasco, Moussab Bennehar, Luis Roldãao, and Dzmitry Tsishkou. Planerf: Svd unsupervised 3d plane regularization for nerf large-scale urban scene reconstruction. In *3DV*, 2024. 5

[38] 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. 1, 2, 3, 6

[39] 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. 2

[40] 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. 1, 2, 3

[41] 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. 3

[42] 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. 2, 6, 8

[43] Lior Yariv, Yoni Kasten, Dror Moran, Meirav Galun, Matan Atzmon, Basri Ronen, and Yaron Lipman. Multiview neural surface reconstruction by disentangling geometry and appearance. *Advances in Neural Information Processing Systems*, 33:2492–2502, 2020. 4

[44] Lior Yariv, Jiatao Gu, Yoni Kasten, and Yaron Lipman. Volume rendering of neural implicit surfaces. In *NeurIPS*, 2021. 1, 2

[45] Zehao Yu, Songyou Peng, Michael Niemeyer, Torsten Sattler, and Andreas Geiger. Monosdf: Exploring monocular geometric cues for neural implicit surface reconstruction. *NeurIPS*, 2022. 2, 5

[46] Zehao Yu, Torsten Sattler, and Andreas Geiger. Gaussian opacity fields: Efficient adaptive surface reconstruction in unbounded scenes. *ACM Transactions on Graphics*, 2024. 3, 6, 7, 8

[47] Jingyang Zhang, Yao Yao, Shiwei Li, Jingbo Liu, Tian Fang, David McKinnon, Yanghai Tsin, and Long Quan. Neilf++: Inter-reflectable light fields for geometry and material estimation. In *ICCV*, 2023. 1

[48] Jingsen Zhu, Yuchi Huo, Qi Ye, Fujun Luan, Jifan Li, Dianbing Xi, Lisha Wang, Rui Tang, Wei Hua, Hujun Bao, and Rui Wang. I2-sdf: Intrinsic indoor scene reconstruction and editing via raytracing in neural sdfs. In *CVPR*, 2023. 1

[49] Michael Zollhöfer, Patrick Stotko, Andreas Görlitz, Christian Theobalt, Matthias Nießner, R. Klein, and Andreas Kolb. State of the art on 3d reconstruction with rgb-d cameras. *Computer Graphics Forum*, 37, 2018. 2